Bachelor Computer Science & Economics

The impact of generative AI on the documentation capability in the public health sector

Manuel Mol

First supervisor: Tyron Offerman
Second supervisor: Joost Visser

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

27/06/2025

Abstract

Background: Documentation in public healthcare, for example within the Salvation Army's LJ&R (Juvenile Protection) department is crucial for patient safety, quality of care, and compliance with auditing bodies. The current documentation process is often time-consuming and manual. This contributes to administrative burden and potential burnout among youth protection professionals. There is a lack of integration between systems and variability in documentation style and detail among colleagues.

Aim: This thesis aimed to investigate how the integration of generative AI (GenAI) can enhance the record-keeping capability for interviews and meetings within public healthcare organizations, focusing on the Salvation Army's LJ&R department. Specifically, the thesis measures the impact of the "Luisterlinie" AI tool on time spent on documentation, the quality and completeness of documentation, user job satisfaction, and to identify potential benefits and challenges.

Method: The study employed an action research-based approach, following cycles of literature review, diagnosis, action planning, action taking, evaluation, and reflection. A pilot study was conducted over approximately four weeks (April 14th to June 1st, 2025) with 47 active youth protection professionals from the Salvation Army. Participants used the "Luisterlinie" web application, which transcribes and summarizes conversations using OpenAI's Whisper and GPT-4o models. Data was collected through pre-pilot interviews, three surveys administered throughout the pilot (kick-off, week 1, week 4), usage analytics from the Luisterlinie app, and qualitative feedback from end-of-pilot meetings.

Results: The pilot showed promising results regarding the impact of GenAI on documentation capability. Participants reported spending less time registering contacts and found their reporting more efficient after using the Luisterlinie app. A key benefit was the reduced need for note-taking during conversations. This allowed for better focus and presence with clients. Documentation generated with the app was perceived as comparable in completeness and quality to human-generated reports. Output accuracy and speaker identification were still the most significant frustration, in line with the literature. User satisfaction with the documentation process improved, and participants overwhelmingly expressed a desire to continue using and recommend the app. Participants found it a valuable addition that reduced their administrative load.

Conclusion: The integration of GenAI through applications like the "Luisterlinie" app has clear potential to enhance record-keeping in public healthcare by improving transcription efficiency, reducing documentation time and administrative burden, and fostering better client engagement and user satisfaction. While AI-assisted documentation quality can be comparable to human efforts, the need for human oversight remains important due to limitations in output accuracy and technical stability. This is in line with findings in existing literature. Despite challenges related to the pilot phase and app stability, the strong positive sentiment from users regarding continued use highlights GenAI's potential to not only optimize workflows but also improve healthcare professionals' daily experience and reduce administrative burden.

Acknowledgements

I would like to express my deepest gratitude to God for His guidance and strength throughout this process of writing my thesis.

Contents

Chapter 1

Introduction

The rapid advancement of generative artificial intelligence (GenAI) is transforming industries, offering unprecedented opportunities for automation and efficiency [12]. The public healthcare sector, often burdened by extensive administrative tasks, stands to benefit significantly from these technological innovations [4]. Within this sector, accurate and timely documentation is essential of patient safety, quality of care, and regulatory compliance [5]. However, current documentation processes are often manual and time-consuming, contributing to a significant administrative burden on healthcare professionals. This not only leads to inconsistencies in records but also increases the risk of professional burnout and diverts valuable time away from direct client interaction [25].

While the potential of GenAI to alleviate these pressures through automated transcription, summarization, and content generation is widely acknowledged [9], a significant research gap exists between its theoretical ideas and its practical implementation. Much of the existing literature focuses on the technical design of GenAI tools or speculates on their potential benefits in controlled settings. There is inconclusive evidence on the real-world impact of integrating these tools into the complex workflows of public healthcare organizations.

To help bridge this gap, this thesis examines a specific environment: the Juvenile Protection (Leger des Heils Jeugdbescherming & Reclassering, or LJ&R) department of the Salvation Army in the Netherlands. As an international organization providing social services, the Salvation Army's LJ&R department works with vulnerable youth and families, a context where record-keeping is not just an administrative task but a critical component of care. This department, like many in the public health sector, faces challenges with manual and inefficient documentation processes, making it an ideal case for exploring the practical implications of technological intervention.

This study therefore investigates the real-world application of a GenAI tool within the Salvation Army's LJ&R department. To understand its practical impact on day-to-day operations, an action research methodology is employed. This approach was chosen because it facilitates solving a practical problem within the organization by actively introducing an intervention. In this case, a GenAI application called Luisterlinie. By focusing on the active implementation into the workflow of the LJ&R department, this research aims to provide insights into how GenAI can enhance transcription and documentation capabilities in a public healthcare setting practically.

1.1 Problem statement

The current documentation process in public healthcare, particularly within environments like the Salvation Army's Juvenile Protection (LJ&R) department, is characterized by manual, inefficient, and often burdensome workflows. This reliance on traditional methods creates a significant administrative load on youth protection professionals, contributing to job dissatisfaction and the risk of burnout—a critical issue in a sector where retaining experienced staff is paramount.

Furthermore, the time consumed by these administrative tasks directly detracts from the time available for direct client interaction. This reduces the opportunity for professionals to build rapport, conduct thorough assessments, and provide the nuanced support that vulnerable youth and their families require.

The problem is increased by inconsistencies in the quality and completeness of documentation among colleagues. Such variability not only complicates case handovers and compromises the continuity of care but also poses a significant risk to regulatory compliance and defensibility during audits.

Consequently, there is an urgent need to explore how emerging technologies like Generative AI (GenAI) can address these deep-seated inefficiencies. While GenAI promises to automate and streamline documentation, its practical applicability and impact within the complex, sensitive workflows of public healthcare remain underexplored. This study, therefore, addresses the need to practically evaluate how GenAI can alleviate administrative burdens and enhance the overall documentation capability, without compromising the quality of care or the integrity of the records.

1.2 Research questions

The main focus of this study is to address the following research question and its sub questions:

How can the addition of GenAI enhance the record-keeping capability for interviews and meetings in public healthcare organizations?

- How can the addition of GenAI enhance the **transcription capability** in public healthcare organizations?
- How can the addition of GenAI enhance the **documentation capability** in public healthcare organizations?

1.3 Overview of the thesis

The first chapter introduces the research by outlining the problem statement and the research questions it aims to answer. Chapter 2 provides the necessary background by delving into the relevant literature on Generative AI, its applications in healthcare, and the specific challenges of documentation in the public health sector. This concludes with identifying the research gap.

Chapter 3 describes the research methodology, explaining the action research framework used for the study. It explains the case selection, the cyclical research design, and the mixed-methods approach to data collection.

The next chapters follow the action research cycle. Chapter 4 covers the Diagnosing phase, presenting findings from pre-pilot interviews to understand the existing documentation process. Chapter 5 details the Action Planning phase, outlining the objectives and design of the pilot study involving the "Luisterlinie" application. Chapter 6 documents the Action Taking phase, providing a chronological account of the pilot's execution. The Evaluation of the collected data is presented in Chapter 7, which analyses the results from surveys, application logs, and meetings.

Chapter 8 contains the Reflecting phase, where the key findings are discussed, interpreted in the context of the research questions, and the study's limitations are addressed.

Finally, Chapter 9 presents the Conclusions, summarizing the main outcomes of the research and their implications. The Appendices at the end of the document provide the interview guides and survey questionnaires used for data collection.

Chapter 2

Background and related work

This chapter examines the intersection of Generative AI (GenAI) and documentation within public healthcare. GenAI tools offer the promise of relieving administrative burdens through automated transcription and summarization. There are however concerns over accuracy, reliability, and safety. To provide a comprehensive foundation for this study, this chapter will systematically review the relevant literature. It will cover GenAI, its role in healthcare, the importance of documentation, and the current state of research on AI-assisted documentation.

2.1 Generative AI

Generative AI, or GenAI is *"a technology that (i) leverages deep learning models to (ii) generate human-like content (e.g., images, words) in response to (iii) complex and varied prompts (e.g., languages, instructions, questions)"* [15]. GenAI technologies are becoming increasingly relevant in organizations, since these models can perform a wide array of different tasks, like summarization, transcription, image generation and speech synthesize [26].

Large Language Models (LLMs) like ChatGPT, are built upon the transformer architecture, which utilizes attention mechanisms to weigh the importance of different input elements during text prediction [36]. These models are then trained on massive datasets [17]. These datasets consist of websites, books, social media platforms and conversational data [17]. This data, among choice of algorithm, product design decisions and policy decisions can introduce biases and limitations into their outputs [17].

Transcription tools like Whisper can be used to transcribe audio into text, which can later be analysed by LLMs. They are built on the same transformer architecture as ChatGPT [31]. Whisper performs well in challenging conditions like background noise or multiple speakers. The model is also capable of transcribing multiple languages. The resulting text output can then be analysed by LLMs. Like other transformer-based models, Whisper's performance can be affected by the quality and characteristics of the training data, and potential biases within that data should be considered. This paper will not touch on other types of GenAI, like image or voice generation.

A lot has changed since ChatGPT's release. Current state-of-the-art models like Claude 3.5 Sonnet [6], ChatGPT 4.5 [29] and Gemini 2.0 Flash [30] have shown remarkable performance in tasks like coding, vision and problem-solving tasks. Reasoning models like OpenAI o3 [28] and Deepseek-R1 [14] show that these language models can perform well on complex problems where reasoning is required.

There are a multitude of tasks where these GenAI models can be very useful to increase productivity. Tasks involving professional writing and content creation, such as writing reports, emails and press releases, received a significant boost in productivity with the assistance of ChatGPT. In this case, the time that was needed to complete the task decreased, and the quality increased [27]. Essays written by students using GenAI also saw an increase in quality [35]. GenAI thus has a great use in assisting in writing tasks.

Another set of tasks where GenAI is very prevalent is programming and software engineering. Here, tasks can be split up into two modes of interaction: acceleration, where the tool is used to speed up a task that is already known, and exploration, where the tool helps explore options when the programmer is unsure how to proceed [7]. When accelerating a task, programmers can use tools like GitHub Copilot to speed up development by handling simple tasks. When exploring, a tool like ChatGPT can be used to explore more options to solve a particular problem [34]. Multiple iterations are however often necessary to produce good working code [33].

2.2 GenAI in healthcare

Generative AI technologies are also beginning to be utilized in healthcare. GenAI models, such as Generative Adversarial Networks (GANs), Diffusion Models, Large Language Models (LLMs), and Variational Autoencoders (VAEs) are being used for a variety of uses, like clinical decision-making, drug design and FAQ systems [4, 12].

One of these new developments is GenAI-powered chatbots in healthcare. These chatbots can help patients with chronic conditions like dyslipidemia by giving personalized advice on lifestyle changes and reminding them to take their medications. Such a chatbot also helps patients access health-related information more easily, by providing them with an easy-to-use interface. GenAI systems that can also understand pictures (Multimodal GenAI) can also be used to analyse images of food and packaging to deliver dietary recommendations [4]. Another application of chatbots is MedBot, which provides accurate medical information to patients. Responses are generated based on patients' healthcare records, medical resources such as books, and information users provide in text, images, or audio formats. MedBot enhances accessibility for users and simplifies the process of understanding medical information. This ultimately helps users make better-informed decisions about their health [16].

Another key area of development are Clinical Decision Support (CDS) systems. These systems are used by clinicians when they make patient-specific decisions. One of the ways that LLMs are used in CDS is to analyse patient data and clinical context more effectively. This can help prevent irrelevant alerts when they come up, and better prioritize patient care [12]. LLMs have also been used to summarize medical texts. This provides clinicians with more concise, relevant information for decision-making [18].

Automatic Speech Recognition (ASR) models in healthcare represent another key development. These models can convert audio into text. The transcribed inputs can then serve as input for a large language model (LLM) to generate a report or analyze the transcript [9]. However, these models struggle with real-world complex conversations due to a lack of context, interruptions, and non-standard speaking styles in the dialogue [23].

While these systems can be very useful, they also present certain risks when implemented in healthcare. One of these risks is that LLMs tend to generate inaccurate or fabricated information. LLMs also rely on their training data, which can be restricted and not recent. is a significant concern in healthcare where accuracy is crusial [18]. Privacy is another significant issue; sensitive data should not be sent to LLM providers. Healthcare providers must ensure that these systems comply with HIPAA or other privacy regulations [4]. Furthermore, LLMs often function as “black boxes,” making it challenging to understand their outputs. Explainability is essential for building trust with both patients and healthcare professionals [12].

2.3 Documentation capability in public healthcare

Documentation in healthcare is important to keep track of patient data. It must be completed to a high standard to ensure patient safety and the quality of healthcare services [5]. Healthcare institutions are transitioning from paper-based systems to electronic medical records (EMR). This is because paper-based systems introduce issues in quality and completeness of medical entries. EMRs have potential to reduce errors by improving quality and completeness [24].

The EMR include all information related to patient care [5]. Data entry in EMRs is typically done through typing, but it is theorized that transcription and voice recognition systems can be used to document faster [5, 8]. Some EMR systems also have features like decision support, designed to aid healthcare professionals in their clinical decision-making processes [8].

Inadequate documentation has some significant consequences. Inadequate documentation can contribute to an increased rate of medical errors and mortality in healthcare settings [24]. It can also cause delayed patient care [24]. It can also lead to inaccurate performance measurement, which can lead to inappropriate allocation of health funding [24]. Inadequate documentation can also make auditing paper-based records difficult [5]. Besides this, the burden of documentation, potentially exacerbated by inadequate or inefficient documentation processes, is a key factor contributing to burnout among healthcare professionals [25].

2.4 GenAI and documentation in healthcare

The main focus of this proposed study will be in healthcare related tasks. GenAI can be used to increase productivity in tasks such as patient facing interactions and administrative work. This allows healthcare providers to spend more time on patient care. A case study showed that it is possible to use natural language processing and automatic speech recognition to transcribe patient-clinician interactions, using Large Language Models to draft clinical notes. This approach improved documentation efficiency and patient-centred care. It also enhanced note quality, showing that there is great potential in LLM supported clinical documentation [9]. However, another study found that clinical notes generated from audio recordings did not consistently meet clinical standards for quality and reliability [22].

Transcription is classically done with the help of Medical Transcriptionists (MTs) [20]. This process involves a doctor using a template from the EPR and dictating the conversation with the patient. This dictation is then sent to the MT for transcription. The MT transcribes the information, and may send a note to the doctor if something is incorrect or unclear. The transcribed document can then be added to the EMR of the patient. This process is quite elaborate.

Breakthroughs in speech recognition software like Whisper have created new opportunities for the usage of GenAI in healthcare. Automatic transcriptions have a number of benefits [23]. First, it reduces physical paperwork by recording and saving everything digitally. It also saves time, both in conversation and after the conversation when reporting. This is great for patient engagement, since healthcare professionals can focus on talking to patients instead of writing things down. Another benefit is the improved workflow. There is no need for Medical Transcriptionists and sending dictations around. This also improves data security and privacy, since speech recognition systems are designed to prioritize data privacy and security. They also adhere to industry regulation and use encryption for data transfer.

AI driven documentation has seen some promising results. There is a growing interest in AI-driven documentation [10]. AI generated documentation often meets or surpasses traditional documentation standards. It is important to keep in mind the presence of "hallucinations", or fictitious information generated, as multiple studies have encountered quality issues related to them [10]. GenAI technologies have demonstrated a substantial potential to improve efficiency in clinical documentation. The most notable improvement was found in complex cases, where efficiency gains were the highest. This time-saving aspect is very important, since it has a direct impact on the workload of healthcare workers. Burnout has also been linked to extensive EMR use. These efficiency gains provided by AI could provide a promising solution to reduce burnout. It is however important to keep in mind the quality of the work produced.

2.5 Research gap

While research has explored the potential of Generative AI (GenAI) in clinical documentation, the findings present a mixed picture. Some studies, like Biswas et al. [9], highlight potential improvements in efficiency and the quality of clinical notes drafted using Large Language Models (LLMs) from transcriptions. Conversely, other research, such as Kernberg et al. [22], raises significant concerns regarding the consistency, accuracy, and overall clinical adequacy of AI-generated documentation derived from audio recordings. This conflict indicates that the practical usability and effectiveness of these tools within the context of public healthcare are not yet fully researched.

There is a notable gap in understanding how effectively current GenAI tools for transcription (like Whisper) and subsequent documentation/summarization (using LLMs) actually enhance the documentation capability within public healthcare organizations. The uncertainty lies in whether the theoretical productivity benefits [27, 9] can be reliably realized without compromising the quality and completeness standards crucial for patient safety [5, 24], especially given the documented risks of inaccuracies and hallucinations [18, 22, 10]. This study aims to bridge this gap by empirically evaluating the impact of integrating GenAI on both transcription processes and overall documentation capability for interviews and meetings in a public healthcare setting, providing needed evidence for organizational decision-making.

Chapter 3

Methodology

This chapter outlines the methodological approach used in this study to investigate the impact of a transcription and summarization application on documentation capabilities at the Salvation Army. It will cover the research design, case selection, data collection & analysis, and the surveys.

3.1 Research Design

This study will use an action research-based approach. The action research methodology provides a structured framework for evaluating the impact of the transcription and summarization application on the documentation capability at the Salvation Army. This method follows a cycle consisting of literature review, diagnosis, action planning, action taking, evaluation, and reflection [13].

Action research was chosen over other research methods like a case study or design science because the goal was to solve a practical problem in the organization. A case study would be too passive, since the goal was not to only observe the existing documentation process, but to actively intervene by introducing GenAI application for interview and meetings and evaluate the changes. Design science was also considered, because the project involves creating a technical artifact. However, the primary focus of this thesis is less on the technical design of the GenAI application, but more on the impact the application has on the organization, and how it integrates into the workflows of Salvation Army. Action research provides a solid framework for this, as it is designed to implement change and study its effects in a real-world setting.

