Text summarisation in healthcare to reduce workload

Summarising patient experiences for healthcare professionals

Jan Mark Dannenberg

TU Delft

Text summarisation in healthcare to reduce workload

Summarising patient experiences for healthcare professionals

Thesis Report

by

Jan Mark Dannenberg

to obtain the degree of Master of Science
at the Delft University of Technology to be defended publicly on July 3 2024 at 16:30

Thesis committee:	
Chair:	Dr. C. Lofi
Supervisors:	Dr. J. Yang
	Dr. ir. J. Jung
Place:	Faculty of Electrical Engineering, Mathematics and Computer Science, Delft
Department:	Web Information Systems
Project Duration:	October, 2023 - July, 2024
Student number:	4889576

Faculty of Electrical Engineering, Mathematics and Computer Science · Delft University of Technology

Contents

Nomenclature

List of Abbreviations

AI	Artificial Intelligence
EHR	Electronic Health Record
EMC	Erasmus Medical Centre
LLM	Large Language Model
ML	Machine Learning

List of Figures

List of Tables

1 Introduction

Lately the medical domain has been in the news often in the Netherlands. Reports of the workload being perceived as very high by healthcare professionals in all different fields of expertise have become common [1]. Healthcare professionals are overwhelmed by the amount of work and the increasing number of patients. In turn, this leads to high burnout rates within this group of healthcare professionals [2]. This has great negative consequences for all stakeholders within the medical domain; medical errors, lower (perceived) quality of care, and decreased patient satisfaction [3, 4]. In addition, increasingly more professionals have chosen to change careers (or stop), which only further contributes to the need for more people within the healthcare domain [5, 6].

As can be seen, the increasingly high work pressure within the healthcare domain is a dangerous development, as it causes problems within one of the most important sectors of a country. And by no means it is limited to the Netherlands. Research has found that this has become a problem in multiple countries, such as the United States [7]. This raises the important question: what is causing this increased workload? The answer is not only the growing older demographic within these countries, but also the rise of Electronic Health Records (EHR) and the documentation tasks related to these records [7, 8]. The amount of additional tasks healthcare professionals have to perform surrounding their 'normal' work activities creates a documentation burden that is one of the main contributing factors that lead to burnouts. The time consumed by these tasks can be tremendous, even close to 50% [9]. In addition to causing stress, this also leads to high turnover rates [6] and low work engagement, as professionals do not enjoy these tedious manual tasks [10].

However, the rise of Artificial Intelligence (AI) might be able to provide solutions, especially for this documentation burden [11]. Large Language Models are becoming increasingly better at a range of natural language tasks and are able to perform quite as well as humans on these tasks. This provides great opportunities to reduce the workload of healthcare professionals by implementing smart strategies to automate parts of the documentation burden. In this thesis research, the focus will be on summarisation tasks surrounding patient interactions of healthcare professionals with a slight emphasis on consultation sessions. Consultation sessions are a perfect example in which all stakeholders are negatively affected by the high workload of the healthcare professional. Only a short time is available for each patient, doctors constantly have to interact with their computer during the consultation session, and have to perform numerous tasks after concluding a session before they can see the next patient. From preliminary experiences, it was found that most of these tasks are tied to text summarisation in some way. Therefore, the use of LLMs for this particular purpose in the healthcare domain will be the main goal of this thesis research. In the following sections of this chapter, the research of this thesis will be outlined and further introduced.

1.1. Scenario

In this section, the basic scenario of a patient interaction will be described to clarify the content of this thesis research by an example. This scenario is visualised in Fig. 1.1.

The core of the example is the patient interacting with the healthcare professional. This could be through any means of communication, such as phone calls, emails, or face-to-face conversations. The patient shares experiences related to his or her disease or treatment, asks questions, and answers questions from

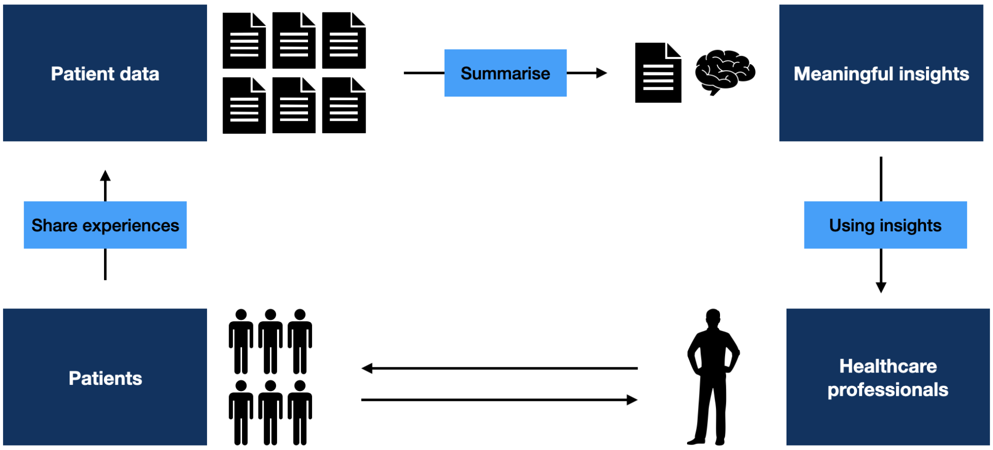

Figure 1.1: Scenario of patient interaction by healthcare professional for which summarisation used

the healthcare professional. Naturally, this leads to patient data in multiple forms; transcripts, notes, emails, etc. Even a 15-minute consultation already leads to quite some patient data that need to be processed in some way by the healthcare professional. The purpose of this processing is mainly to obtain and store particular insights that the healthcare professional deems important or relevant to the current treatment and well-being of the patient. In order to do this, the healthcare professional uses summarisation of patient data, with the most obvious example result being medical notes [12]. This particular step is the one on which this research will focus as it can be time-consuming [13] and create a high workload [8], but it is necessary for both the healthcare professional and the patient.