The six steps of action research are explained in more detail below:

- **Literature review:** The first step is to review the literature. Research will be done on the topics of action research and GenAI in healthcare. This research will support the future action planning and diagnosis phases.
- **Diagnosing:** The diagnosing stage will involve describing the context and situation. The problem will be clearly defined at this point. Further relevant information is gathered about the current situation by interviewing stakeholders. This phase also includes collecting data from those involved in the pilot to establish a baseline.
- **Action planning:** In this step, a pilot will be developed based on the findings from the literature review and interviews with stakeholders and employees. The goal of the pilot is to observe how documentation capabilities change when GenAI tooling for interviewing and meetings is available. This step also includes planning the action, such as determining the format of the pilot, the data to be collected, and the methods for collecting this data.
- **Action taking:** At this stage, the planned actions are implemented. The pilot will be conducted, and data will be collected. That data that will be collected is detailed in Section 3.3.
- **Evaluating:** After the action is taken, the data will be analysed to assess the effects of the intervention. This process will involve comparing metrics related to documentation efficiency, quality, and user satisfaction. Both quantitative measures, such as application logs, and qualitative feedback will be considered.

- **Reflecting:** In this final stage, we interpret the evaluation results within the broader context of the research questions. This includes identifying key insights, limitations of the study, and implications for practice. Conclusions are drawn from this reflection, which will also inform recommendations for future research and practical implementation.

3.2 Case Selection

This research has been conducted at the Salvation Army. The Salvation Army is an international movement and belongs to the universal Christian church. Its message is based on the Bible. Its mission is to preach the Gospel of Jesus Christ and to alleviate human suffering in His name without any form of discrimination [2]. I will conduct my research in collaboration with the AI team at Salvation Army, as well as the LJ&R (Juvenile protection). The AI team is working on a GenAI transcription and summarization application. The LJ&R (Juvenile protection) department will be one of the end users of this application. Juvenile protection focuses on the safety of children. The aim is to remove threats to a child's development. They provide guidance on supervision orders, guardianship measures, and coercive measures [1]. When parents fail to provide safe and responsible care for a child, the court can intervene and mandate assistance from youth protection services. Salvation Army then steps in to help the parents and child. The Juvenile protection division is a national division spread over different cities in the Netherlands. A pilot will be conducted in which the LJ&R department will utilize the Luisterlinie application for interviews and meetings. The Luisterlinie application records, transcribes and summarizes conversations and meetings into the correct form of documentation, reducing the documentation burden. A more detailed walkthrough of the Luisterlinie app is located in Section 5.3.

The research was conducted at the Salvation Army because they are developing innovative technology for healthcare. This thesis presents a unique opportunity to pilot this technology scientifically, something that has not been done before within the Salvation Army. As a result, the Salvation Army benefits from a scientific analysis of the tools it is developing. This analysis also addresses the research gap presented in Chapter 2 and will contribute to the scientific field of documentation and GenAI in a practical application.

3.3 Data Collection and Data Sources

This research will employ triangulation to enhance the validity and reliability of the findings. Four data collection methods will be utilized: diagnostic interviews, surveys, usage data from the Luisterlinie app, and qualitative feedback from the end-of-pilot meeting with participants. The variety of collection methods will ensure that the conclusions drawn are properly substantiated. The data collection will measure the following key metrics to address the research question:

1. **Time spend:** The impact of "Luisterlinie" on the time spent by healthcare professionals on transcription tasks
2. **Quality & completeness:** The quality and completeness of the documentation generated by the "Luisterlinie" application compared to human-written documentation
3. **Job satisfaction:** The job satisfaction of the users of the application
4. **Benefits and challenges:** Potential benefits and challenges that may arise when integrating "Luisterlinie" into the existing workflows

Four collection methods will be used to gather data to measure these key metrics:

1. **Pre-pilot interviews:** These interviews are primarily used for diagnosing the problem and will be utilized to find potential challenges and benefits in the current record-keeping capability.
2. **Surveys:** Three surveys will be conducted to gather data during the pilot phase. The first survey will collect baseline data about documentation within the Salvation Army. The second survey will assess the short-term effects and gather initial feedback after the tool is introduced. The final survey will evaluate the long-term effects of the application.
3. **Luisterlinie analytics:** Analytics from the Luisterlinie application will be collected during the pilot.

4. **End-of-pilot meeting:** The end-of-pilot meeting will be an opportunity for participants to give feedback on the pilot and the app after the pilot.

The key metrics will be measured using the collection methods in the following way:

- **Time spend:** The surveys, combined with the pre-pilot interviews, will be used to measure the time spend on documentation tasks. The pre-pilot interviews will give us insight into the steps of the documentation process for youth-protection professionals. The surveys will be used to determine if there are changes in reported time spend on documentation after the Luisterlinie application is introduced. Analytics from the Luisterlinie application will also be used to gather data about usage patterns.
- **Quality and Completeness:** The surveys, Luisterlinie analytics, and end-of-pilot meeting will provide valuable insights into the quality and completeness of the generated documentation. Surveys will measure users' perceptions of the quality and completeness of the Luisterlinie-generated documentation compared to the kick-off meeting and diagnosing phase interviews. The analytics gathered by the Luisterlinie application can serve as indirect quality metrics. Measures of how frequently and extensively users edit the transcript indicate potential quality issues. The end-of-pilot meeting will provide more detailed feedback on the perceived quality and completeness of the generated documentation.
- **Job Satisfaction:** The surveys and end-of-pilot meeting will be used to measure job satisfaction with the documentation process. The kick-off survey will give a baseline of the overall satisfaction with the documentation process. Surveys during the pilot will monitor the change in satisfaction. The end-of-pilot meeting will give in-depth feedback on the satisfaction changes.
- **Benefits and Challenges:** Benefits and challenges will come from the diagnosing interviews, surveys and end-of-pilot meeting. The pre-pilot interviews will identify potential benefits and challenges based on the current approach. Surveys will monitor any emerging benefits and challenges that might come up in the pilot. The end-of-pilot meeting will give participants an opportunity to share their experiences and benefits and challenges.

3.4 Data analysis

This study will use a mixed-methods approach to data analysis, using both quantitative and qualitative data. This approach allows for the triangulation of findings, where insights from one source are combined with another, enhancing the overall results.

- **Quantitative Analysis**: Data from the three surveys (specifically the Likert scale questions) and the Luisterlinie application logs will be analysed quantitatively. Descriptive statistics, including means, standard deviations, frequencies, and percentages will be calculated to summarize responses and usage patterns. This analysis will help quantify the impact of the Luisterlinie app on metrics such as time spent, perceived efficiency, and user satisfaction. Results will be presented in tables and charts.
- **Qualitative Analysis**: Qualitative data from the pre-pilot interviews, open-ended survey questions, and the end-of-pilot meetings will be analysed using thematic analysis. This process involves several steps: first, reading transcripts and responses. Second, generating codes that identify interesting features, and third, collecting codes into themes. This method will be used to identify patterns, user experiences, specific benefits, challenges, and suggestions for improvement that are not captured by quantitative measures.
- **Integration and Triangulation**: Triangulation will involve integrating the quantitative and qualitative findings. For instance, a quantitative result from the surveys showing a high mean score for "reduced administrative burden" will be explained by qualitative quotes from interviews where participants describe exactly how the app saved them time. By comparing the different data sources, a more robust and nuanced understanding of the GenAI tool's impact on documentation capability can be achieved.

3.5 Surveys

Throughout the four-week pilot, participants will be asked to complete three surveys: a survey right at the kick-off meeting, a survey after the first week of use and a survey in the last week of the pilot. These surveys over time will give insight into the change that occurs when the Luisterlinie app is introduced. These surveys will serve slightly different purposes. The surveys mainly aim to:

1. Monitor reported changes in the time spent on documentation tasks compared to previous methods.
2. Assess participants perceptions of the quality and completeness of the documentation generated by the Luisterlinie application.
3. Monitor changes in job satisfaction concerning the documentation process while using the Luisterlinie application.
4. Identify benefits or challenges encountered during the week's usage, including usability issues or workflow impacts.

The different surveys will serve different goals in the data collection process:

- **Kick-off survey:** The kick-off survey will serve as a baseline to gather initial data on the documentation process inside LJ&R. The questions in this survey are largely based on the diagnosing phase. Which gave us more insight into the inner workings of the documentation process of LJ&R. The survey is deliberately kept short, since the survey will be taken at the kick-off meeting, and shouldn't take up a lot of time. The Likert scale was used for most questions because these types of questions are generally less exhausting to answer and provide a good source of data [19]. The study by [21] was also used as a reference when designing the survey. The kick-off survey can be found in appendix B.
- **Week 1 survey:** The week one survey will provide insights into the initial changes following the introduction of the Luisterlinie app. In addition to the main goals outlined above, it will help identify any initial issues and problems that participants may encounter while using the app for the first few times. It will also measure their initial impression and struggles with the Luisterlinie app in actual usage. This survey is quite a bit longer than the kick-off survey, since people can complete it in their own time. The full survey used can be found in appendix C.
- **Week 4 survey:** The week four survey will be the last survey that the participants will complete. The results of this survey will provide insight in the long term usage of the Luisterlinie app. In addition to the main goals outlined above, it will explore if the Luisterlinie app has got integrated into the working process in long term use. This survey will follow a similar format as the survey from week 1. The full survey used can be found in appendix D

Chapter 4

Diagnosing

To diagnose the current situation at Salvation Army regarding their documentation capability, semi-structured interviews were held. These interviews were taken in the beginning of April. Three people from LJ&R were interviewed: A youth-protection professional, a primary process team leader and the administration team leader. This gave a comprehensive look into the administration process from different levels at Salvation Army. The interviewees were asked about their current documentation process, their satisfaction and challenges with this process, and their perspective on the use of GenAI in the workplace. The current situation regarding the documentation capability at Salvation Army was determined based on these interviews. The interview questions used can be found in Appendix A.

4.1 Background

These interviews gave a good insight into the inner workings of the administrative process of the Salvation Army. The documentation is done in a patient registration system called WIJZ. A contact journal is created in WIJZ when a youth-protection professional interacts with a client. Examples of these interactions are phone calls, in-person meetings and email conversations. Every contact, or interaction, should be logged as a contact journal in WIJZ. This makes sure that everything is logged. Documenting these interactions in this way is important for several reasons. Firstly, documentation needs to be up-to-date and accurate to handle complaints and facilitate file access. They are regulated and audited by numerous different parties (estimated at 11-12, like the Keurmerk Instituut), which check if dossiers are in order and traceable. These parties each have their own rules and requirements. It was noted that meeting all these disparate rules is extremely difficult and time-consuming, leaving little room or capacity to focus on optimizing efficiency or working "smarter." The primary focus becomes simply ensuring compliance with all external demands. Secondly, in case a youth-protection professional gets ill, there should be sufficient documentation about the clients of the youth-protection professional to make sure the work can be picked up by another youth-protection professional.

The number of contacts varies, but is estimated to be between 5 and 15 per day. This includes both physical meetings and phone calls. These contacts can range from very short check-ins to longer, more formal discussions. It was mentioned that there is never a workday without some form of client contact.

A significant portion of these contacts are done over the phone. The estimated ratio between telephonic and physical contact is approximately 75% telephonic to 25% physical. Phone calls are often used for quick check-ins or immediate reactions to situations. The daily amount of calls fluctuates significantly, ranging from just a few calls on some days to as many as 30-40 on others.

The planned, more formal meetings (like evaluations or home visits) typically last between 45 minutes and 1.5 hours. Phone calls or ad-hoc contacts can range from 2 minutes to half an hour. The longer planned meetings usually require more extensive documentation afterwards.

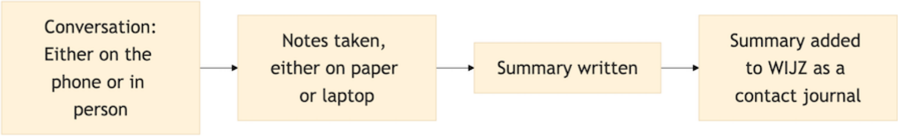

Conversation: Either on the phone or in person
Notes taken, either on paper or laptop
Summary written
Summary added to WIJZ as a contact journal

Figure 4.1: The documentation process

4.2 Current documentation process

As mentioned, documentation about patients happens in a patient registration system called WIJZ. When on the phone, the youth-protection professional described that they open the client's file in WIJZ to type the contact journal simultaneously. Notes are written down during physical meetings. Some colleagues use a laptop to type notes during the meeting. After the meeting, the youth-protection professional creates a contact journal in WIJZ with a summary of the conversation. The format of this summary is up to the youth-protection professional. In practice, not every interaction is documented as a contact. The reason for this is that, with a high volume of daily contacts, immediately writing a report for each is challenging. It feels less necessary for very brief interactions (like confirming a child is going to school). This makes the documentation process vulnerable. If the youth-protection professional were to become ill or leave, colleagues wouldn't have a clear, documented history of events. This make continuity difficult.

Estimates for documenting an individual contact, such as a phone call or a brief interaction, ranged from 2 to 10 minutes on average. For documenting longer, planned meetings (like evaluations or home visits), it takes between 15 and 30 minutes afterwards to write the report. It was emphasized that the time spent depends on the complexity of the contact or the client/family involved. If a family is known to be prone to complaints, the interaction is documented more extensively, which takes more time.

The primary tool for documentation used is WIJZ, the patient registration system. For note-taking during meetings or calls, pen and paper or laptops are used. Afterwards, a computer is used to enter the summarized information into WIJZ. Audio recording is not used as a standard method. It is only done in very rare cases if explicitly requested by the parents.

There is currently no detailed, step-by-step manual that describes how to conduct the documentation process itself (e.g., "step A: take notes on X, step B: use format Y for the report"). There is an expectation and requirement that all contacts must be documented in WIJZ. The process of entering data into the system is standardized, with fields like "subject" and "description". The specific method of note-taking or the level of detail in the summary report can vary. A "contact journal" (which used to be in Word but is now part of the WIJZ system) is created for every contact, including internal meetings discussing clients.

4.3 Challenges and successes

The current documentation process works quite well. Some aspects are already automated (automatically prefilling certain fields that are already in WIJZ). There are however some challenges. A major area of improvement would be a link between email and WIJZ. This would allow emails to be automatically attached or linked to client files, eliminating the need for manual copy-pasting. There is also a significant amount of variance in how different colleagues document conversations and how much detail they include. Furthermore, too much time is spent on the documentation process, particularly on writing out the reports after meetings. The underlying challenge could be that there is little room or capacity to focus on optimizing efficiency or working "smarter", because of the rules and auditing requirements of different parties. The primary focus becomes simply ensuring compliance with all external demands.

4.4 Summary

To summarize, the main problem that was found after the interviews is that people spend too much time on documentation. This has a variety of underlying reasons, such as: the high volume and sometimes short nature of daily contacts; the dependence on manual note-taking during physical meetings followed by a separate entry into the system (WIJZ); the lack of integration between email and the WIJZ system, resulting in a lot of manual copy-pasting; variation in the level of detail and format used by different professionals when creating contact journals; and the substantial time and capacity required to meet the requirements of numerous auditing and regulatory bodies. This external pressure prioritizes compliance over the optimization of internal efficiency.

The importance of accurate documentation is clearly understood within LJ&R. The current processes in WIJZ combined with manual methods are seen as inefficient. There is a tension between complete logging and efficiency.

The diagnostic interviews showed that some basic automation in WIJZ exists, but the overall process still relies on manual effort. The documentation takes significant time for youth-protection professionals. These inefficiencies represent key opportunities for potential improvements. This could include the introduction of intelligent automation solutions like GenAI, which could aim to reduce the administrative burden while maintaining or enhancing the quality and completeness of documentation.

Chapter 5

Action planning

In the action planning phase of the research, the primary objective was the creation of the pilot. The pilot will consist of a 4-week period where we will measure what effect the AI tool developed by the Salvation army, called "Luisterlinie" will have on the record-keeping capability of the Salvation army. The Luisterlinie application is a web-app that is able to transcribe and summarize both interviews and meetings. A user can start a recording with consent of the participants. After the recording has finished, a transcript is created using the OpenAI Whisper model. The transcript is then analysed using the GPT-4o model to extract relevant information from it. This information is then presented to the user in the form of a webpage.

5.1 Objectives of the pilot

The overarching aim of this pilot is to investigate **how the incorporation of GenAI can enhance the record-keeping capability for interviews and meetings within public healthcare organisations**, specifically focusing on the Salvation Army's LJ&R (Juvenile protection) department. This objective is aligned with the central research question. The pilot, which will run for a period of four weeks, intends to measure the effect of the "Luisterlinie" AI tool, developed by the Salvation Army, on the record-keeping abilities of the participating department.

More specifically, the pilot seeks to achieve several key objectives. Firstly, it aims to assess the impact of "Luisterlinie" on the time spent by healthcare professionals on transcription tasks. Secondly, it will focus on evaluating the quality and completeness of the summaries generated by the "Luisterlinie" application compared to human written summaries. Finally, understanding the job satisfaction of the users of the application is also a key objective as this will provide insight into the tool's usability and integration into their daily work, which also impacts the effectiveness of the tool.

Furthermore, the pilot aims to identify both the potential benefits and the challenges that may arise when integrating "Luisterlinie" into the existing workflows within the public healthcare setting.

5.2 Format and Scope

The pilot takes place from the 14th of April till the 23rd of May. This timeframe allows us to gather the data needed to answer the research question and objectives. We will have about 60 participants divided into two groups. This arrangement helps reduce the load during the kick-off meetings, which would otherwise be quite large. These participants are part of the juvenile protection division of the Salvation Army, specifically the youth-protection professionals. Youth-protection professionals from multiple different cities will be participating in the pilot. These cities and professionals were selected by the Salvation Army.

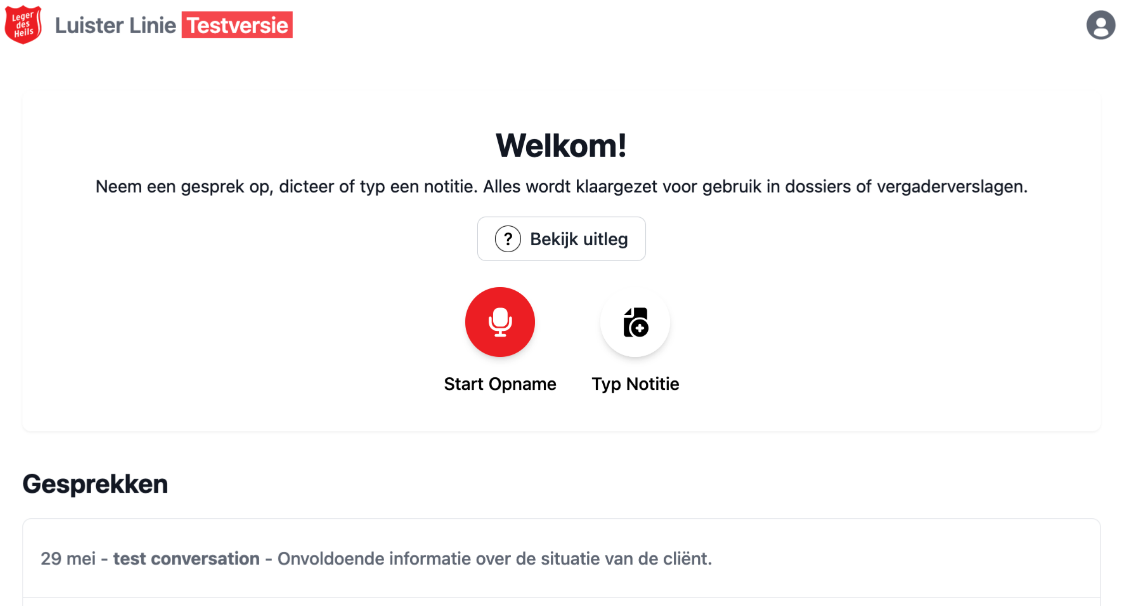
Participants will access the Luisterlinie application on both their phones and computers as a web app. They will primarily use their phones to record conversations. Once the recording is complete, the transcription and analysis process begins. During this time, the participant can close the website on their phone. They can then open the website on their computer to document the conversation with the assistance of the Luisterlinie platform.

The types of conversations that will be included in the pilot will be conversations between the youth-protection professionals and their clients. The youth-protection professional will always ask for permission before recording the conversation. Conversation audio is never saved.

The participants might run into certain issues during the pilot. To make sure these issues are reported and solved rapidly, a Microsoft Teams channel will be set up. This Teams channel will include the participants of the pilot, members of the technical team, and supervisors from LJ&R. The participants can ask questions in the Teams channel. The Teams channel will also be used for announcements and to inform participants.

5.3 The Luisterlinie app

The Luisterlinie app is a web app that can both be used on mobile and laptop/desktop computers. Its main purpose is to record conversations and then transcribe and summarize them in a format useful for the users. The following section will walk through recording a conversation to show how the app works and what it looks like.



Welkom!

Neem een gesprek op, dicteer of typ een notitie. Alles wordt klaargezet voor gebruik in dossiers of vergaderverslagen.

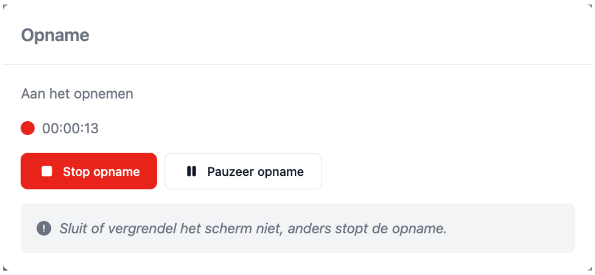
Gesprekken

29 mei - **test conversation** - Onvoldoende informatie over de situatie van de cliënt.

Figure 5.1: Home screen of the Luisterlinie app

Once the user logs in with their Salvation Army account, they land on the home screen. On the home screen, they can start a recording or manually type a note. The home screen also displays previous conversations and provides quick access to an explanation of the app and its functions.

Once the user clicks the button to start a recording, a pop-up will appear. In this pop-up, the user can select the conversation type (either a conversation with a client or a meeting) and indicate whether they want to record the conversation or take notes afterward. The user can then enter the client's first name and confirm that the client has given consent to be recorded. Additionally, users can choose to record their computer audio if they are in an online meeting.

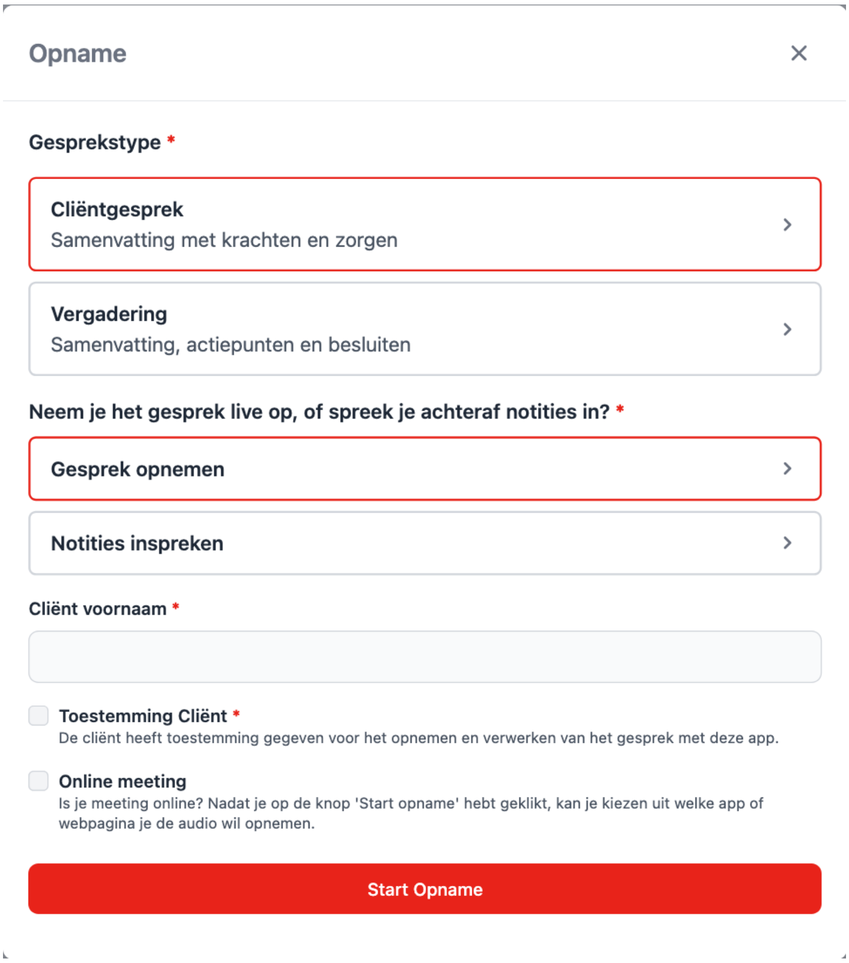

Opname

Aan het opnemen

● 00:00:13

Sluit of vergrendel het scherm niet, anders stopt de opname.

Figure 5.2: The Luisterlinie app has started recording


Opname

Gesprekstype *

Cliëntgesprek
Samenvatting met krachten en zorgen

Vergadering
Samenvatting, actiepunten en besluiten

Neem je het gesprek live op, of spreek je achteraf notities in? *

Gesprek opnemen

Notities inspreken

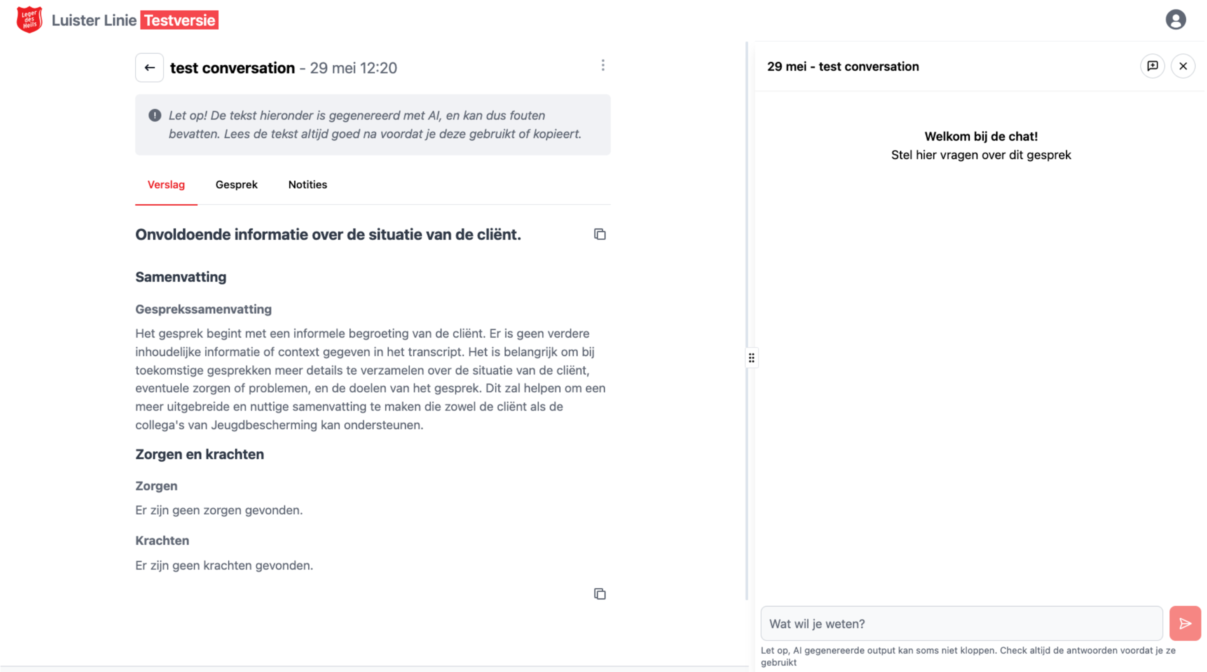
Cliënt voornaam *

□ Toestemming Cliënt *
De cliënt heeft toestemming gegeven voor het opnemen en verwerken van het gesprek met deze app.

□ Online meeting
Is je meeting online? Nadat je op de knop 'Start opname' hebt geklikt, kan je kiezen uit welke app of webpagina je de audio wil opnemen.

Figure 5.3: The recording pop-up in the Luisterlinie app

When the user clicks the "Start recording" button, the recording begins. A new pop-up will appear with options to stop recording or pause. There will also be a timer and a red circle to indicate that the recording is active.



test conversation - 29 mei 12:20

Let op! De tekst hieronder is gegenereerd met AI, en kan dus fouten bevatten. Lees de tekst altijd goed na voordat je deze gebruikt of kopieert.

Onvoldoende informatie over de situatie van de cliënt.

Samenvatting

Gesprekssamenvatting

Het gesprek begint met een informele begroeting van de cliënt. Er is geen verdere inhoudelijke informatie of context gegeven in het transcript. Het is belangrijk om bij toekomstige gesprekken meer details te verzamelen over de situatie van de cliënt, eventuele zorgen of problemen, en de doelen van het gesprek. Dit zal helpen om een meer uitgebreide en nuttige samenvatting te maken die zowel de cliënt als de collega's van Jeugdbescherming kan ondersteunen.

Zorgen en krachten

Zorgen

Er zijn geen zorgen gevonden.

Krachten

Er zijn geen krachten gevonden.

29 mei - test conversation

Welkom bij de chat!
Stel hier vragen over dit gesprek

Let op, AI gegenereerde output kan soms niet kloppen. Check altijd de antwoorden voordat je ze gebruikt

Figure 5.4: The conversation analysis

After the recording is finished, it will be transcribed and summarized. The user will then be directed to a page containing the analysis and summary of the conversation, as illustrated in Figure 5.4. Additionally, there is a chat window where the user can ask specific questions about the conversation or request further analysis. The transcript can also be edited, and users can add additional notes to the conversation.

There is some other functionality besides the functionality shown above. The user can add written notes, which will be summarized/analysed in the same way. It is also possible to change the conversation type after the fact.

5.4 Data collection

As detailed in Section 3.3, data collection will be done through four different methods: pre-pilot interviews, surveys, analytics from the Luisterlinie app and the end-of-pilot meeting. The diagnosing interview questions can be found in Appendix A, and the results of those interviews can be found in Chapter 4.

The surveys can be found in Appendix B, C and D. The first survey will take place right before the kick-off meeting. The second survey will be administered one week after the kick-off meeting, and the final survey will occur in the last week of the pilot.

The logs will be collected using a tool called Application Insights by Azure. This is integrated into the Luisterlinie application and allows monitoring of specific user actions, such as starting a recording or sending a chat message. The contents of these conversations or chat messages are neither logged nor analysed, since they contain very private information.

The primary purpose of the end-of-pilot meeting is to inform participants about the next steps for the Luisterlinie app and to gather their feedback on its usage and experience.

More information about the purpose of these collection methods can be found in Section 3.3.

5.5 Survey distribution

The three surveys, as detailed in Chapter 3, will be distributed using Qualtrics. This software, provided by the University, facilitates the easy creation and distribution of surveys. Initially, the surveys will be created in PDF format and then transferred to Qualtrics. The surveys are located in Appendix B, C, and D.

Qualtrics offers various methods for distributing surveys. Two methods will be used. For the kick-off survey, a URL will be shared during the Teams meeting, and participants are expected to complete the survey while in the meeting. The other surveys will be distributed based on attendance from the kick-off meeting. A list of email addresses will be compiled, and emails containing the surveys will be sent to the participants.

5.6 Kick-off Meeting

The primary purpose of the kick-off meeting is to inform the participants of the pilot about how the pilot will be conducted, how the Luisterlinie application works and what participants should expect. The kick-off meeting is planned to take one hour and will take place online using Microsoft Teams, since participants come from all over the country. The team-lead and at least one person from the technical team will be there. The kick-off will be led by the team-lead. The kick-off meeting will include the following segments, summarized in Table 5.1. A more detailed explanation is below

Duration	Activity	Description
5 min	Introduction	The team introduces themselves and there is a quick round of introductions
5 min	Survey & consent form	Participants fill out both the survey and consent forms
20-min	Demo & questions	A demo of the Luisterlinie app is given. There is an opportunity during the demo to ask questions
15 min	Hands on	Participants are encouraged to try the Luisterlinie app by recording the kick-off meeting. Questions can also be answered
10 min	Communication	Participants are informed about the Teams channel and expectations
5 min	Final notes	Participants are informed about the duration of the pilot and encouraged to use the Luiserlinie app as much as possible

Table 5.1: The agenda of the kick-off meetings

5.6.1 Welcome, Explanation pilot and expectations

The kick-off session will start with a brief welcome and an introduction of the team lead and technical team. Following that, there will be a concise explanation of the pilot's purpose.

5.6.2 Explanation what we are measuring

After the welcome, there will be a brief explanation of what will be measured. This will provide an overview of the surveys and data collection in the Luisterlinie application. This approach ensures that participants clearly understand what is expected of them and what data will be collected. Afterward, participants will have a moment to complete the kick-off survey and fill out the consent form.

5.6.3 Short demo and hands on

This short explanation is followed with a quick walkthrough through the app. A comprehensive tutorial is given on how to operate the app. After this, the URL to the Luisterlinie application is shared in the chat. Participants are then guided to help them put the application on their home screen for easy access. Following this, the participants are then encouraged to go use the Luisterlinie app and record the kick-off meeting. They will be given the opportunity to share their experience with the Luisterlinie application. They will also get the opportunity to ask any questions they might have about how to use the Luisterlinie application.

5.6.4 Explanation of communication to participants

A brief explanation is provided about the Microsoft Teams channel, where participants can ask questions during the pilot. They will receive access to this channel. The expectations from participants are also clearly communicated. Participants are expected to actively use the Luisterlinie application throughout the 4-week period. Participants should also ask for help if something isn't working. Finally, they are expected to complete the three surveys.

5.6.5 Preparing for end of pilot (scaling up expectation)

Participants will be informed that the Luisterlinie application is not fully ready for use after the pilot, helping to temper their expectations.

5.7 End-of-Pilot Meeting

The primary purpose of the end-of-pilot meeting is to inform participants about the next steps for the Luisterlinie app and to gather their feedback on its usage and experience. These meetings are expected to last about an hour. To avoid scheduling conflicts, two separate meetings will be held. The agenda includes the following segments:

5.7.1 Welcome

The meeting will begin with a brief welcome to explain its purpose. The team will express its desire for feedback and would appreciate hearing both positive and negative experiences with the app.

5.7.2 Feedback & Questions

Questions for the meeting will be prepared in advance based on the survey results. These questions aim to provide insight into participants' experiences with the Luisterlinie app, and to clarify usage patterns and survey results that were previously unclear. Participants will also have the opportunity to provide open feedback and ask questions.

5.7.3 Final Remarks

After feedback has subsided or if time is running short, final remarks will be made. It will be noted that the Luisterlinie app will shut down after the pilot. Participants will be thanked for their involvement and reminded that they can always reach out to a member of the pilot team with any questions.