1.2. Motivation

There were several reasons to focus on this particular problem within the healthcare domain. First, previous research conducted by students with the same supervisor focused on similar aspects within the healthcare domain. This work can be seen as a preparation for more practical research towards a real-life solution to assist healthcare professionals. Furthermore, personal interest in applicable solutions within a particular domain matched with this particular problem setting. In addition, the supervisor's connections with the Erasmus Medical Centre (EMC) also facilitated this research.

1.3. Scope

Given the time constraints of a Computer Science thesis, it is important to determine the scope of the research; what it touches upon and what it does not touch upon. Naturally, it limits itself to the healthcare domain where healthcare professionals and patients are the main stakeholders, with a slight focus on a hospital setting. In addition, it focusses on patient interactions data, especially consultation session transcripts. Furthermore, it aims to provide a solution specifically for the summarisation task with regard to patient interaction data.

The goal of this research, which will be further elaborated on in Chapter 2, is to define the needs & requirements of healthcare professionals and translate them into a system that meets these needs. However, it is important to note that the aim is not to build a technically perfect system. Naturally, this will be tried as much as possible, but the lack of good training data (and the lack of time to obtain such a dataset) limits this goal. In addition, voice-to-text techniques will not be covered by this research, but it will assume that these techniques are good enough to provide correct transcripts of, for example, consultation sessions. This technique is not needed for every summarisation task (email, patient history), but is crucial to the performance of the system for other tasks (consultation session, phone call).

1.4. Structure

The purpose of this report is to inform the reader about the research on text summarisation within the healthcare domain to reduce the workload of healthcare professionals that has been conducted for this master's thesis. The content of this report will discuss the methodology used in the research and the way it was applied. In addition, research findings will be presented, including a functional prototype and its evaluation.

The structure of this thesis report is as follows. First, the main problem that this research touches upon will be discussed, to further motivate and validate the thesis research. In Chapter 2 the problem and knowledge gap will be formally defined, which creates a basis for the research questions posed in Chapter 3. In addition, in Chapter 3, the research methodology, Design Science Research Method (DSRM) [14, 15], used to carry out this research will be discussed. Then, in Chapter 4, a literature review will be used to deep dive into the important concepts that are related to this research and that are necessary to carry out meaningful studies. Chapter 5 will discuss the way interviews (using an initial prototype) with healthcare professionals were organised to identify the needs of a text summarisation system. Subsequently, Chapter 6 will analyse these interviews and define the system requirements that were found. In Chapter 7 the prototype of the system will be further constructed and improved using the system requirements by translating them into practical improvements and additions. The constructed and improved prototype, along with the system requirements, are evaluated in Chapter 8 to prove their relevance and usefulness. Limitations, personal reflections, and possible future research directions will be discussed in Chapter 9. Finally, Chapter 10 concludes by presenting the key findings of this thesis research and showing how these findings answer the defined research questions.

2 Background

This chapter will dive into the background of the research problem and define it more formally. First, work pressure within the healthcare domain will be discussed. Then, the current status of summarisation of patient interactions will be described, which is the setting that the research aims to address. Some of the observations about work pressure, manual summarisation, and noise present in patient interactions have been based on preliminary experiences with healthcare professionals. The stakeholders in the problem will be discussed in combination with the negative consequences they face because of this problem. Subsequently, the knowledge gap that defines the research problem will be defined along with the value that a solution would provide to stakeholders.

2.1. Work pressure in healthcare

As mentioned previously in the introduction, work pressure and work load appear to be at an all-time high for healthcare professionals [1]. The many tasks they have to perform in a limited time per patient have a huge impact on professionals in various ways, including an alarming high rate of burnout [6, 2].

Especially the use of electronic health records (EHR) has contributed to the increase in different documentation tasks that the healthcare professional is required to perform. This can be a great burden for healthcare professionals as it leaves them with less time for the same number of patients [8]. The amount of time spent on these tasks that involve the use of a computer can range from 50% to the work time of the professional [9]. This increased work load not only negatively affects healthcare professionals [6, 4], but also affects patients in dangerous ways [3].

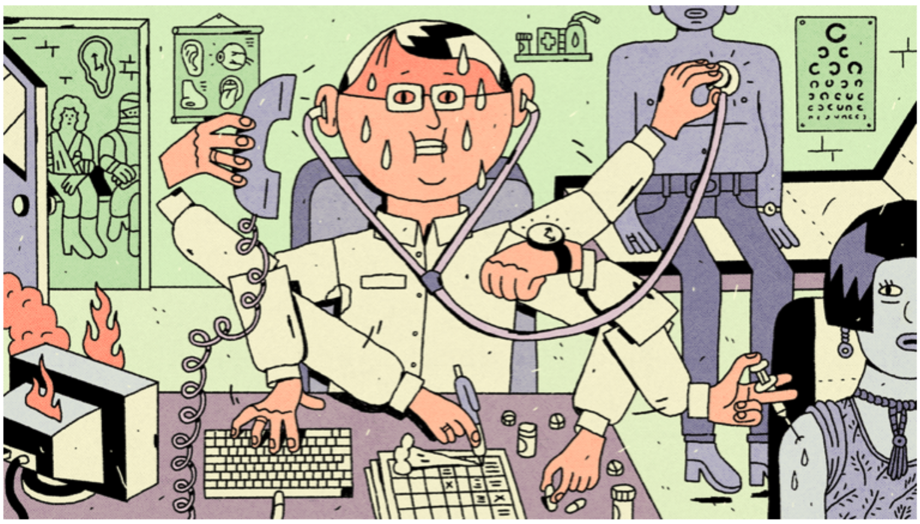

Figure 2.1: Healthcare professionals have an increasing high workload due to the many tasks surrounding patient interactions.

One of the most task-intensive patient interactions that healthcare professionals have during their work is a consultation session (illustrated in Fig. 2.1. This can range from a monthly check-up to an intake session or even a post-surgery check-up. The duration of these consultations also differs; it can range from 10 minutes to an hour, but usually takes around 20 minutes [16]. During and primarily surrounding these consultation sessions, healthcare professionals have to perform numerous tasks in addition to listening and talking to the patient. Many of these tasks originate from the need for EHRs or take longer due to the existence of EHRs. First, preparation before a consultation session can be lengthy, as the healthcare professional will have to go through the (possibly long) patient history to some extent to gain some background knowledge about the patient [17]. During the session, the patient may need to be physically examined, and the results of the tests should be registered. Afterwards, the healthcare professional has to schedule possible next sessions and medical examination appointments. In addition, the professional might have communicated with other healthcare professionals about the patient or the decisions that have been made (for example, the local general practitioner). Finally, an important task also is to document the consultation session using some form of medical notes, which is a time-consuming task [18].

As mentioned in Chapter 1, high work pressure and work load negatively impact the healthcare professional. Major medical errors, less patient happiness, less patient time, and burnout are only a few of the serious implications [3, 7]. This research aims to offer solutions specifically targeted towards (semi)-automated text summarisation, which is why the following section will focus on the current state of summarisation within the medical domain when it comes to patient interactions.

2.2. Manual summarisation in patient interactions

Going back to the scenario described in Section 1.1 and the description of tasks surrounding the consultation sessions in Section 2.1, it becomes clear that some time-consuming tasks are characterised by a summarisation process by the healthcare professional. Especially the preparation and documenting step of the consultation session rely heavily on some form of summarisation; extracting the most important/relevant information from (a set of) document(s). This section will discuss how this is currently performed manually by healthcare professionals.

Before, during and after a consultation session, there are numerous tasks in which a healthcare professional has to apply summarisation. This has mainly to do with extracting key information from some (spoken) data to document it for future reference [19]. A consultation session is not only a physical consultation session, but could also be an online session, a phone call, or an email. Documenting this information causes work pressure and is simply not pleasant work for healthcare professionals [8]. Furthermore, especially during contact with the patient, it has numerous negative effects on the perception the patient has of the quality of the healthcare. For example, research found that patients stopped talking while the doctor was typing on their computer [20]. Additionally, non-verbal communication such as eye contact is crucial in consultation sessions which is also hindered by the documenting tasks the professional has to perform [21].

The preparation before a consultation session consists mainly in summarising the relevant pieces of the patient's medical history. This is done by healthcare professionals in their own way and structure, which was observed in preliminary experiences. The summarisation during and after the consultation session of the session itself is more structured and, in the Netherlands, follows the SOAP protocol (or SOEP in Dutch) of subjective, objective, assessment, and plan format [19]. However, based on preliminary experiences, healthcare professionals often summarise in a non-overlapping structure, and their summaries often differ greatly in length. Healthcare professionals often have to rush the process of summarisation during and after a consultation session because the next patient is already waiting and may be late on schedule. The summarisation before the consultation session is time-consuming and requires the healthcare professionals to prepare on time.

In addition, noise is an important factor in the summarisation process as it is widely present in the medical domain [22]. As patient interactions are based on some form of human dialogue, although through written communication, it can naturally contain irrelevant information that will be labelled *noise*. This noise is not bad; one could even argue that it is a necessary part of patient interactions. However, in this research, it is labelled as noise, as it is not relevant to the healthcare professional when considering the treatment or disease of the patient. This also means that this noise is not necessarily all the nonmedical information shared by the patient or healthcare professional, such as the vacation to Germany that the patient went on. It could also be medical facts or medical information the patient shares that is not relevant to this particular healthcare professional. A good example of this kind of noise is a report on a toe surgery the patient had 10 years ago, if the patient is undergoing a lung cancer treatment today. Another example is for the patient to mention all types of small pain that healthcare professionals do not consider relevant to the current disease the patient experiences. This inherently makes the noise within the medical domain subjective, as one healthcare professional might find the family situation of a patient relevant to his treatment, whereas another might think this is not relevant at all. Therefore, filtering through this subjective noise is a complicating factor which makes summarisation difficult, especially from a software perspective.

2.3. Problem

Using the findings of the previous sections, this thesis research focusses on the problem of the manual time-consuming summarization healthcare professionals have to perform in the context of patient interactions, which contributes to increased work pressure and work load they have to deal with. These tasks are manual and repetitive and are not enjoyed by healthcare professionals. In addition, they contribute to the high burnout and turnover rates among them. The other stakeholders, patients, are also negatively affected by these time-consuming tasks as they may have to wait longer for a response and have less interactive time with the healthcare professional. These negative consequences are visually summarised in Fig. 2.2). The problem on which this research is based is formulated more formally as follows:

Problem

The patient experiences shared during consultation sessions, emails and other means of communication can be lengthy and include a lot of noise.
Healthcare professionals need to spend a lot of time summarising these to gain meaningful insights for diagnosis and treatment, which causes work pressure and increases the administrative, repetitive work.