5.8 Ethical Considerations

Several ethical factors were considered when planning the pilot. Firstly, all audio is never saved, and transcriptions are only accessible to the employee who creates them. This approach minimizes the sharing of sensitive information. A Data Protection Impact Assessment (DPIA) has been conducted on the Luisterlinie application to ensure compliance. Patients participating in the conversation must provide consent before any recording takes place. Participants in the pilot are clearly informed about what data will be collected and how it will be processed. Usage analytics from the Luisterlinie application will be anonymized. Informed consent will be obtained from participants through an informed consent form.

Chapter 6

Action taking

This chapter outlines the execution phase of the action research, specifically focusing on the 4-week pilot of the Luisterlinie app with the LJ&R department. The pilot ran from April 14th till the 30th of May 2025, a week longer than planned. A total of 47 active participants participated in the pilot. The actions followed the plan outlined in Chapter 5.

Table 6.1: Weekly Active Users

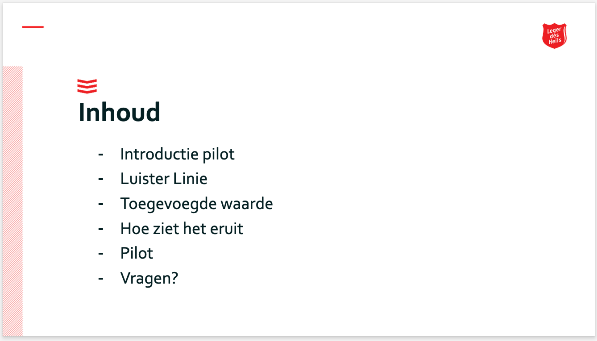
Week	Active Users
Week 1: 2025-04-14/2025-04-20	52
Week 2: 2025-04-21/2025-04-27	44
Week 3: 2025-04-28/2025-05-04	44
Week 4: 2025-05-05/2025-05-11	37
Week 5: 2025-05-12/2025-05-18	55
Week 6: 2025-05-19/2025-05-25	46
Week 7: 2025-05-26/2025-06-01	27

Table 6.1 displays the weekly active users for each week. This provides context regarding usage patterns throughout the pilot period.

6.1 Week 1: kick-off meetings (Apr 14th - 20)

The 14th of April was the first week of the pilot. The pilot week consisted of three kick-off meetings, with a forth in week two. Eleven people attended the initial kick-off meeting, six were present at the second, ten at the third, and sixteen at the final meeting.

6.1.1 Kick-off meetings



Inhoud

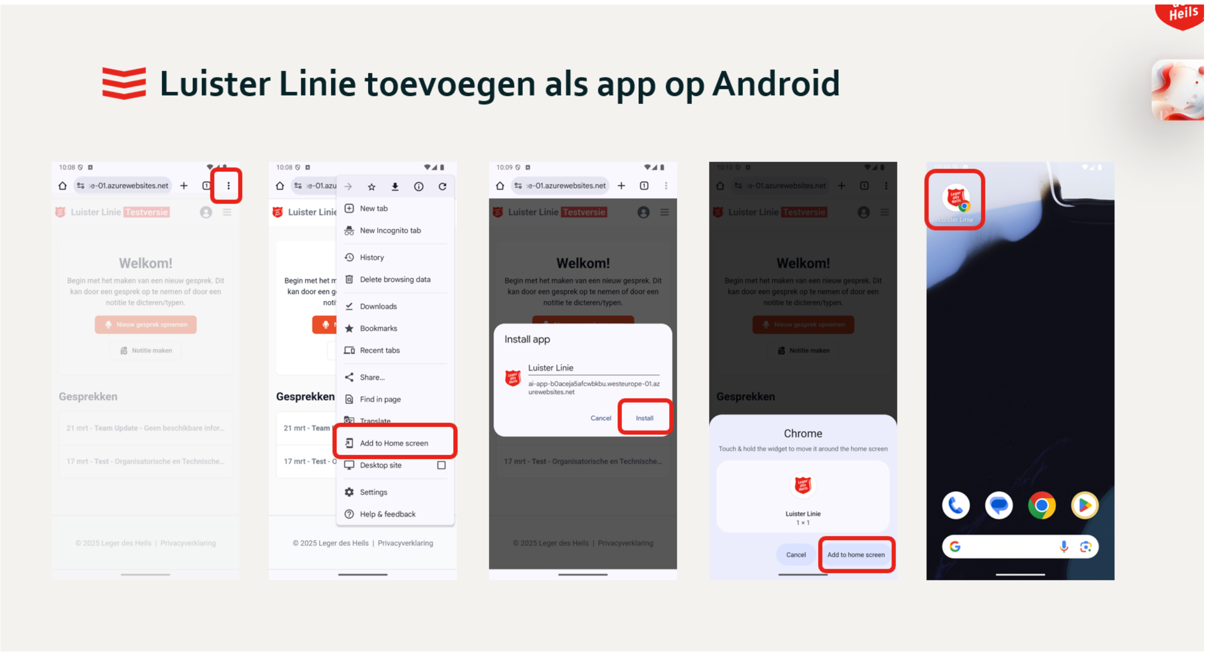
- Introductie pilot
- Luister Linie
- Toegevoegde waarde
- Hoe ziet het eruit
- Pilot
- Vragen?

Figure 6.1: First slide of the kick-off presentation showing the contents

The kick-off meetings took place on the 14th, 15th, 16th and 24th of April online in Microsoft Teams. The agenda for the kick-off meetings can be found in Table 5.1. All participants who had registered for the pilot were present at the kick-off meetings, with some joining a later kick-off meeting if they couldn't make the first. The kick-off started with an overview of the contents of the meeting and a brief introduction of those in attendance, which can be seen in Figure 6.1. This included the team leader from administration, who primarily presented, along with two developers from the team responsible for building the Luisterlinie app. They were available to answer technical questions. Following that, the participants completed the kick-off survey found in appendix B. Although the survey was expected to take about three minutes, it took closer to five minutes. After that, a quick demo was given of the Luisterlinie app, during which people could ask questions. Multiple technical questions were answered by the developers and the team leader. Time was reserved after the demo for the pilot participants to log into the app and attempt to record the team meetings they were attending. This process went relatively smoothly. Those who struggled could share their screens to receive assistance from us, which also allowed us to help others. After this, the consent form was handed to the participants to fill out via a PDF in the first kick-off meeting. This was found to be a struggle, so future kick-off meetings switched to integrating the consent form directly after the kick-off survey. This proved to be much easier and less time-consuming for participants. Following this, the participants were informed that the pilot would run for a month and were encouraged to use the app as much as possible. A participant asked if it was okay to try the app during their free time. This was encouraged and explicitly mentioned to be allowed and encouraged by the team leader in the subsequent kick-off meetings. Another participant asked what types of filters the AI model employs. This question arose after we mentioned that the model might not generate a summary if the transcript contains vulgar, sexually explicit, or harmful text. An example was provided, along with some solutions for instances when the model's filters are triggered, such as removing or rewriting inappropriate language from the text.

6.1.2 After the kick-off meeting

The days following the kick-off meetings went smoothly. Participants could use the established Teams channel for additional assistance and received a reference guide detailing the most important functionalities. This guide included information about the various features of the app and the locations of specific buttons or functions. It also provided instructions on how to add the app to the home screen of a phone. An example of the reference guide can be found in Figure 6.2. As shown in Table 6.1, the app was used by 52 people.



Luister Linie toevoegen als app op Android

Welkom!

Begin met het maken van een nieuw gesprek. Dit kan door een gesprek op te nemen of door een notitie te dicteren/typen.

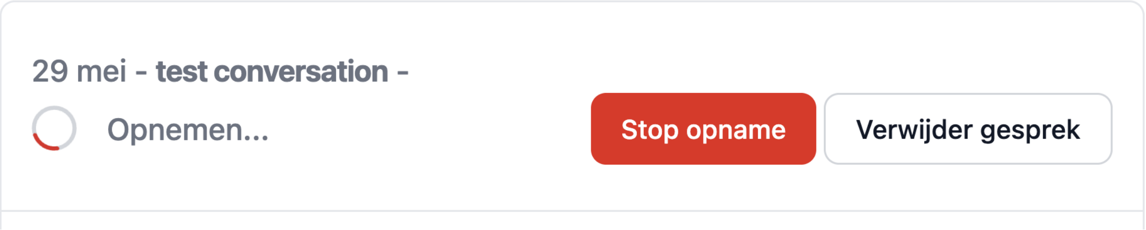
Touch & hold the widget to move it around the home screen

Figure 6.2: Page from the reference guide showing how to install the Luisterlinie app to your phone

6.2 Week 2: bug fixes (Apr 21st - 27)

The second week included the final kick-off meeting, which followed the same format outlined in Table 5.1. This meeting proceeded without any issues. At the beginning of week 2, some participants mentioned in the Teams channel that the app had been removed from their work phones. An initial investigation revealed that the application maintenance department had not whitelisted the app, resulting in its automatic deletion. This deletion led to some recordings not processing correctly or being cut off because the app was removed during a conversation. Consequently, these conversations remained in a recording state with no straightforward way to exit that state.

Initially, we resolved this issue by entering a special command in the browser console, which often restored the conversation and recording. Once we discovered that this was a widespread problem, we implemented a fix that added a button to the conversation whenever the recording state was detected, as shown in Figure 6.3. This feature allowed participants to manually stop the conversation if the recording stalled for any reason. By the end of the week, all these issues were resolved. As shown in Table 6.1, the app was used by 44 people.



29 mei - **test conversation** -

Opnemen...

Figure 6.3: A conversation that is stuck in the "recording" state with a button besides it that stops the recording

Some participants provided feedback regarding the app icon, which featured the Salvation Army logo. They noted that this logo was also used for another app, leading to potential confusion. To address this issue, the app logo was changed to a new design that was more distinct from the previous one.

Five people couldn't make it to the kick-offs, but still wanted to participate in the pilot. These people were given the presentation and documents from the pilot personally and got set up in week 2.

Week 2 was originally scheduled for the distribution of the second survey. However, this was postponed by a week due to the Easter weekend, as most people had not yet had the opportunity to use the Luisterlinie app. This was evident in the usage dashboard, as shown in Figure 6.4. Additionally, the decision was made to extend the pilot by a week due to Easter weekend and the May holiday.

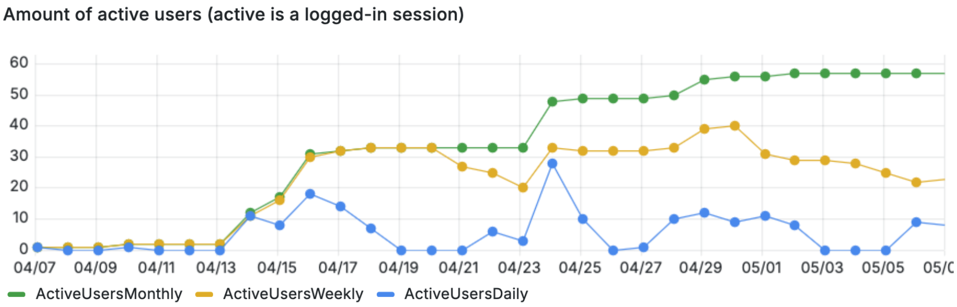


Figure 6.4: Part of the usage dashboard showing the daily, weekly and monthly active users

6.3 Week 3: second survey (Apr 28th - May 4)

The third week was spent distributing the week 1 survey, which can be found in Appendix C. The primary goal of this survey was to assess initial usage and identify any issues that could be addressed before the pilot's conclusion. During the kick-off meetings, a list of participants was maintained and used to send the survey via email. Surveys were distributed exactly two weeks after the kick-off session to ensure that those who started using the Luisterlinie app later had equal usage time. A reminder email was scheduled for one week later. Initially, the response to the survey was slow, but it began to increase after the reminder emails were sent. The team leader's administration also called some individuals who had not responded to the survey to confirm their participation in the pilot. Survey results can be found in Section 7.2. App usage was strong, with approximately 44 users, the same as in week 2.

6.4 Week 4, 5 & 6: steady usage (May 5th - 25)

Weeks 4, 5 and 6 were relatively uneventful. In week 4, app usage decreased to 37 people, as shown in as shown in Table 6.1. This decrease in app usage can be explained by the national two-week school holiday in the Netherlands known as "Mei vakantie." Participants communicated that they would be using the app less during this time because they were on holiday, which explains this dip in usage. After the May holiday, the app usage increased slightly, and major issues, such as the app missing from participants' phones, were resolved. Participants discovered that the microphone permission was sometimes revoked on their phones, prompting a message in the team chat to address this issue. After the 11th of May, week 5, a spike in usage was observed as more conversations were recorded and more users used the app. This increase likely occurred because the 11th of May marks the end of a national two-week school holiday.

6.5 Week 7: end-of-pilot meetings (May 26th - Jun 1)

This week, the end-of-pilot survey was sent out to all participants. The survey can be found in Appendix D. This survey was mostly meant to determine the long term effect of the Luisterlinie app on the work of the youth-care workers. The survey was sent out in one go, since the end-of-pilot meeting were also done in two big groups, instead of the kick-off groups. This approach was chosen due to the fact that some people from the kick-off groups could attend on one day but not the other. The app's usage also decreased, primarily because the pilot period concluded.

6.5.1 End-of-pilot meetings

The end of pilot meeting, as detailed in Section 5.7, was an opportunity to gather feedback and answer questions from the participants. Two meetings were held in week 7, to give participants the opportunity to attend one of them and to make it easier to avoid scheduling conflicts. The end of pilot meetings resulted in a ton of valuable input, in addition to the feedback given in the surveys and throughout the pilot. Details and results from the meeting can be found in 7.5. These meetings also marked the last step in the pilot.

Chapter 7

Evaluation

This section will evaluate the impact of the "Luisterlinie" GenAI tool on documentation capabilities, using data from the surveys and logs generated during the action-taking phase. A total of three surveys were conducted, and logs were collected throughout the entire duration of the pilot. A total of 47 active participants participated in the pilot. The pilot ran from April 14th till June 1ste.

7.1 Profile of participants and kick-off survey

As described in the action taking chapter, a total of 4 kick-offs were conducted. The participation in these kick-offs can be seen in table 7.1.

Table 7.1: Kickoff Survey Participants

Kickoff	Participants
Kickoff 1 (April 14th)	11
Kickoff 2 (April 15th)	3
Kickoff 3 (April 16th)	13
Kickoff 4 (April 24th)	15
Other	5
Total	47

As you can see, an "other" row has been added. This is to account for five people that couldn't make it to the kick-offs, but still wanted to participate in the pilot. These people were given the presentation and documents from the pilot personally. All participants work as youth care workers.

The results of the kick-off survey can be seen in table 7.2 and 7.3. The kick-off survey itself can be found in appendix B.

Table 7.2: Descriptive Statistics for Q4 and Q5

	Q4	Q5
mean	45.62	18.52
std	17.25	11.59
min	10.00	2.00
max	80.00	60.00

As shown in table 7.2, individuals report spending an average of 45.6% of their time on documentation, with a maximum of 80% and a minimum of 10%. Participants report spending an average of 18 minutes on each client conversation. The standard deviation is notably high in both questions, indicating a significant spread of percentages among individuals. This aligns with the diagnostic interviews, where participants reported having about 10 to 15 conversations each day. This suggests that conversations occupy approximately 42.8% to 64.2% of their time during an 8-hour workday, excluding a one-hour break, which aligns with the survey results.

Table 7.3: Evaluation of the survey responses by kickoff session (1=strongly disagree and 5= strongly agree)

	KO 1 (n=11)		KO 2 (n=3)		KO 3 (n=13)		KO 4 (n=15)		KO 5 (n=5)		Total (n=48)	
Question	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2	$\bar{x}$	σ^2
WIJZ Time Burden (Q1)	3.3	1.4	5.0	0.0	4.2	0.8	3.5	0.8	4.0	0.0	3.8	1.0
In-Meeting Note-taking (Q2)	3.6	0.7	3.3	2.3	3.3	1.4	3.0	1.6	4.2	0.7	3.4	1.3
Timely Admin Completion (Q3)	3.3	0.8	2.3	0.3	2.4	1.3	2.8	1.3	2.8	1.7	2.8	1.2
Per-Contact Journaling (Q6)	3.7	1.0	3.7	0.3	3.6	1.1	3.2	0.9	4.0	0.0	3.6	0.8
WIJZ Data Completeness (Q7)	3.3	0.8	4.0	1.0	3.4	0.8	3.2	0.7	3.6	0.3	3.3	0.7
WIJZ Missing Data (Q8)	3.1	0.5	2.0	0.0	2.9	0.9	3.1	0.6	3.2	0.7	3.0	0.7
WIJZ Recording Ease (Q9)	3.5	0.5	2.7	0.3	3.2	0.8	2.8	1.0	3.0	0.5	3.1	0.8
WIJZ Document Quality (Q10)	3.5	0.3	4.0	0.0	3.2	0.5	3.0	0.6	3.8	0.2	3.3	0.5
Administration Ease (Q11)	3.5	0.5	3.0	1.0	3.2	0.9	3.2	0.6	3.2	0.7	3.2	0.6
Administration Effort (Q12)	2.5	0.5	2.7	0.3	3.2	1.4	2.8	1.0	2.8	1.2	2.8	1.0

7.1.1 Time spend

According to the results shown in table 7.3, participants tend to agree they spend more time than desired on administration (Q1). Participants also lean towards agreeing they are busy making notes during client meetings (Q2). Participants do slightly disagree that they can complete administration on time. This indicates that timeliness is an issue (Q3). This data suggests that time is a concern. Participants generally feel they spend too much time on administration (Q1) and struggle to complete it on time (Q3). They also report being busy with note-taking during sessions (Q2).