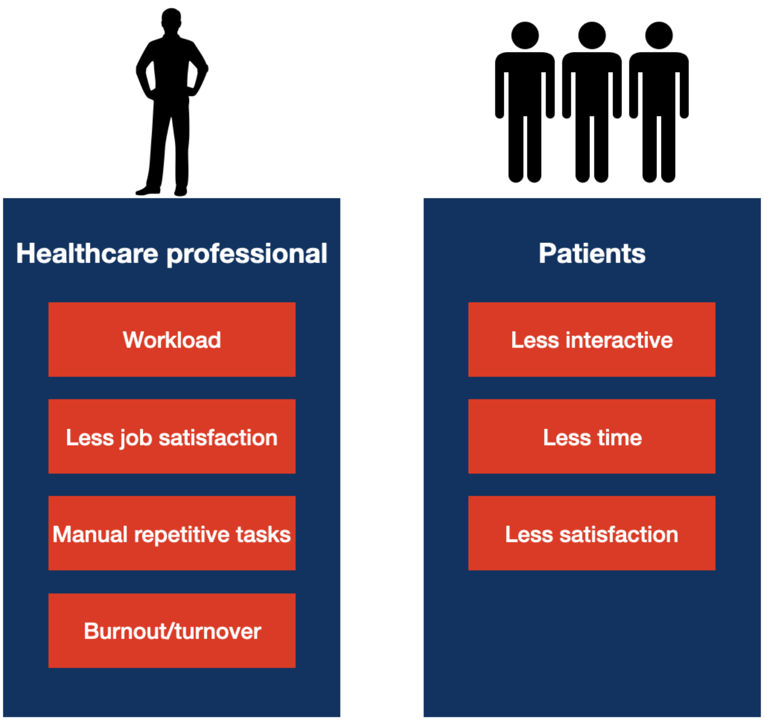

Figure 2.2: Negative consequences for stakeholders due to a lot of time being spent on summarisation by healthcare professionals

2.4. Knowledge Gap

Using the problem statement defined in Section 2.3 it becomes clear that there is some missing knowledge to tackle this problem. In this case, the missing knowledge can be seen as a missing system which can help healthcare professionals with the tasks described in the problem statement. This knowledge gap has been defined as:

Knowledge Gap

There is no system which fits the need of the healthcare professionals and can summarise text patient experience documents.

This research aims to fill the missing knowledge of the problem and the objectives and design of a system described in the knowledge gap. This will create value for both stakeholders, healthcare professionals, and patients. First, the healthcare professional will have more time to focus on other tasks. Second, it increases the job satisfaction of the healthcare professional as the work load negatively impacts job engagement and satisfaction [2, 10]. It also allows the professional to spend more time on intelligent tasks and patient interactions. Lastly, the system should also provide in-depth summaries that contribute to the quality of EHRs. In terms of patients, they benefit from the healthcare professional not having to split his time between a screen and the patient, but he can interact more with the patient. Furthermore, since a part of the tasks of the healthcare professional can be performed more efficiently, there will be more time left for the patient. Lastly, the quality of the EHRs will also positively benefit the quality of healthcare, including diagnosis and treatment. The value in solving the problem and the knowledge gap has been visually summarised in Fig. 2.3

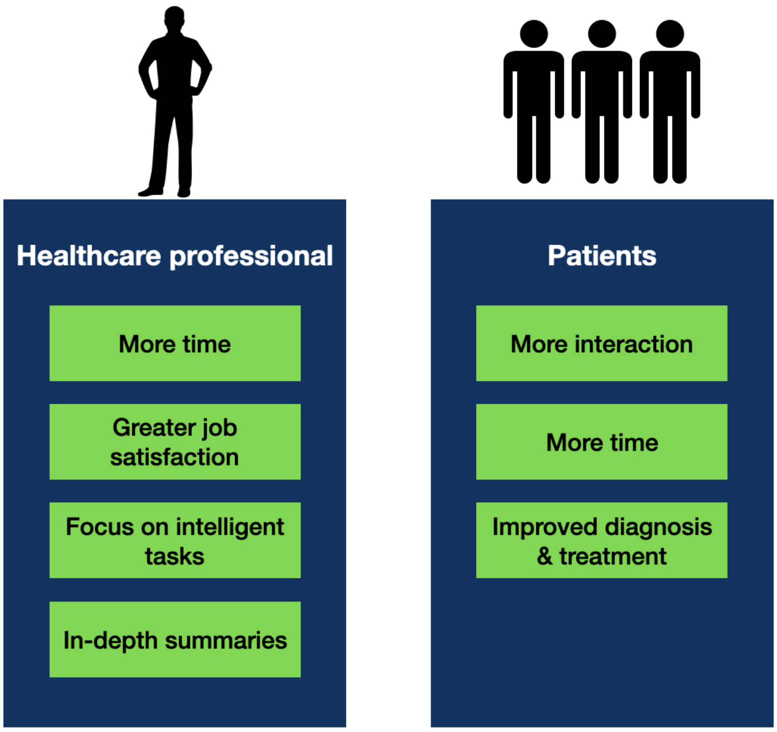

Figure 2.3: Value for stakeholders gained by solving the problem

3 Methodology

In this chapter, the methodology used to carry out the research will be introduced and discussed. First, research questions originating from the problem statement and the knowledge gap described in Section 2.3 and Section 2.4 will be defined to guide research throughout the course of this thesis. The research questions will then be related to the different methods and components used in this research in Section 3.2. Then, the Design Science Research Framework will be introduced in Section 3.4 and applied to this research thesis in Section 3.3. The use of semi-structured interviews with stimuli to support the research conducted will be explained further and discussed in Section 3.5. Finally, some ethical considerations will be highlighted in Section 3.6.

3.1. Research Questions

As discussed in the previous chapter, this research focusses on the summarisation of patient experience documents. The purpose of this research is to find a solution to the problem statement formulated in Section 2.3 by exploring the problem setting and researching the needs of healthcare professionals for an automated summarisation system. Together with the formulated knowledge gap, this problem can be translated into the primary research question:

Primary Research Question

How can the summarisation of patient experience documents for healthcare professionals be automated by AI to reduce their workload?

In order to answer this research question, the research is divided into multiple smaller parts. This also further structures the research and provides clear goals to focus on. This thesis defines five sub-research questions to guide the research:

Research Question 1

What are the needs of healthcare professionals to trust and use an AI tool for summarising patient experience documents?