7.1.2 Quality and completeness

In the quality and completeness section, participants generally agree that they make contact journals for each client contact (Q6). They slightly agree that administration is generally complete (Q7) but are, however, undecided or have mixed experiences regarding missing data in WIJZ (Q8). Participants are neutral about the difficulty of administrative work, finding it neither particularly easy nor difficult (Q9). They report a slightly positive sentiment towards the overall quality of the documentation that is currently stored in WIJZ (Q10). To summarize, participants tend to follow procedures, but perceptions are mixed on the documentation quality and ease. There's no strong consensus on whether important data is often missing, though KO2 felt strongly that data was not missing.

7.1.3 Documentation satisfaction

Participants agree slightly with Q11, suggesting that administering is an easy task. This is slightly in contrast with Q9. There is a slight disagreement that administering costs a lot of effort (Q12). This does align with Q11.

7.1.4 Variability

Most variances from the kick-off survey are around 0.5 to 1.4, suggesting a moderate spread of opinions. Low variances (e.g., 0.0, 0.2, 0.3) in KO2 and KO5 are largely due to small sample sizes and high agreement within those specific small groups. These groups showed more extreme views on certain aspects, but this can also be explained by the small number of participants in these groups.

7.1.5 Opportunities and concerns

Participants identified numerous opportunities for using AI in documentation. Their responses were divided into five categories, as shown in Figure 7.1. It is important to note that some responses fit into multiple categories, so the total number of responses across all categories will not equal the total number of survey participants.

Participants see a significant opportunity in AI's potential to save time and enhance efficiency. They commonly point out that the documentation process will become easier and faster, reducing the likelihood of work piling up. Another opportunity is the improved quality and completeness of documentation,

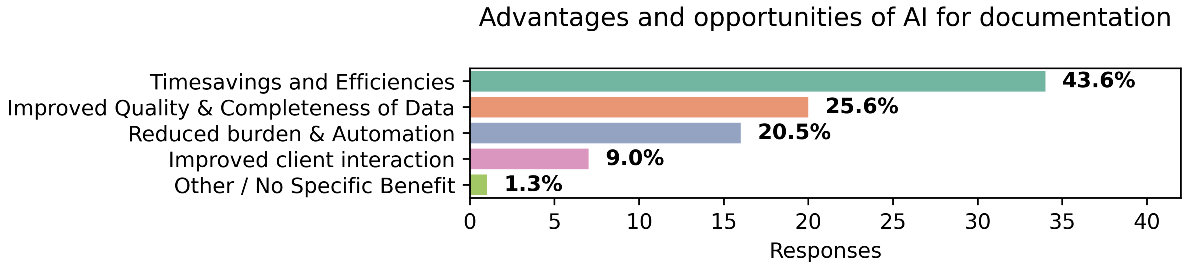


Figure 7.1: Q13: Advantages and opportunities for AI

which makes it easier to review previous discussions. Additionally, they believe AI will reduce the administrative burden through automation. Lastly, they recognize a substantial chance for more personalized client interactions.

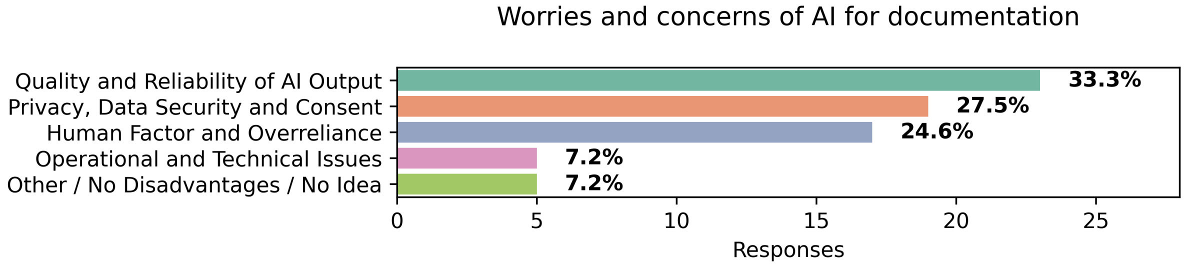


Figure 7.2: Q14: Worries and concerns for AI

Participants have expressed concerns about the use of AI, as illustrated in Figure 7.2. The quality and reliability of AI outputs are significant issues. Additionally, there are worries about privacy and security implications, as well as the need to obtain consent from clients before recording. Participants also fear the drawbacks of excessive reliance on AI, which include diminished human control and critical thinking, a loss of personal interaction and human touch, standardization, and potential impacts on human skills and alertness. Furthermore, they are concerned that AI may actually consume more time instead of saving time due to technical difficulties and operational challenges.

7.2 Week 1 survey

The Week 1 survey was scheduled to be sent out one week after each kick-off. However, due to Easter weekend, the survey was intentionally delayed by an additional week to give participants more time to use the app. The surveys were sent out on April 28th, 29th, and 30th, as well as on May 9th. A total of 28 responses were received, with 22 respondents having actually used the Luisterlinie app since the kick-off (Q1). The 6 other respondents gave a few reasons why they didn't use the app (Q2): 3 people were on holiday, one didn't need the Luisterlinie app, one didn't think to use the app, and one didn't use the app due to language issues, where the conversations were in two different languages.

Descriptive statistics, including mean ($\bar{x}$), variance (σ^2), and the number of respondents (N) for each relevant question, are detailed in Table 7.4. Initially, 22 participants who reported using the app (Q1, not shown in table) responded to the first set of questions. The number of respondents slightly decreased for later questions, as indicated by N values in the table. A mean around 3 is neutral, means above 3.5 indicate general agreement, and means below 2.5 indicate general disagreement. A lower variance (lower than 0.5) indicates strong concensus, whereas higher variance (higher then 1.0) indicates more diverse opinions.

7.2.1 Time spend

Participants' perceptions of the time spent on administrative tasks after using the Luisterlinie app were generally positive. On average, they felt they spent less time registering contacts in WIJZ (Q3).

Question	Mean ($\bar{x}$)	Variance (σ^2)	N
Registration time (Q3)	3.65	0.77	20
Reporting efficiency (Q4)	3.65	0.98	20
Double work? (Q5)	2.75	0.83	20
Less note-taking? (Q6)	3.90	0.94	20
More journals? (Q7)	3.15	1.29	20
Summary completeness (Q8)	3.37	0.91	19
Add missing info? (Q9)	2.84	0.92	19
Report completeness vs human (Q10)	3.53	0.93	19
Report quality vs human (Q11)	3.74	0.43	19
Reporting enjoyment (Q13)	3.72	1.04	18
Admin burden reduced? (Q14)	3.56	0.73	18
Regret non-use? (Q15)	3.83	1.09	18
App pleasantness (Q16)	3.78	0.54	18
Recommend app? (Q17)	4.33	0.35	18
Continue use? (Q18)	4.67	0.35	18

Table 7.4: Week 1 Statistics for Likert Scale Questions

Additionally, participants agreed that their administration in WIJZ became more efficient after using the Luisterlinie app (Q4). When asked if the app led to significant double work, participants generally disagreed (Q5). There was strong agreement that using the Luisterlinie app reduced the need for note-taking during client conversations (Q6). However, participants were neutral when asked if they created more contact journals than before using the Luisterlinie app (Q7).

7.2.2 Quality and completeness

The survey also assessed perceptions of the quality and completeness of reports generated by the Luisterlinie app. When asked if the summaries contained all important aspects of a conversation (Q8), the responses indicate a neutral to slightly positive perception of completeness. Regarding the need to add important missing information to make the app-generated reports more accurate (Q9), participants slightly disagreed with the statement that they *often* had to add missing information. This implies that frequent additions were not a big issue. Participants generally agreed that the completeness of reporting generated by the Luisterlinie app was comparable to human-generated reporting (Q10). Similarly, they agreed that the quality of reporting was also comparable to human-generated reporting (Q11). The low variance for Q11 suggests a high level of consensus on the quality of reporting.

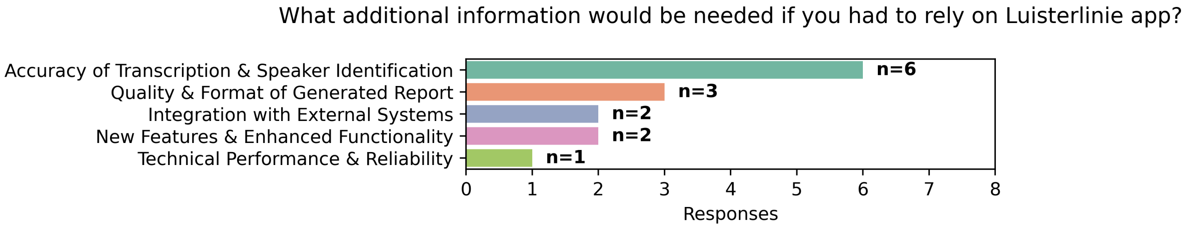


Figure 7.3: What additional information or changes would be absolutely necessary if you had to rely solely on the reports generated by the Luisterlinie app? Why? (Q12)

Figure 7.3 displays the results of Q12, with a total of 14 participants providing feedback. Six participants indicated that most issues arose from the application mishearing inputs, leading to incorrect transcripts. Three participants mentioned problems with the quality and format of the generated report. These issues included ethical blocks, where the model refrained from responding due to ethical guardrails, or used incorrect terminology. Two individuals requested a more seamless integration of Luisterlinie with WIJZ. Additionally, two participants suggested incorporating features such as phone call recording and real-time translation. One participant reported stability issues with the app, specifically that recordings occasionally failed.

7.2.3 User satisfaction

User satisfaction with the Luisterlinie app was generally high. Participants found reporting more enjoyable when using the app compared to their previous experiences (Q13). They also felt that the administrative burden had decreased since they began using the app (Q14). Additionally, they expressed regret at the thought of not being able to use the Luisterlinie app during conversations (Q15). The app was perceived as pleasant to use, with relatively low variance, indicating strong consensus among users (Q16). The intention to recommend the app to colleagues for administrative tasks was very strong, also showing low variance and high agreement (Q17). Participants expressed an extremely strong desire to continue using the app after the pilot phase, with similarly low variance (Q18). This was the highest-rated item, demonstrating a strong and consistent positive sentiment toward continued use.

The responses to Q19 (How many times did you want to use the Luisterlinie app?) were varied. Some participants reported an amount of times they wanted to use the app, 3.81 times on average. Four participants indicated they wanted to use the app with every conversation. In contrast, three other participants mentioned they did not have enough conversations, only used the app when it seemed useful and when they forgot to turn it on.

Table 7.5: Descriptive Statistics for Q20, Q21, and Q22

	mean	std	min	max	count
Q20: Could use Luisterlinie app (%)	62.83	31.54	0.00	100.00	18.00
Q21: Asked for permission (%)	73.11	40.43	0.00	100.00	18.00
Q22: Permission granted (%)	71.72	36.62	0.00	100.00	18.00

The results for questions 20, 21, and 22 are shown in Table 7.5. When asked what percentage of the time participants could use the app when they wanted to, the response was approximately 62% (Q20). In those instances, they requested permission 73% of the time (Q21) and received it 71% of the time (Q22).

7.2.4 Benefits and challenges

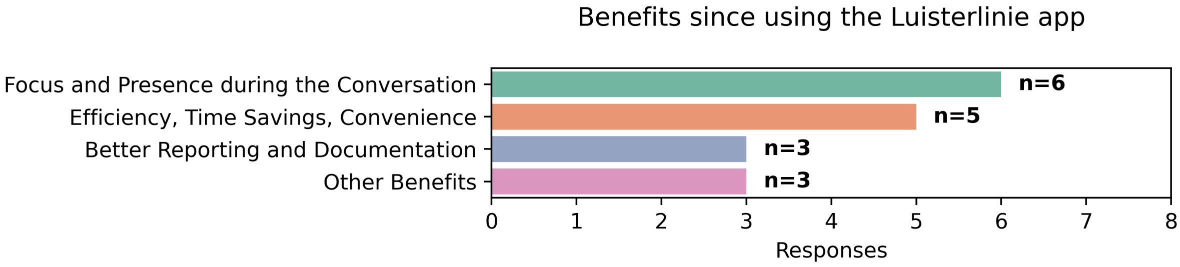


Figure 7.4: Q23: Biggest benefits when using the Luisterlinie app

The benefits reported by users of the Luisterlinie app can be categorized into four groups as shown in Figure 7.4. The largest group, consisting of six people, indicated that their main advantages were increased focus and improved presence during conversations. Following this, five participants mentioned time savings, efficiency, and convenience as key benefits of using the app. Additionally, three participants noted enhancements in documentation and reporting quality. Finally, three individuals provided feedback that did not fit into the previous categories. They mentioned features such as translation, using the app as a backup while taking notes, and utilizing it afterward for a quick summary of the conversation.

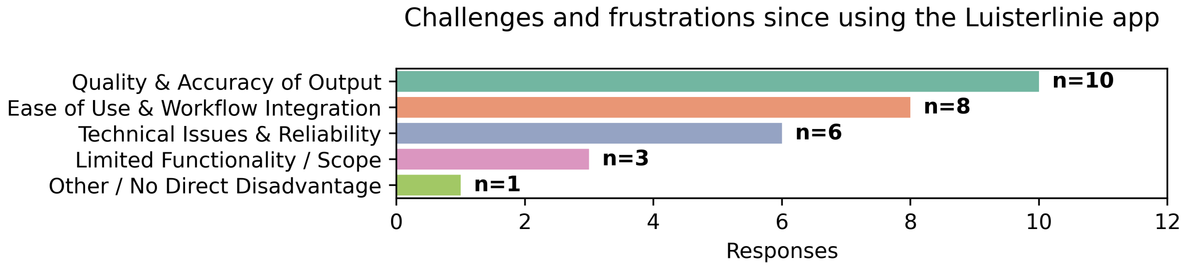


Figure 7.5: Q24: Biggest challenges and frustrations when using the Luisterlinie app

The disadvantages reported by participants can be categorized into five groups, as shown in Figure 7.5. The most significant frustration, noted by 10 individuals, was the quality and output of the app. Complaints included inaccurate transcription of words, incorrect interpretation of texts, and undesirable phrases, such as improper tense or factual inconsistencies. The second major frustration, mentioned by 8 participants, was related to ease of use and integration into their workflow. This encompassed the need to ask for permission during conversations and the time-consuming process of checking transcripts and summaries. The third frustration, identified by 6 people, involved technical issues and reliability. This included bugs in the Luisterlinie app, such as missing recordings and audio quality problems. Additionally, 3 participants expressed disappointment over missing features, such as the lack of translation support and the inability to transcribe specific conversations, like certain meetings. Lastly, one participant stated that they recommend the app to their colleagues and do not perceive any disadvantages.

7.2.5 Stability and usability

Approximately 44.44% of respondents encountered issues with the Luisterlinie app (Q25). These problems stemmed from bugs present in the initial weeks of the app, most of which were later resolved. Four participants reported that transcriptions failed after recording conversations. Three participants noted that the app was removed from their phones. Additionally, one participant mentioned that the microphone did not adequately capture voices in the room (Q26).

Regarding the other feedback mentioned in Q28, one participant expressed her disappointment about not being able to use the app after the pilot. Another participant suggested that she would like to see increased support for meetings and various types of conversations.

7.3 Week 4 survey

The week 4 survey was scheduled to be sent out four weeks after the pilot began. This was intentionally delayed to provide participants with additional time to use the app. The survey was distributed to all groups simultaneously on May 22nd. A total of 23 responses were received, with 19 participants having actually used the app. Some participants partially completed the survey. The reasons four individuals did not use the app were as follows: one used a colleague's laptop, two lacked the time due to being too busy, and one switched to a different role. People that did use the app had on average 10 conversations where they used it. However, this varied wildly between one and forty conversations.

Question	Mean ($\bar{x}$)	Variance (σ^2)	N
Registration time saved (Q5)	3.50	0.74	18
Documentation efficiency (Q6)	3.83	0.74	18
Review time saved (Q7)	3.83	0.97	18
Client attention (Q8)	4.00	1.06	18
Summary completeness (Q9)	3.71	0.35	17
Fewer corrections needed (Q10)	2.94	0.56	17
Report completeness (Q11)	3.76	0.44	17
Report quality (Q12)	3.41	0.76	17
Documentation satisfaction (Q14)	3.88	0.36	17
Value in routine (Q15)	3.94	0.68	17
Admin load reduced (Q16)	3.71	0.60	17
Continue app use (Q17)	4.65	0.24	17
Ease of use (Q18)	4.41	0.26	17
Recommend to colleagues (Q19)	4.71	0.22	17
App stability (Q24)	4.06	0.68	17
Pilot support clarity (Q25)	4.12	0.99	17
Instruction clarity (Q26)	4.47	0.26	17

Table 7.6: Week 4 Statistics for Likert Scale Questions

Table 7.6 shows the mean, variance and amount of answers for the Likert scale questions from the survey.

7.3.1 Time spend

Participants generally agreed that the app saved them time when registering contacts (Q5). They expressed even stronger agreement that the app made their documentation process more efficient (Q6). Additionally, there was significant consensus that reviewing and adjusting app summaries was quicker than manual writing (Q7). Participants strongly agreed that the app enabled them to focus more on clients by reducing the amount of note-taking required (Q8).

7.3.2 Quality and completeness

Users generally agreed that summaries captured important aspects (Q9). However, the response was quite neutral when asked if needed to make fewer corrections as they used the app longer (Q10). This suggests the app's output might not have improved significantly with user experience, or that initial corrections remained necessary. The variance (0.56) is moderate, indicating some spread, but generally clustered around neutral, which is notable. Participants agreed that app-generated reports were comparably or more complete than manual ones after 4 weeks (Q11). The app-generated report quality was comparable or better, though this agreement is slightly less strong than for completeness (Q12). This might indicate that the quality of the output of the app is less than completeness.

As shown in Figure 7.6, speaker assignment and role clarification is still the biggest change that participants needed to make when using the Luisterlinie app. This ties in with the results from week 1, where speaker identification and accuracy of transcription was also the biggest issue. Accuracy of the content and facts, completeness and style, structure and wording follow, which can also be classified as quality of the report. Participants also gave feedback about the process around the pilot and the integration of the Luisterlinie app in their workflow.

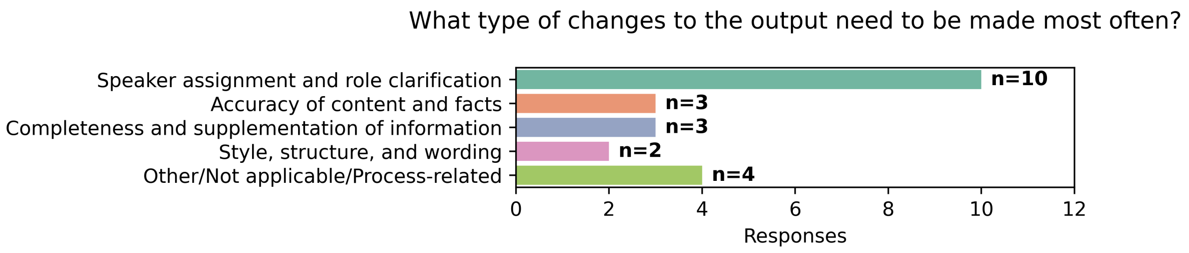


Figure 7.6: Looking back over the past four weeks, what types of information or adjustments were most often needed in the text generated by Luisterlinie? (Q13)

7.3.3 User satisfaction

User satisfaction with the Luisterlinie app was again notably high. Participants strongly agreed that their overall satisfaction with the documentation process had improved since using the app (Q14) and that the app had become a valuable addition to their daily work routine (Q15). There was also agreement that the app helped reduce their administrative load (Q16).

Intentions for future use and advocacy were exceptionally strong. Participants expressed a very strong desire to continue using the Luisterlinie app after the pilot (Q17). They also strongly agreed that the app was easy to use after an initial period of familiarization (Q18). The highest level of agreement was again observed for recommending the app to colleagues performing similar work (Q19). The low variances for Q17, Q18, and Q19 indicate a high degree of consensus among users on these positive aspects.

7.3.4 Benefits & challenges

Participants were again asked about the advantages and challenges they experienced when using the Luisterlinie app.

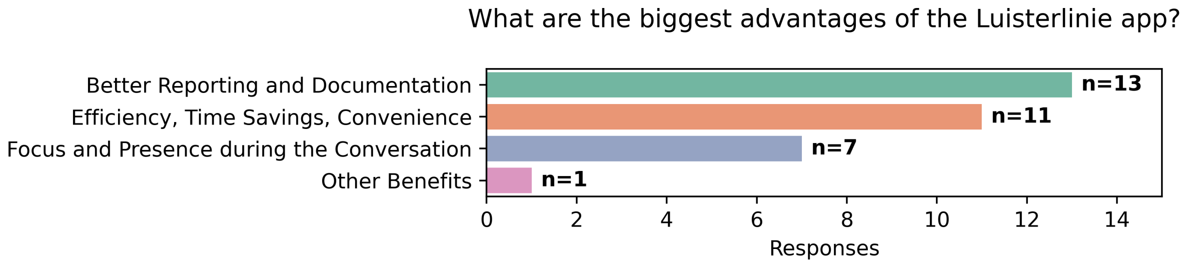


Figure 7.7: Biggest advantages of the Luisterlinie app (Q22)

As shown in Figure 7.7, the most significant advantage reported by participants was improved reporting and documentation. They noted a more uniform quality, as well as more detailed and well-organized records. A close second was the increased efficiency, time savings, and convenience. Participants also experienced greater focus and presence during conversations.

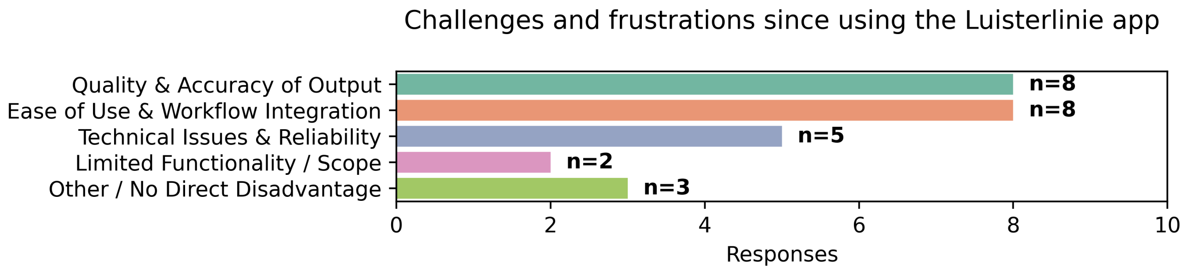


Figure 7.8: Biggest challenges and frustrations using the Luisterlinie app (Q23)

Participants identified quality and accuracy, as well as ease of use and workflow integration, as the biggest challenges or frustrations, as shown in Figure 7.8. Quality issues included problems with inaccurate transcriptions, where words were misheard, and summaries that failed to accurately reflect conversations. Integration and workflow challenges involved frustrations related to obtaining permission and the necessity of verifying the output. Additionally, some participants reported technical issues, such as the need for a reliable Wi-Fi connection and the app's inconsistent performance at the beginning of the pilot. Lastly, two participants expressed a desire for additional functionality. They requested a ChatGPT-like interface that could handle more than just conversations and accommodate other types of interactions.

7.3.5 App stability

The technical aspects of the app and the support provided during the pilot were well-received. Participants strongly agreed that the Luisterlinie app was generally stable and reliable in use over the four weeks (Q24). They also indicated strong agreement that it was clear where they could direct questions about the pilot (Q25) and that the instructions provided during the kick-off meeting gave them sufficient clarity to begin using the app (Q26).

Two participants reported experiencing problems while using the Luisterlinie app (Q27). Both encountered technical issues where their speech was recognized, but the conversation was not recorded. One participant also noted that the app sometimes struggles to extract information from lengthy conversations (Q28).

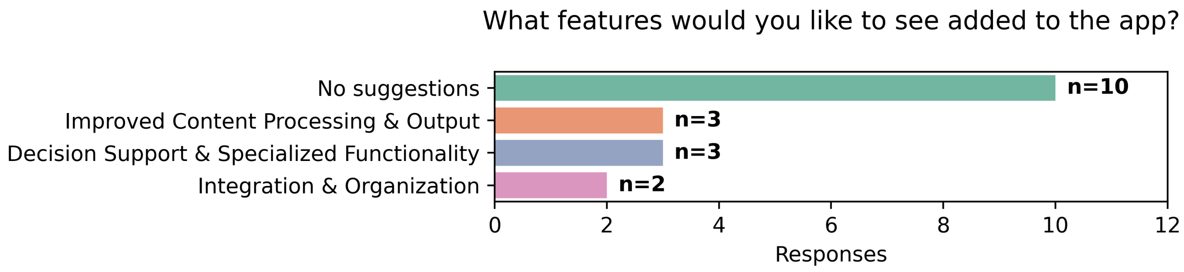


Figure 7.9: What features would you like to see added to the app? (Q30)

Most participants are satisfied with the app as it is and did not request additional features when asked (Q30), as shown in Figure 7.9. Three individuals expressed a desire for improvements in the processing of summaries and outputs. Additionally, three participants requested specific functionalities tailored to their roles, such as particular conversation types. Two others would like to see better integration of the Luisterlinie app into the organization and their workflow.

When asked for additional feedback (Q31), one participant requested custom conversation types instead of the pre-selected options currently available in the app. Another participant suggested that consent for recording should be automatically granted. Additionally, one participant mentioned that the pilot was relatively short and expressed a desire for more time to try the Luisterlinie app.

7.4 Luisterlinie logs

The logs of the Luisterlinie app provide insight into the actual usage patterns of the participants. We regularly monitored these logs to identify errors and track app usage.

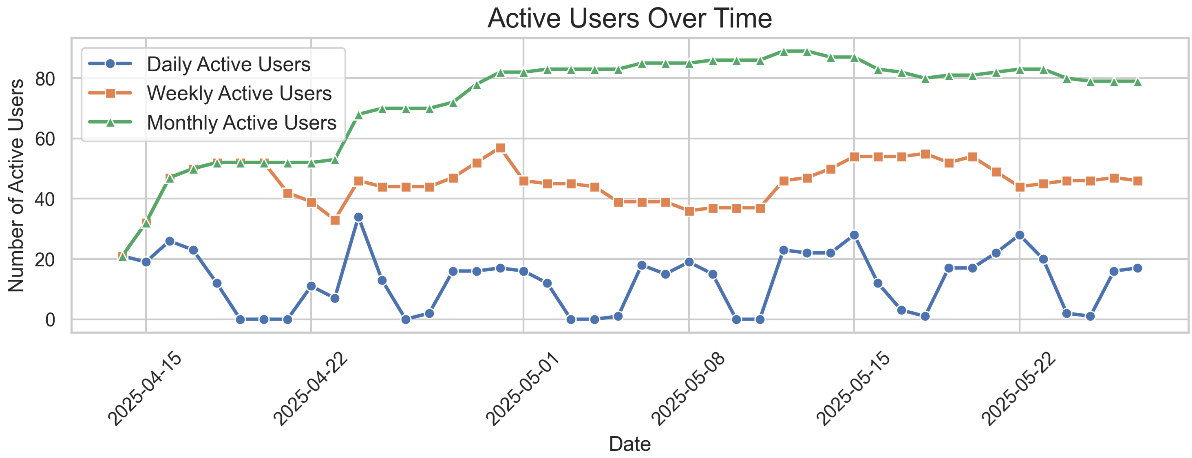
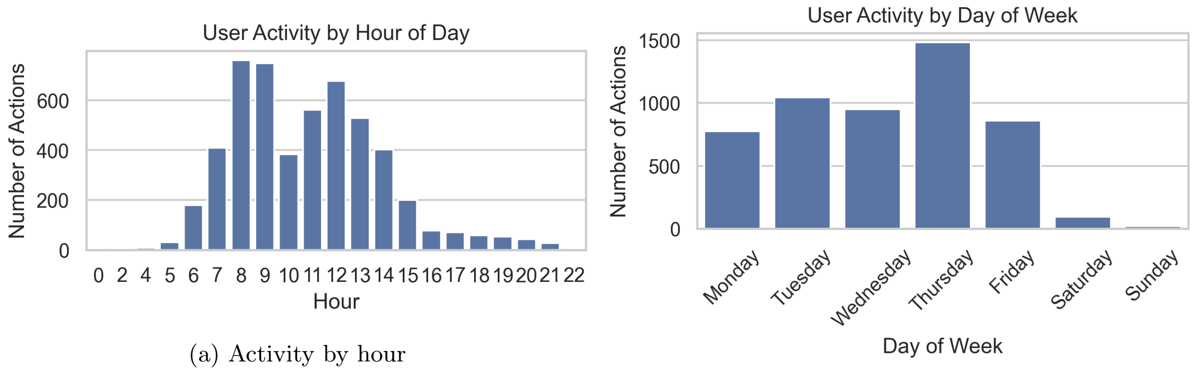


Figure 7.10: Daily, weekly and monthly active users for the luisterlinie app during the pilot period

Figure 7.10 illustrates the daily, weekly, and monthly active users since the pilot began. It shows a steady increase in monthly active users. The number of weekly active users plateaued shortly after the kickoffs. Additionally, the impact of the May holiday is evident, as there is a spike in weekly active users after the holiday concludes.



(a) Activity by hour

(b) Activity by week

Figure 7.11: A figure with two subfigures

Usage spikes occur between 8:00-9:00 and 11:00-13:00, as shown in Figure 7.11. The app is most popular on Thursday and least popular on Sunday. This trend also accounts for the lower number of users during the weekend, which can be seen in Figure 7.10.

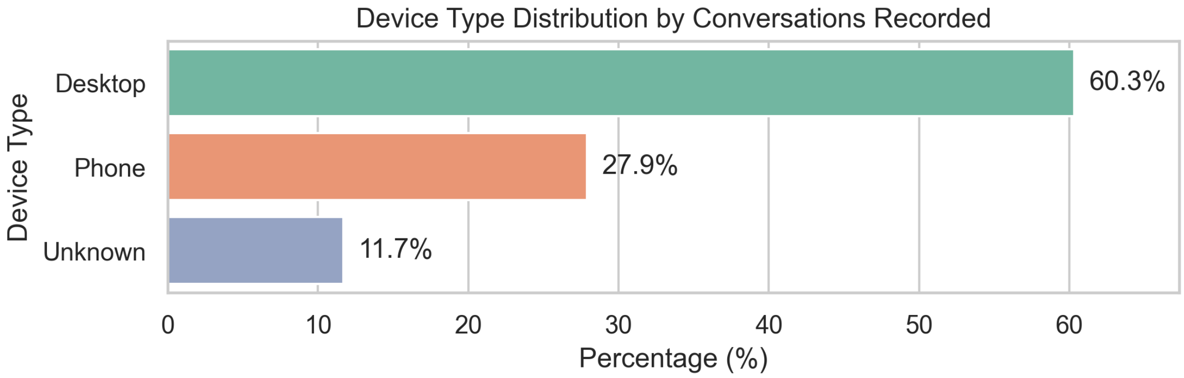


Figure 7.12: Device usage per conversations (%)

As you can see in Figure 7.12, most conversations were recorded on desktop. Only 27.9% of conversations were recorded on a phone, which was quite surprising, since high phone use was expected.

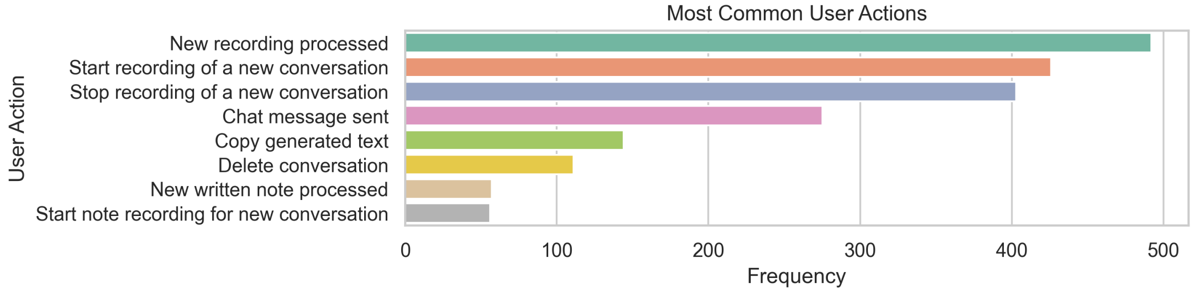


Figure 7.13: Top user actions

In Figure 7.13, the most common user actions are displayed. The primary actions include starting (426 times), stopping (403 times), and processing conversations (492 times). The next most frequent action is sending chat messages to the integrated chatbot, with 275 messages sent. Users also frequently copied the generated text (144 times), suggesting it was valuable for use in WIJZ or other contexts. Additionally, conversations were often deleted (111 times).

7.5 End-of-pilot meeting

The end-of-pilot meeting gave valuable insight into the experience of participants with the Luisterlinie app and the pilot. It also presented an opportunity to gather feedback in an open conversation. The meeting started with a quick welcome and reintroduction of the team. It was also explained that the purpose of the meeting was to gather feedback and answer any questions that might come up. It was also mentioned that the pilot team was curious for what improvements the participants would like. The meeting was started by asking, "If you were Hugo, our director, and had to decide at the end of the pilot, based on your experience, what would you do as a follow-up?" Most participants answered that they would invest further in Luisterlinie due to its benefits in reducing administrative load and its overall potential. Some participants urged immediate wider rollout, and dealt with "teething problems" later, others felt that the app required more polish and development before a wider rollout, particularly regarding the reliability. Participants also noted that summaries should be more reliable and comprehensive for a wider rollout. Participants were also asked what made them decide to use the app, and how they made that consideration. They mentioned that they were still integrating the app into their routine. Others pointed out a practical barrier, being the fact that phone calls can't be recorded from only a phone. Recording a phone call involved putting the phone in speaker mode and recording the conversation with their laptop. This became an obstacle when trying to use the Luisterlinie app in the office, as participants didn't want to disturb others that were working. Participants were also asked what would help them integrate the Luisterlinie app more in their work. One participant suggested training, while others suggested physical, hands-on demonstrations. Participants also mentioned that the app should stay a supportive tool and not a mandatory part of their workflow. When asked about future worries, participants mentioned lack of precision when checking AI output, leading to errors. Others worried about the step of obtaining client consent being overlooked for convenience. After the feedback had subsided, participants were made aware they were allowed to keep using the app, as the team looked at a wider rollout. Participants were excited that they could keep using the app. Participants were also invited to give additional feedback after the pilot if they encountered any issues.

Chapter 8

Reflecting

The aim of the reflection section is to critically analyse the entire action research process, the findings, and their implications. This research sought to determine how the integration of GenAI can enhance record-keeping for interviews and meetings in public healthcare organizations. The evaluation focused on four key metrics: time spent on administration, the quality and completeness of records, job satisfaction, and the potential benefits and challenges of incorporating such an AI application into existing workflows. According to these metrics, a pilot was designed that would gather data for the metrics through surveys and logs, as well as feedback. This pilot was then carried out from April 14th till June 1st. A total of 47 active participants participated in the pilot. The data gathered in the pilot was analysed and showed promising results. Participants reported similar quality, reduced note-taking during conversations, and increased efficiency along with time savings.