Research Question 2

What are the current barriers to use an AI tool for summarising patient experience documents?

Research Question 3

What amount and types of noise are present in patient experience documents?

Research Question 4

How can an AI tool be designed to fit the needs of healthcare professionals?

Research Question 5

How can the designed AI tool be evaluated and improved upon?

The first two sub-research questions, 1 *(user)* and 2, focus on the needs and barriers to such a system. 3 is closely related to this as it investigates the noise that can be present during patient interactions and the way doctors perceive and handle it. The fourth research question focusses on the design and implementation of a system that adheres to the knowledge/requirements found via the first three research questions. The last question, 5, then aims to answer the evaluation part of this system.

3.2. Process

To answer the question posed in the previous section, this thesis research will use several different methods in the various stages of the research. Generally, research can be divided into three different phases; a discovery phase, a design phase, and an evaluation phase. The first phase corresponds to the first three research questions in which the current state of the problem setting is investigated and requirements & barriers to a solution should be discovered. Then, in the second phase, a system is defined according to the knowledge collected in the first phase (4). Lastly, the proposed solution is evaluated in the final phase, which corresponds to the last research question.

Throughout these phases, various methods have been chosen to conduct the research:

- **Literature review:** This method is present throughout the entire research timeline, but will mostly be used in the first phase of the research as it is very knowledge intensive. Literature will be used to provide a solid foundation for the research and by using existing practices and methodologies throughout the thesis.
- **Interviews:** Interviews will be used gather knowledge from expert within the domain and problem setting. Interviews are the most commonly used data collection technique [23]. This is especially useful in this setting as the interviewees will be the key users of a proposed solution [24]. Interviews will be used in this thesis as a means to guide the design of the solution. Section 3.5 will dive deeper into why semi-structured interviews are used in this thesis.
- **Prototype:** A prototype of the researched solution will be developed to aid the research and make the output tangible. It will be used in combination with interviews to stimulate the participants and show them what is possible [25]. Furthermore, it will be developed in an iterative fashion.
- **Evaluation:** This is the key to measuring the quality of the output of the research. First, the prototype will be evaluated where possible and in addition the key findings of the thesis will also be evaluated.

These different elements belong to different phases of the research, although some might be used in multiple phases. By relating them to the phases, they are also directly related to the five research questions. This has been visualised in Fig. 3.1

3.3. Design Science Research Framework

To conduct the research in this thesis in a structured and meaningful way, the three core concepts of the Design Science Research Framework [14], will be explained and applied to the content of this research. The DSR Framework, depicted in Fig. 3.2, was designed to show how information systems (IS) research is embedded in a certain environment (context) and knowledge base. By being aware of this and applying this to information system research, the utility and relevance of the research output will be ensured. In addition, the authors of the framework also provided seven guidelines for Design Science Research [14]. Although these guidelines will not be mentioned explicitly, they are interwoven in the following sub-sections of this section.

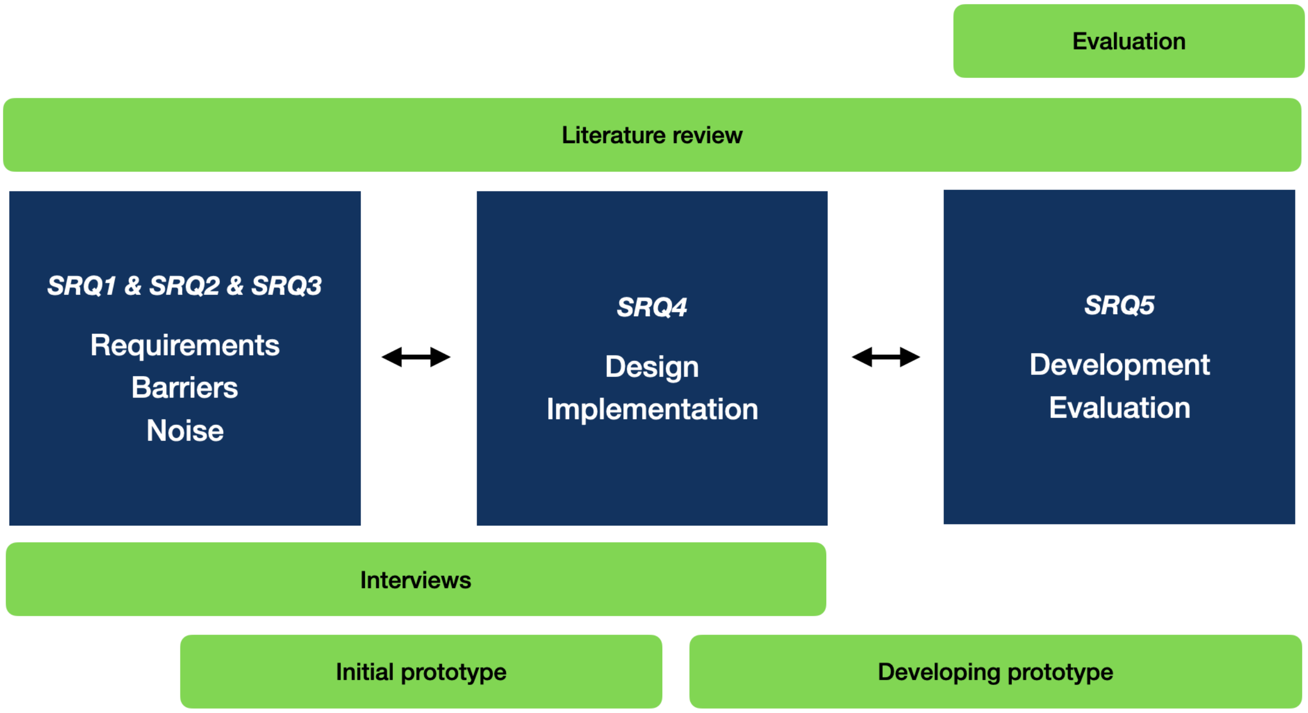

Figure 3.1: The relation of the research questions with respect to each other and the different methods used in the three phases of the research.

As mentioned, the three different pillars of the Design Science Research Framework are environment, IS research, and knowledge base. Each of those pillars provides an integral part to the research and will be discussed in more detail and related to the thesis research in the following sections.

3.3.1. Knowledge Base

The knowledge base pillar of the DSR Framework is the foundation that provides methodologies, theories, and knowledge that serves as a basis for research [14]. In short, it consists of the existing knowledge and literature on which the research is built, which ensures reserach rigour. The theory is used mainly for the *develop / build* phase of the research and could be seen as a kick-start to the research. Methodologies can be used in the *justify/evaluate* phase, as they are proven and validated and justify the validity of the research conducted.

For this thesis research, the knowledge base is consulted primarily in the literature review part of the research. It provides an understanding of the current state-of-the-art technologies for text summarisation and the current state of summarisation within the healthcare domain (which can be found in Chapter 4. In addition, the knowledge base is used to find techniques and methodologies to conduct interviews, develop a prototype, interact with LLM, and perform a meaningful evaluation.

3.3.2. Environment

The environment encapsulates the problem context, which consists of people, organisations, and technologies that interact with each other in the problem setting [14]. Considering this aspect of the problem ensures that the research has a meaningful application to the 'real world'. The environment also defines the business needs that the research output should meet. The requirements formulated by the environment impact the *develop/build* phase of the research. Furthermore, in the *justify/evaluate* the output should be sufficient for the stakeholders in the environment, which ensures the relevance of the problem and therefore the solution.

The environment within this thesis consists of healthcare professionals in the medical domain that interact with patients and have to perform some form of text summarisation. This environment will be

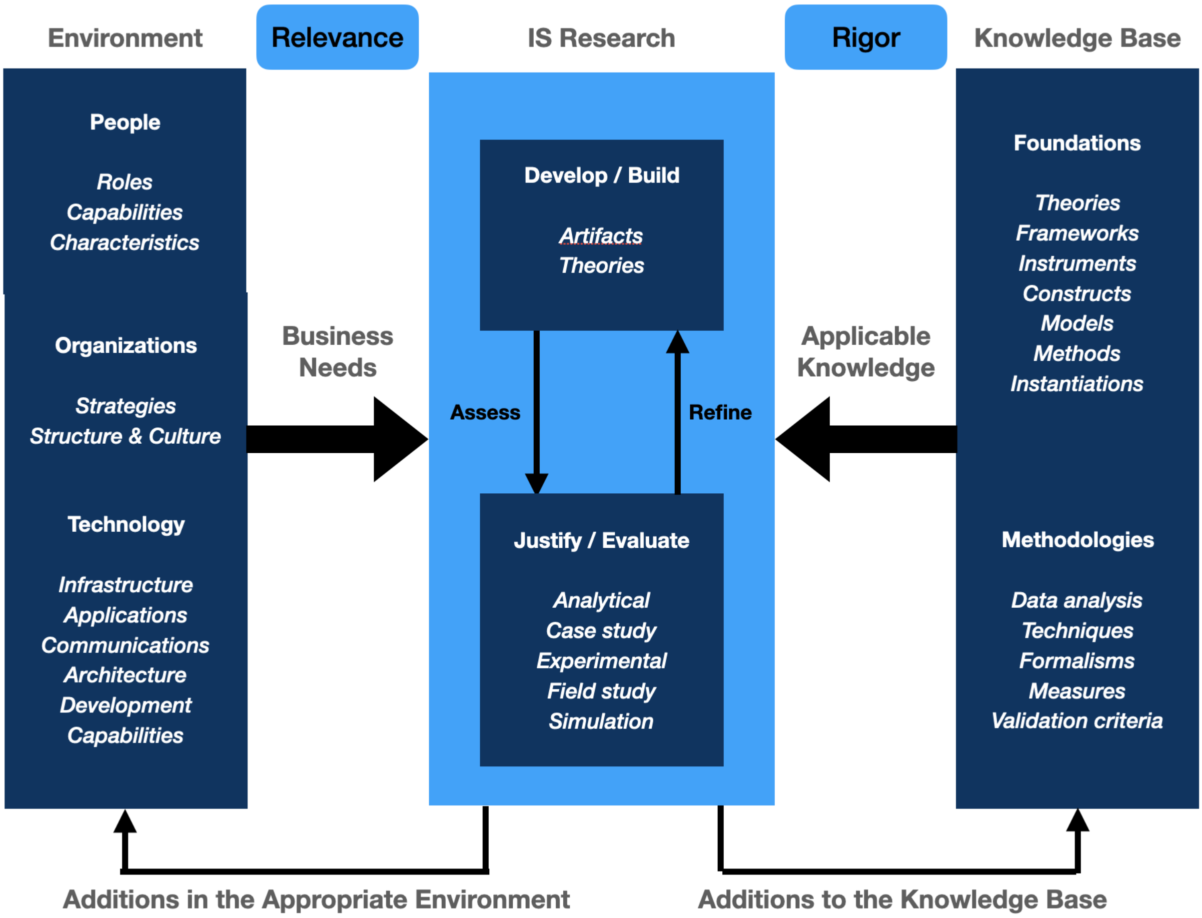

Figure 3.2: Three pillars of Design Science Research in IS, adapted from [14]

closely investigated by conducting interviews and evaluation with experts in this environment, which shapes the output prototype of this research.