8.1 Key findings & discussion

This section summarizes the core findings from the evaluation phase and addresses the study's research questions. Both the effect of GenAI on the transcription capability and the documentation capability will be addressed. It will also address why some results might have occurred.

8.1.1 Enhanced transcription capabilities

Transcriptions are a key part of the Luisterlinie app. They provide the input for the summaries that will be used by youth care workers. Before the Luisterlinie app, no transcripts were made. The quality of the transcripts were mixed. The week 1 survey results (Section 7.2) showed that "quality & accuracy of output" was the most significant frustration (Fig 7.5), with specific mentions of inaccurate transcription of words and misinterpretation of inputs. There were also some reliability issues in the Luisterlinie app related to the transcript, though many of these were resolved during the pilot.

While Luisterlinie app offered a new way to capture conversational data, the perceived and actual accuracy of the transcripts was a significant concern for pilot participants. The transcription capability was present, but its reliability and quality need improvements to be considered a good enhancement. This result aligns with the findings of the study by Kumar [23], which also identified reliability as an issue in real-world contexts. The primary value of the transcripts is the input for subsequent summarization and analysis, rather than a standalone perfect transcript. However, if the quality of the transcript is lacking, the quality of the summarization and analysis will also suffer.

The lack of quality and accuracy in the transcripts can stem from several factors. First, the Whisper model used is not specifically trained for the Dutch language or for difficult-to-hear audio, which may limit its effectiveness. A different model could enhance performance. Another significant factor affecting accuracy and quality is the method of audio recording. The conversations were primarily recorded on laptops, which typically have directional microphones that are not effective at capturing sound from all directions, like in a physical meeting or when sitting opposite to each other. Lastly, the quality and accuracy issues may also arise because some clients of youth protection do not speak fluent Dutch or English, making it challenging for a transcription model like Whisper to accurately understand the dialogue.

8.1.2 Enhancement of documentation capability

Enhancements in documentation were measured according to four metrics: time spent on administration, the quality and completeness of records, job satisfaction, and the potential benefits and challenges of incorporating such an AI application into existing workflows. These key metrics give a full picture of how documentation can be enhanced using GenAI in public healthcare. The Luisterlinie app showed significant and positive perceived impact here.

Reduced documentation time. Participants initially reported spending a significant portion of their time on documentation, averaging 45.62% of their time and 18 minutes per conversation on this task (Section 7.1). After using the Luisterlinie app, users reported spending less time registering contacts in WIJZ (Section 7.2 and 7.3) and found their reporting more efficient. This is in line with existing research, where it was found that automatic note generation can significantly save doctors time [3].

Increased presence during conversations. A key contributor was the reduced need for extensive note-taking during conversations, which was also highlighted as a major benefit in allowing better focus and presence (Fig 7.4). The reduced need for note-taking can be attributed to the ability to read back the transcript instead of having to write things down. This enhancement in attention and presence with the client was not significantly reflected in earlier research. Further studies may be necessary to determine the long-term effects of this increased presence on both clients and healthcare workers.

Mixed quality and completeness. While administration in WIJZ was generally considered complete (Section 7.1), there were neutral or mixed experiences regarding missing data. Summaries generated by Luisterlinie were generally perceived to contain important aspects of conversations (Section 7.2 and 7.3). Crucially, participants felt the completeness and quality of Luisterlinie assisted documentation were comparable to human-generated reports. Despite comparability, "quality & accuracy of output" remained a top frustration (Fig 7.5). This indicates that human oversight is still very important, which was also mentioned in the end-of-pilot meetings. These mixed feelings about quality and completeness were somewhat expected. The study from Kernberg et al. [22] stated that documentation from recordings did not meet clinical standards, while a study from Biswas and Talukdar [9] demonstrated great potential. As with the issues found in other studies [10], quality problems also emerged in the pilot due to hallucinations and misinterpretations of situations and conversations. These hallucinations will probably become less pronounced with new innovations in both transcription and language models, as well as better prompts. We can conclude that for the Salvation Army's application, where clinical standards are not required and Luisterlinie serves as a supportive tool, the quality issues were not as pronounced as in other studies. The technology still demonstrates great potential.

Increased satisfaction with documentation. Participants found reporting more enjoyable than before using the Luisterlinie app. They also reported experiencing a reduced administrative burden. Participants also overwhelmingly expressed a desire to continue using the app post-pilot and recommend it to colleagues for reporting purposes. The logs (Fig 7.13) showed frequent use of "Copy generated text". This suggests that users were integrating the AI's output into their existing documentation systems like WIJZ. Documentation is a major factor contributing to burnout in healthcare [25]. Therefore, it is great that participants now find reporting more enjoyable since using Luisterlinie, and also report a reduced administrative burden.

Concerns about quality. Participants expressed concerns about the quality and reliability of AI output, as well as operational and technical issues before the pilot in the kick-off survey (Fig 7.2). The Week 1 survey did reveal that the primary challenges and frustrations were related to output quality, technical issues, reliability, and the app's ease of use (Fig 7.5). Some of these issues, like reliability and technical issues, were due to the Luisterlinie app being in a pilot phase rather than a fully finished product. Quality issues like hallucination, factual incorrectness and limited contextual understanding were consistent with other studies [10, 32, 9].

8.1.3 Unexpected outcomes

The evaluation stage revealed some unexpected outcomes. Participants expressed unanimous agreement regarding their desire to continue using the Luisterlinie app and to recommend it to their colleagues. We anticipated that the most significant measurable increase would be in time, rather than job satisfaction. Another surprising finding was that participants primarily used desktop devices instead of phones (7.12). The expectation was that most conversations to occur outside the office, making phones more convenient for recording. However, this might not be the case.

8.2 Main research question

To address the main research question, "How can the addition of GenAI enhance the record-keeping capability for interviews and meetings in public healthcare organizations?" we find that integrating GenAI presents significant opportunities to improve record-keeping. It notably increases efficiency and reduces the administrative burden on employees, making administration more streamlined.

Key enhancements include time savings from automated transcription and summarization. This allows employees to spend less time on manual note-taking and more time engaging directly with clients or colleagues, thereby improving focus and presence during interactions. Furthermore, using GenAI can enhance job satisfaction by simplifying and speeding up the documentation process. Finally, GenAI can lead to more complete and consistent records by providing comprehensive summaries.

There are some caveats. Human oversight is needed to verify the generated documentation to prevent risks and inaccuracies. The output remains a concern, with AI hallucinations and misinterpretations sometimes leading to bad results. Technical stability and reliability are important for a tool that is used for documentation, as people quickly depend on it. User training to manage expectations and ensure good use might be needed to integrate GenAI responsibly in a healthcare environment

8.3 Limitations

This research has four limitations concerning the scope of the pilot, the number of participants, the Luisterlinie app and the data collection methods.

The first limitation of this research is its narrow focus. The study concentrated on the LJ&R department of the Salvation Army. To address this limitation, participants were selected from various regions, which added diversity to the group. This narrow focus also facilitated direct comparisons of survey results, as all participants held the same role as youth protection workers with the Salvation Army. The smaller very specific scope also makes the study less generalizable since this research examined a very specific case. Further research is needed to assess the extent to which these findings can be applied to other public healthcare settings or other AI tools.

The second limitation is that the number of participants was slightly smaller than planned. The initial aim was for 60 participants, but ultimately, 47 active individuals took part. This limitation was addressed by collecting data from a lot of various sources, which allowed us to extract insights despite the lower-than-expected participation.

The third limitation was the Luisterlinie app itself. As a pilot version, the app was not as stable as a fully tested application, which may have influenced the findings. It also caused frustration among users since the app sometimes malfunctioned or failed to record conversations correctly. This limitation was partly addressed by establishing a feedback mechanism through Microsoft Teams, allowing participants to quickly reach the development team and resolve issues efficiently. While this approach did not eliminate all errors and instability, it did resolve most issues for other users.

Lastly, the data collection methods primarily relied on self-reported data. This reliance may have affected the results, as participants knew they were part of a pilot study and self reported the data. Additionally, the study lacks quantitative data, mainly due to time constraints and pressure on participants. Implementing a more quantitative method would have required more time and placed further pressure on them. It was concluded that this approach was not worth pursuing for this pilot. Future research could focus on gathering more quantitative data.

Chapter 9

Conclusions

This thesis explored how GenAI could improve record-keeping in public healthcare, specifically at Salvation Army's LJ&R (Juvenile Protection) department, through a pilot of the "Luisterlinie" application. Using an action research methodology, the study evaluated the impact on transcription and documentation through metrics like time efficiency, record quality, user satisfaction, and identified benefits and challenges.

In conclusion, the thesis confirms that GenAI, via applications like Luisterlinie, can enhance public healthcare record-keeping by improving transcription efficiency, reducing documentation time, fostering better client engagement, and increasing user satisfaction. However, current GenAI limitations, particularly output accuracy, highlight the need for human oversight.

The pilot revealed significant potential for GenAI to transform documentation. AI-assisted processes drastically reduced administrative time, boosting efficiency and allowing youth care professionals to focus more on client interactions. While AI-assisted documentation quality matched human-generated reports, the study highlighted the continued need for human oversight to ensure accuracy and clinical accuracy, repeating existing literature.

A key positive outcome was improved job satisfaction, with participants overwhelmingly wanting to continue using and recommending the Luisterlinie app. This suggests GenAI can not only optimize workflows but also improve healthcare professional's daily experience and potentially reduce administrative burnout. Challenges primarily involved output accuracy and technical stability.

The Luisterlinie app addressed issues like excessive documentation time and inconsistent output. Unexpectedly, desktop usage for recordings exceeded phone usage, and the strong positive sentiment for continued app use surpassed initial expectations.

The main limitations of this study are the specific context, relatively short duration of the pilot, the piloting of a not quite stable application and the reliance on self reported data. These limitations can be explored further in future research

Bibliography

[1] Jeugdbescherming - Leger des Heils — legerdesheils.nl. https://www.legerdesheils.nl/zorg/jeugdbescherming, . [Accessed 18-06-2025].

[2] Leger des Heils: Missie en visie — legerdesheils.nl. https://www.legerdesheils.nl/wie-we-zijn, . [Accessed 18-06-2025].

[3] Asma Ben Abacha and Wen-wai Yim. An Investigation of Evaluation Metrics for Automated Medical Note Generation.

[4] Jihyun Ahn and Bokyoung Kim. Application of Generative Artificial Intelligence in Dyslipidemia Care. *Journal of Lipid and Atherosclerosis*, 14(1):77, 2025. ISSN 2287-2892, 2288-2561. doi: 10.12997/jla.2025.14.1.77. URL https://e-jla.org/DOIx.php?id=10.12997/jla.2025.14.1.77.

[5] Laila Akhu-Zaheya, Rowaida Al-Maaitah, and Salam Bany Hani. Quality of nursing documentation: Paper-based health records versus electronic-based health records. *Journal of Clinical Nursing*, 27(3-4), February 2018. ISSN 0962-1067, 1365-2702. doi: 10.1111/jocn.14097. URL https://onlinelibrary.wiley.com/doi/10.1111/jocn.14097.

[6] Anthropic. Introducing the next generation of claude, 2024. URL https://www.anthropic.com/news/claude-3-family.

[7] Shraddha Barke, Michael B. James, and Nadia Polikarpova. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proceedings of the ACM on Programming Languages*, 7 (OOPSLA1):85–111, April 2023. ISSN 2475-1421. doi: 10.1145/3586030. URL https://dl.acm.org/doi/10.1145/3586030.

[8] Lisa Ann Baumann, Jannah Baker, and Adam G. Elshaug. The impact of electronic health record systems on clinical documentation times: A systematic review. *Health Policy*, 122(8):827–836, August 2018. ISSN 01688510. doi: 10.1016/j.healthpol.2018.05.014. URL https://linkinghub.elsevier.com/retrieve/pii/S0168851018301635.

[9] Anjanava Biswas and Wrick Talukdar. Intelligent clinical documentation: Harnessing generative AI for patient-centric clinical note generation. pages 994–1008, may 2024. doi: https://doi.org/10.38124/ijisrt/IJISRT24MAY1483.

[10] Aisling Bracken, Clodagh Reilly, Aoife Feeley, Eoin Sheehan, Khalid Merghani, and Iain Feeley. Artificial Intelligence (AI) – Powered Documentation Systems in Healthcare: A Systematic Review. *Journal of Medical Systems*, 49(1):28, February 2025. ISSN 1573-689X. doi: 10.1007/s10916-025-02157-4. URL https://link.springer.com/10.1007/s10916-025-02157-4.

[11] Arthur H. Brayfield and Harold F. Rothe. An index of job satisfaction. *Journal of Applied Psychology*, 35(5):307–311, 1951. ISSN 1939-1854(Electronic),0021-9010(Print). doi: 10.1037/h0055617. Place: US Publisher: American Psychological Association.

[12] Sandeep Chataut, Sirjana Bhatta, Bishwambhar Dahal, Grishma Ojha, Bigyan Subedi, and Bijay Bastakoti. Advancements and Applications of Generative AI in Healthcare. *European Journal of Theoretical and Applied Sciences*, 2(6):873–895, November 2024. ISSN 2786-7447. doi: 10.59324/ejtas.2024.2(6).77. URL https://ejtas.com/index.php/journal/article/view/1364.

[13] Robert Davison, Maris G. Martinsons, and Ned Kock. Principles of canonical action research. *Information Systems Journal*, 14(1):65–86, January 2004. ISSN 1350-1917, 1365-2575. doi: 10.1111/j.1365-2575.2004.00162.x. URL `https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2575.2004.00162.x`.

[14] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, and Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. URL `http://arxiv.org/abs/2501.12948`. arXiv:2501.12948 [cs].

[15] Fabrizio Dell'Acqua, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *SSRN Electronic Journal*, 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4573321. URL `https://www.ssrn.com/abstract=4573321`.

[16] Department of Information Science and Engineering, RV College of Engineering, Bangalore, India and Pratham Agarwal. MedBot : A GenAI based Chatbot for Healthcare. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 08(06):1–5, June 2024. ISSN 25823930. doi: 10.55041/IJSREM35757. URL `https://ijsrem.com/download/medbot-a-genai-based-chatbot-for-healthcare/`.

[17] Emilio Ferrara. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. *First Monday*, November 2023. ISSN 1396-0466. doi: 10.5210/fm.v28i11.13346. URL `http://arxiv.org/abs/2304.03738`. arXiv:2304.03738 [cs].

[18] Betina Idnay, Zihan Xu, William G. Adams, Mohammad Adibuzzaman, Nicholas R. Anderson, and Bahroos. Environment Scan of Generative AI Infrastructure for Clinical and Translational Science. *npj Health Systems*, 2(1):4, January 2025. ISSN 3005-1959. doi: 10.1038/s44401-024-00009-w. URL `http://arxiv.org/abs/2410.12793`. arXiv:2410.12793 [cs].

[19] Institute of Education Sciences (IES). Creating Effective Surveys. November 2021.

[20] Johansen Monika A., Pedersen &Aring;se-Merete, and Ellingsen Gunnar. The Role of Medical Transcriptionists in Producing High-Quality Documentation. In *Studies in Health Technology and Informatics*. IOS Press, 2015. doi: 10.3233/978-1-61499-574-6-114. URL `https://www.medra.org/servlet/aliasResolver?alias=iospressISBN&isbn=978-1-61499-573-9&spage=114&doi=10.3233/978-1-61499-574-6-114`.

[21] K. Kelley. Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, 15(3):261–266, May 2003. ISSN 13534505, 14643677. doi: 10.1093/intqhc/mzg031. URL `https://academic.oup.com/intqhc/article-lookup/doi/10.1093/intqhc/mzg031`.

[22] Annessa Kernberg, Jeffrey A Gold, and Vishnu Mohan. Using ChatGPT-4 to Create Structured Medical Notes From Audio Recordings of Physician-Patient Encounters: Comparative Study. *Journal of Medical Internet Research*, 26:e54419, April 2024. ISSN 1438-8871. doi: 10.2196/54419. URL `https://www.jmir.org/2024/1/e54419`.

[23] Yogesh Kumar. A Comprehensive Analysis of Speech Recognition Systems in Healthcare: Current Research Challenges and Future Prospects. *SN Computer Science*, 5(1):137, January 2024. ISSN 2661-8907. doi: 10.1007/s42979-023-02466-w. URL `https://link.springer.com/10.1007/s42979-023-02466-w`.

[24] Noloyiso Makeleni and Liezel Cilliers. Critical success factors to improve data quality of electronic medical records in public healthcare institutions. *SA Journal of Information Management*, 23(1), March 2021. ISSN 1560-683X, 2078-1865. doi: 10.4102/sajim.v23i1.1230. URL `http://www.sajim.co.za/index.php/SAJIM/article/view/1230`.

[25] M. Hassan Murad, Brianna E. Vaa Stelling, Colin P. West, Bashar Hasan, Suvyaktha Simha, Samer Saadi, Mohammed Firwana, Kelly E. Viola, Larry J. Prokop, Tarek Nayfeh, and Zhen Wang. Measuring Documentation Burden in Healthcare. *Journal of General Internal Medicine*, 39(14):2837–2848, November 2024. ISSN 0884-8734, 1525-1497. doi: 10.1007/s11606-024-08956-8. URL `https://link.springer.com/10.1007/s11606-024-08956-8`.