3.3.3. IS Research

The pillar in the middle of the DSR framework is the core of the research, it is the place where the design and evaluation of the so-called design artefact are carried out [14]. It consists of the two, previously mentioned stages; the *develop/build* and the *justify/evaluate* phase. These phases are not linear, but influence each other in an iterative fashion. The output of these iterations is the design artefact, which can range from theory, models, constructs, guidelines to an actual product/system. If done correctly, the artefact will contain additions both to the environment (meeting the business needs) and to the knowledge base.

In this thesis, IS research will consist mainly of developing the prototype and evaluating it. Furthermore, the key findings of this research might also be considered a design artefact and output of the research.

3.4. Design Science Research Process Model

The Design Science Research Process Model [15], illustrated in Fig. 3.3, builds on the DSR Framework discussed in the previous section. This model defines a sequential number of phases that encapsulate the different elements and guidelines of the DSR framework. As this model provides further justification for the process discussed in Section 3.2 and can be used as a basis for planning, this section will shortly discuss the different phases and apply them to this research.

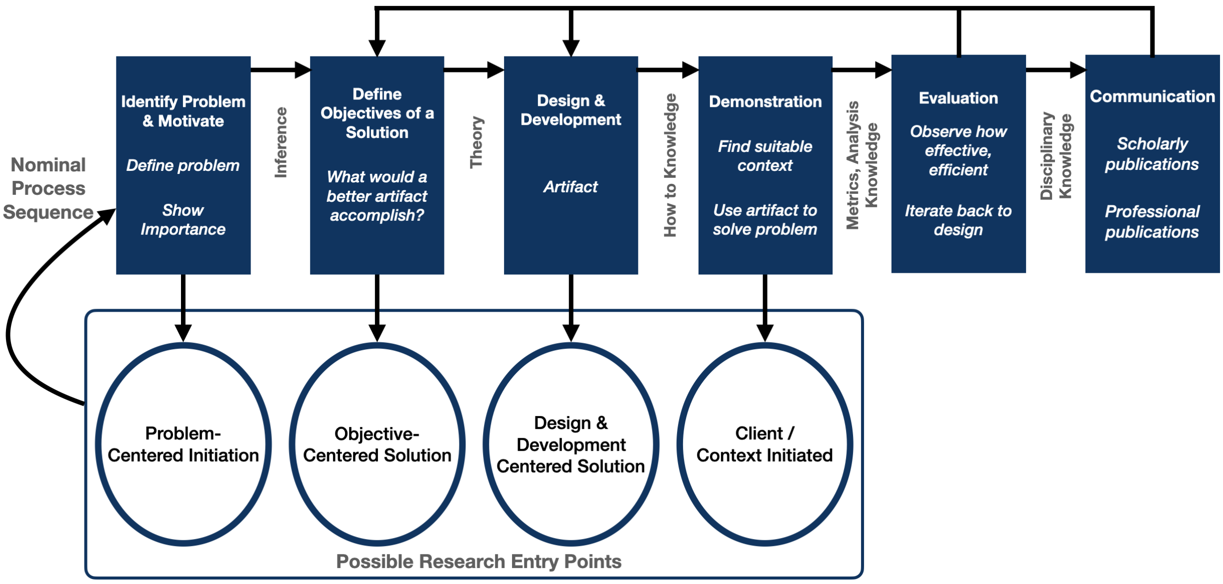

Figure 3.3: DSRM Process Model adapted from [15]

3.4.1. Problem Identification & Motivation

The first stage of the model focusses on the discovery of the setting of the problem and its relevance. This ensures the validity of the research and the relevance of a possible solution. In this stage, the problem is defined and the importance is discussed.

In this thesis, this stage mainly consists of defining the problem and further validate and dive into it using interviews. In addition, the literature review is used to investigate the context of the problem.

3.4.2. Solution Objectives

The second stage is closely related to the first stage as the outline of a solution is defined. The objectives and requirements for a solution are defined in this stage by investigating the environment.

In this thesis, the interviews serve as the basis for this stage, as experts, which are also the key users of the system, will be able to voice their requirements (and barriers) to a solution. Furthermore, this is also used to investigate what exactly is needed to provide a better solution than the current systems that are in place.

3.4.3. Design & Development

This stage consists of actually designing and developing the design artefact that the research aims to deliver. Naturally, this is done in an iterative fashion with the other stages of the model.

For this research, this step consists of designing and improving the prototype as it is the main output of the thesis. The use of interviews, feedback, and evaluation ensures that this is done iteratively.

3.4.4. Demonstrate & Evaluation

Although these steps are separate in the original model, they are closely related (especially for this thesis). These steps ensure that the design artefact is demonstrated in the environment in which it can be evaluated. The output of this step is then also fed back into the previous step.

In this thesis report, these steps consist of letting healthcare professionals use the prototype and evaluate it (through interviews). The result of these demonstrations is then used to further improve the prototype.

3.4.5. Communication

The last stage of the model consists of communicating the findings and artefacts of the research to appropriate audiences. This is done most commonly through publications or presentation talks.

This thesis will likely not fully incorporate this step of the model as it might require additional work and transformation of the thesis report document. However, the research will be made public by hosting it in the TU Delft archive.

3.5. Semi-structured Interviews with Stimuli

As mentioned, semi-structured interviews are used as the main means of collecting data and interacting with the environment of the problem setting. The main reason for this choice is due to the fact that the solutions objectives are usually defined by experts in the field. When dealing with human perception and their opinions/views, it is generally a good idea to use interviews for data collection [26]. Semi-structured interviews provide a good way to collect data in a uniform way, while still allowing for dynamic interviews where necessary [26]. They are widely used when follow-up questions are desirable to extend knowledge to particular bits of information [27].