[26] Voraprapa Nakavachara, Tanapong Potipiti, and Thanee Chaiwat. Experimenting with Generative AI: Does ChatGPT Really Increase Everyone's Productivity? *SSRN Electronic Journal*, 2024. ISSN 1556-5068. doi: 10.2139/ssrn.4746770. URL `https://www.ssrn.com/abstract=4746770`.

[27] Shakked Noy and Whitney Zhang. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence. pages 187–192, July 2023. doi: 10.1126/science.adh2586. URL `https://www.science.org/doi/10.1126/science.adh2586`.

[28] OpenAI. OpenAI o3-mini. URL `https://openai.com/index/openai-o3-mini/`.

[29] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[30] Sundar Pichai. Introducing Gemini 2.0: our new AI model for the agentic era. 12 2024. URL `https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#gemini-2-0-flash`.

[31] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision.

[32] Sajani Ranasinghe, Daswin De Silva, Nishan Mills, Damminda Alahakoon, Milos Manic, Yen Lim, and Weranja Ranasinghe. Addressing the Productivity Paradox in Healthcare with Retrieval Augmented Generative AI Chatbots. In *2024 IEEE International Conference on Industrial Technology (ICIT)*, pages 1–6, Bristol, United Kingdom, March 2024. IEEE. ISBN 9798350340266. doi: 10.1109/ICIT58233.2024.10540818. URL `https://ieeexplore.ieee.org/document/10540818/`.

[33] Danissa V Rodriguez, Katharine Lawrence, Javier Gonzalez, Beatrix Brandfield-Harvey, Lynn Xu, Sumaiya Tasneem, Defne L Levine, and Devin Mann. Leveraging Generative AI Tools to Support the Development of Digital Solutions in Health Care Research: Case Study. *JMIR Human Factors*, 11:e52885, March 2024. ISSN 2292-9495. doi: 10.2196/52885. URL `https://humanfactors.jmir.org/2024/1/e52885`.

[34] Benjamin Semujanga and Patrick Mikalef. Exploring the Productivity Impacts of Generative AI in Organizations. In Rogier Van De Wetering, Remko Helms, Ben Roelens, Samaneh Bagheri, Yogesh K. Dwivedi, Ilias O. Pappas, and Matti Mäntymäki, editors, *Disruptive Innovation in a Digitally Connected Healthy World*, volume 14907, pages 103–111. Springer Nature Switzerland, Cham, 2024. ISBN 978-3-031-72233-2 978-3-031-72234-9. doi: 10.1007/978-3-031-72234-9_9. URL `https://link.springer.com/10.1007/978-3-031-72234-9_9`. Series Title: Lecture Notes in Computer Science.

[35] Dimitrios Tsekouras, Rodrigo Belo, and Philipp Cornelius. Generative AI and Student Performance: Evidence from a Large-Scale Intervention. 11, 2024. URL `https://aisel.aisnet.org/icis2024/learnandiscurricula/learnandiscurricula/11`.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need.

Appendix A

Diagnosing interview

Mijn naam is Manuel en ik studeer Informatica & Economie hier aan de Universiteit Leiden. Daarnaast werk ik bij het Leger des Heils als AI/BI-specialist. Voor mijn scriptie onderzoek ik wat voor effect AI-applicaties hebben op documentatie binnen publieke zorg zoals het Leger des Heils.

Je deelname aan dit interview is vrijwillig en je mag op ieder moment aangeven als je ermee zou willen stoppen of als je een vraag liever niet zou beantwoorden. Daarnaast mag je tot 1 maand na dit interview aangeven als je liever toch niet hebt dat ik jouw antwoorden gebruik voor mijn onderzoek.

Dit interview zal ongeveer 1 uur duren. Om alles goed te kunnen analyseren wil ik je vragen of ik dit gesprek zou mogen opnemen, zodat ik later eventueel dingen zou kunnen terugluisteren. De opname wordt alleen door mij gebruikt en alles wat wij hier bespreken wordt geanonimiseerd. De data van dit interview wordt verwerkt volgens de richtlijnen van Universiteit Leiden. Ben je akkoord met de opname?

1. Professionele Achtergrond/introductie:

1. Zou u zichzelf kunnen voorstellen?
2. Hoe groot is de afdeling waarmee u werkt? Hoeveel mensen nemen interviews af?
3. Hoe belangrijk is het documenteren binnen LJ&R?
4. Hoeveel interviews worden er gemiddeld afgenomen/neemt u gemiddeld af (hangt af van wie er wordt geïnterviewd)

2. Huidige Documentatieprocessen (SITUATIE):

1. Kan u de stappen beschrijven die je doorloopt als u een interview houd en documenteert?
2. Hoeveel tijd besteedt u meestal aan deze stappen?
3. Welke tools of methoden gebruikt u momenteel voor het documenteren van interviews en vergaderingen (bijv. handmatig notities maken, audio-opname, transcriptiediensten, specifieke software)?
4. Is er een proces/handleiding voor documenteren, en in hoeverre wordt dit gevolgd?

3. Tevredenheid en Uitdagingen met Huidige Processen:

1. Wat gaat er goed in het documentatieproces?
2. Wat kan er beter in het documentatieproces?

5. Algemene Vragen:

1. Bent u beschikbaar voor eventuele vervolgvragen indien nodig?
2. Heeft u nog aanvullende opmerkingen of suggesties?

Appendix B

Kick-off survey

The Likert scale was used for most questions because these types of questions are generally less exhausting to answer and provide a good source of data [19]. The study by [21] was also used as a reference.

Measuring time spend

1. "Ik besteed momenteel meer tijd dan dat ik zou willen aan het administreren in WIJZ" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
2. "Ik ben tijdens een cliëntgesprek veel bezig met notities maken" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
3. "Ik ben in staat om mijn administratie na gesprekken met cliënten op tijd af te ronden" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
4. Welk percentage van een gemiddelde werkweek wordt gebruikt voor het administreren van client gesprekken in WIJZ? (0% - 100%)
5. Hoeveel minuten ben je gemiddeld bezig om een cliëntgesprek van een uur te verwerken in WIJZ?

Measuring quality and completeness

6. "Ik maak voor elk contact met een client een contact journaal" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
7. "De administratie die ik teruglees in WIJZ is over het algemeen volledig" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
8. "Er missen vaak belangrijke gegevens in de documentatie die ik in WIJZ teruglees" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
9. "Het is makkelijk om alle relevante administratie vast te leggen in WIJZ" (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
10. Hoe beoordeel je de algehele kwaliteit (bijv. nauwkeurigheid, duidelijkheid, bruikbaarheid) van de documentatie die je momenteel in WIJZ vastlegt? (Zeer Slecht tot Zeer Goed)

Documentation satisfaction

11. "Ik vind het administreren van gesprekken een gemakkelijke taak." (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)
12. "Het administreren van gesprekken kost mij veel moeite." (Helemaal oneens, Oneens, Neutraal, Eens, Zeer eens)

Kansen en risico's

13. Wat zijn volgens jou de belangrijkste voordelen en kansen van AI binnen jou werk?

14. Wat zijn volgens jou de belangrijkste risico's en nadelen van AI binnen jou werk?

After the survey, a consent form was included to make sure participants were aware that they were participating in both a pilot and academic research.

Appendix C

Week 1 survey

The survey consists of open and closed questions. These surveys will measure time spend, quality and completeness, user satisfaction and benefits and challenges. It will also include some questions about the usability of the Luisterlinie application, to catch any important errors that might come up in the pilot that can influence the outcome of the survey. The Likert scale was used for most questions because these types of questions are generally less exhausting to answer and provide a good source of data [19]. The study by Kelley [21] was also used as a reference.

Starting questions:

1. Heb je de Luisterlinie app gebruikt? (Selecteer één optie: ja, nee)
2. In het geval van niet, waarom heb je deze week de Luisterlinie app niet gebruikt? (tekstveld)

Measuring time spent

3. In vergelijking met de periode voordat ik de Luisterlijn app gebruikte, hoeveel tijd besteed ik nu aan het registreren van contacten (contactjournalen) in WIJZ...? (Scale: veel meer tijd - veel minder tijd)
4. "Verslaglegging in WIJZ was deze week efficiënter doordat ik gebruik maakte van de Luisterlinie app" (Helemaal oneens - Helemaal eens)
5. "De introductie van de Luisterlinie app heeft geleid tot dubbel werk bij verslaglegging (bijv. AI-samenvattingen moeten nakijken en dan opnieuw documenteren)" (Helemaal oneens - Helemaal eens)
6. "Door het gebruiken de Luisterlinie app maakte ik minder aantekeningen tijdens mijn client gesprekken" (Helemaal oneens - Helemaal eens) [10]
7. "Ik leg meer contact journalen vast nu ik gebruik maak van de Luisterlinie app" (Helemaal oneens - Helemaal eens)

Measuring quality and completeness

8. "De samenvattingen/inzichten die door de Luisterlinie app werden gegenereerd bevatte alle belangrijke aspecten van het gesprek" (Helemaal oneens - Helemaal eens)
9. "Ik moest vaak belangrijke ontbrekende informatie toevoegen aan de verslaglegging die door de Luisterlinie app werd gegenereerd om deze accuraat/bruikbaar te maken" (Helemaal oneens - Helemaal eens)
10. "Ik vind de **volledigheid** van de verslaglegging die door de Luisterlinie app wordt aangemaakt vergelijkbaar met die van door mensen gemaakte verslaglegging" (Helemaal oneens - Helemaal eens)
11. "Ik vind de **kwaliteit** van de verslaglegging die door de Luisterlinie app wordt aangemaakt vergelijkbaar met die van door mensen gemaakte verslaglegging" (Helemaal oneens - Helemaal eens)

12. Als je alleen zou moeten vertrouwen op de verslaglegging die wordt gegenereerd door de Luisterlinie app, welke aanvullende informatie of wijzigingen zouden absoluut noodzakelijk zijn? Waarom? (open vraag)

User satisfaction

13. Ik vind verslagleggen leuker nu ik de Luisterlinie app gebruik dan voordat ik de Luisterlinie app gebruikte. (Helemaal oneens - Helemaal eens)[11]
14. "Ik ervaar dat mijn administratielast is verminderd sinds het gebruik van de Luisterlinie app" (Helemaal oneens - Helemaal eens)
15. "Ik vind het jammer als ik de Luisterlinie app niet kan gebruiken binnen een gesprek" (Helemaal oneens - Helemaal eens)
16. "Ik vind de Luisterlinie app fijn in gebruik" (Helemaal oneens - Helemaal eens)
17. "Ik zou de Luisterlinie app aanbevelen aan andere collega's voor hun administratietaken" (Helemaal oneens - Helemaal eens)
18. "Ik zou de Luisterlinie app graag willen blijven gebruiken na de pilot" (Helemaal oneens - Helemaal eens)
19. Hoe vaak had je de Luisterlinie app afgelopen week willen gebruiken?
20. In hoeveel procent van de gevallen kon je de Luisterlinie app toen ook gebruiken?
21. Kunt u een schatting geven: In welk percentage van de gesprekken waar het mogelijk was om de Luisterlinie app te gebruiken heeft u (gemiddeld) om toestemming gevraagd om het gesprek op te nemen?
22. Denkend aan de keren dat u om toestemming vroeg voor een opname: In welk percentage van die gevallen werd de toestemming (gemiddeld) gegeven? (percentage)

Benefits & challenges

23. Wat zijn de grootste voordelen die je hebt ervaren bij het gebruik van de Luisterlinie app? (Open tekstveld)
24. Wat zijn de grootste uitdagingen of frustraties die je bent tegengekomen bij het gebruik van de Luisterlinie app? (Open tekstveld)

Stability/usability

This section is here to check if there are any technical issues that might have affected the experiment.

25. Heb je problemen ervaren bij het gebruik van de Luisterlinie app? (Selecteer één optie: Ja, Nee)
26. (Indien Ja op vorige vraag) Zo ja, kun je kort beschrijven welk(e) probleem/problemen je bent tegengekomen? (Open tekstveld)
27. (Indien Ja op vorige vraag) Ondervond je dit probleem voornamelijk bij het gebruik van de Luisterlinie app op uw telefoon, computer, of beide? (Opties: telefoon, computer, beide)
28. Heb je andere feedback op de Luisterlinie app?

Appendix D

Week 4 survey

The survey consists of open and closed questions. These surveys will measure time spend, quality and completeness, user satisfaction and benefits and challenges. It will also include some questions about the usablility of the Luisterlinie application, to catch any important errors that might come up in the pilot that can influence the outcome of the survey. The Likert scale was used for most questions because these types of questions are generally less exhausting to answer and provide a good source of data [19]. The study by Kelley [21] was also used as a reference.

Starting questions:

1. Wat is je naam?
2. Heb je de Luisterlinie app gebruikt? (Selecteer één optie: ja, nee)
3. In het geval van niet, waarom heb je de Luisterlinie app niet gebruikt? (tekstveld)
4. In het geval van wel, hoeveel gesprekken heb je ongeveer vastgelegd met Luisterlinie? (tekstveld)

Measuring time spent

5. "Sinds ik de Luisterlinie app gebruik (nu 4 weken), besteed ik over het algemeen minder tijd aan het registreren van contacten (contactjournalen) in WIJZ." (Helemaal oneens - Helemaal eens)
6. "De Luisterlinie app heeft mijn documentatieproces over de afgelopen 4 weken efficiënter gemaakt."
7. "Het controleren en eventueel aanpassen van de door Luisterlinie gegenereerde samenvattingen kostte mij significant minder tijd dan het volledig zelf schrijven van een verslag."
8. "Ik kon door het gebruik van Luisterlinie meer aandacht besteden aan de cliënt tijdens gesprekken, omdat ik minder hoefde te noteren."

Measuring quality and completeness

9. "De samenvattingen/inzichten die door de Luisterlinie app werden gegenereerd bevatte alle belangrijke aspecten van het gesprek" (Helemaal oneens - Helemaal eens)
10. "Naarmate ik de Luisterlinie app langer gebruikte, hoefde ik minder vaak belangrijke ontbrekende informatie toe te voegen of correcties aan te brengen aan de gegenereerde verslaglegging." (Helemaal oneens - Helemaal eens)
11. "De volledigheid van de door Luisterlinie gegenereerde verslaglegging is, na 4 weken gebruik, goed vergelijkbaar met of beter dan mijn handmatige verslaglegging." (Helemaal oneens - Helemaal eens)
12. "De kwaliteit (bijv. accuraatheid, duidelijkheid) van de door Luisterlinie gegenereerde verslaglegging is, na 4 weken gebruik, goed vergelijkbaar met of beter dan mijn handmatige verslaglegging." (Helemaal oneens - Helemaal eens)
13. Als je kijkt naar de afgelopen 4 weken, welke typen informatie of aanpassingen waren het vaakst nog nodig bij de door Luisterlinie gegenereerde tekst? (open vraag)

User satisfaction

14. "Mijn algehele tevredenheid over het documentatieproces is verbeterd sinds ik de Luisterlinie app gebruik." (Helemaal oneens - Helemaal eens)
15. "De Luisterlinie app is een waardevolle toevoeging geworden aan mijn dagelijkse werkroutine." (Helemaal oneens - Helemaal eens)
16. "Ik heb het gevoel dat de Luisterlinie app mijn administratieve last daadwerkelijk heeft verminderd over de afgelopen 4 weken." (Helemaal oneens - Helemaal eens)
17. "Ik zou de Luisterlinie app na deze pilot graag willen blijven gebruiken." (Helemaal oneens - Helemaal eens)
18. "Ik vond de Luisterlinie app na enige gewenning gemakkelijk te gebruiken." (Helemaal oneens - Helemaal eens)
19. "Ik zou de Luisterlinie app aanbevelen aan collega's die vergelijkbaar werk doen." (Helemaal oneens - Helemaal eens)
20. Denkend aan de afgelopen 4 weken: In welk percentage van de gesprekken waar het mogelijk was om Luisterlinie te gebruiken, heb je (gemiddeld) om toestemming gevraagd aan de cliënt/betrokkenen om het gesprek op te nemen? (percentage)
21. Denkend aan de afgelopen 4 weken: In welk percentage van die gevallen werd de toestemming (gemiddeld) gegeven? (percentage)

Benefits & challenges

22. Wat zijn, na 4 weken gebruik, de grootste BLIJVENDE voordelen die je ervaart bij het gebruik van de Luisterlinie app?
23. Wat zijn, na 4 weken gebruik, de grootste BLIJVENDE uitdagingen of frustraties die je ervaart bij het gebruik van de Luisterlinie app?

Stability/usability

This section is here to check if there are any technical issues that might have affected the experiment.

24. "De Luisterlinie app was over de afgelopen 4 weken over het algemeen stabiel en betrouwbaar in gebruik." (Helemaal oneens - Helemaal eens)
25. "Het was duidelijk waar ik terechtkon met vragen over de pilot" (Helemaal oneens - Helemaal eens)
26. "De instructie in de kick-off meeting gaf mij genoeg duidelijkheid om aan de slag te gaan met de Luisterlinie app" (Helemaal oneens - Helemaal eens)
27. Heb je problemen ervaren bij het gebruik van de Luisterlinie app? (Selecteer één optie: Ja, Nee)
28. (Indien Ja op vorige vraag) Zo ja, kun je kort beschrijven welk(e) probleem/problemen je bent tegengekomen? (Open tekstveld)
29. (Indien Ja op vorige vraag) Ondervond je dit probleem voornamelijk bij het gebruik van de Luisterlinie app op uw telefoon, computer, of beide? (Opties: telefoon, computer, beide)
30. Welke functionaliteiten zou je graag toegevoegd zien aan de Luisterlinie app om het nog nuttiger te maken voor jouw werk? (open tekstveld)
31. Heb je andere feedback op de Luisterlinie app?