The core of semi-structured interviews is the interview guide, which is a general outline of the questions to be asked with additional follow-up questions where necessary [28]. Verbal and non-verbal actions can be used to *probe*, which can allow the interviewee to further extend on the view or feeling expressed [26, 29]. The interview guide should consist of at least five or six questions that make up the main body of the interview [24]. To construct such an interview guide, systematic methods have been developed consisting of various steps to create an interview guide [29]. These steps will be used in Chapter 5 and Chapter 8 to build the interview guides needed. The main challenge of semi-structured interviews is to keep the interview dynamic and open ended, while still being able to perform a comparative analysis on the results of the interviews [30]. The main output of the interviews are the interview transcripts, which can be further used for (coding) analysis.

In addition to semi-structured interviews, the prototype will be used as a stimuli for participants to further ideate about the requirements and possible improvements to a text summarisation system. Stimuli can be effective in creating more participation and gaining more in-depth information about the research topic [31, 25]. Especially, creating engagement for healthcare professionals by showing them what is possible is important as it keeps them hooked and interested in the results of the research. It also adds a tangible aspect to an otherwise 'vague' and theoretic study. In Chapter 5, the interviews conducted for this research will be discussed in more detail.

3.6. Ethical considerations

Since this research is dealing with real people, patients, and healthcare professionals, within a complex domain, the ethical considerations of this research should be discussed. These ethical considerations are not the main priority of this research and may not be fully addressed by the results of this research. However, it is important to identify these risks and take them into account as much as possible, although a part of them will be left to future research.

The main ethical consideration of this research is the risk of medical (documentation) errors due to the use of a (semi) automated summarisation system and the responsibility that comes with this risk. As the output of this research (or future research) could be used by healthcare professionals instead of manual summarisation, the risk is always present that important information is left out. The loss of these data or the addition of incorrect data (which is directly related to the risk of hallucination with LLMs [32]) could negatively impact patients. This research tries to focus on the responsible role of the healthcare professional, which is also explained and highlighted in Chapter 6 and Chapter 7.

Another consideration is the privacy of the patients. As the summarisation tooling will be processing patient experience data, this might be leaked to unintended people, which might harm the patient (or healthcare professional). This research tries to choose and use technology in a way that it can be easily made as secure and private as possible. An example of this, which will be further extended in Chapter 7, is the choice of using a private open source LLM that can be trained and used privately within a medical organisation.

4 Literature Review

As mentioned in Section 3.3, the knowledge base is an important part of the Design Science Research model [14]. To cover this aspect of information systems research, a literature review has been conducted in the field of text summarisation. This chapter aims to provide a solid foundation for the further contents of this thesis and guides the choices made for the design of the prototype.

This chapter has been divided into two main sections; the first discusses long-text summarisation and current methods aimed at performing this. The second narrows the scope to text summarisation in the medical domain. As an output of this chapter, two tables will be provided which show the different approaches found in the literature and compare the different aspects of them. This will be used to base the decisions on the implementation of the prototype.

4.1. Automated text summarisation

This section discusses long text summarisation from a basic computer science perspective, without focussing on a particular field or domain. Specifically, long text summarisation techniques were chosen because patient experience text documents can be lengthy and do not have predefined constraints. A system which automates the summarisation process has to be able to handle long texts while still providing an accurate summary. Dealing with long texts can be problematic for language models, due to memory considerations, and are difficult to scale.

Upon searching for long-text summarisation literature, five main approaches were chosen to consider for this review of the literature. These approaches were selected on the basis of the date of publication to ensure that state-of-the-art methods are considered. Furthermore, the number of times the papers were referenced and in which research they were referenced was also used. Lastly, whether the technique fit the research context of this thesis was also considered. For example, if a technique had a limit of a certain number of characters, this resulted in the technique not being selected here as it would not be possible to use it in this research. All the approaches reviewed in this section are summarised in Table 4.1. For each approach considered, the architecture and long document mechanism are described. Furthermore, whether it was created using public data is reported in this table as this influences the applicability of the approach to this thesis. Furthermore, whether the approach is tailored towards conversational data (much patient experience text is conversational), if it supports human-in-the-loop (likely will be necessary for legal and ethical reasons), and the evaluation metrics used by the approach are shown.

In the field of text summarisation, there are two main approaches to the summarisation problem; extractive and abstractive. The extractive approach aims to identify the most important sentences/parts (keyphrases) out of a text and build a summary by combining them in some way. Using such an approach, all output data can be mapped back onto the input data in some way. The other approach, the abstractive approach, aims to create a summary by creating new sentences and words based on the identified key ideas and elements it captures [33]. Both approaches have their uses, but for this research abstractive summarisation especially seems to best fit the research context. The summaries written by healthcare professionals are written from a different perspective and use a different vocabulary compared to the original patient experience text. For example, a long explanation of a patient during a consultation of a certain symptom can be reduced to one medical term by the healthcare professional. Furthermore, the input structure and context can differ from the output summary. However, for long text summarisation,

Model (Architecture, Year)	Long Document Mechanism	Public data	Dialogue	Human -in-the -loop	Evaluation
TLM+Ext [33] *(Transformer, 2020)*	Content Selection + Discourse Bias	✓	✗	✗	ROUGE + Human
DANCER [34] *(Transformer, 2020)*	Content Selection + Discourse Bias	✓	✗	✗	ROUGE
SEAL [35] *(Transformer, 2020)*	Content Selection + Segment-wise Scorer	✓	✗	✗	ROUGE
LoBART [36] *(Transformer, 2021)*	Content Selection + Efficient Attention	✓/✗	✓/✗	✗	ROUGE + Memory usage
PG-Net [37] *(Transformer, 2019)*	Topic Aware Attention	✗	✓	✗	ROUGE

Table 4.1: Comparison table of current approaches to the summarisation of long documents.

often a hybrid approach is used where the extractive part is used to reduce the text or divide the input text in some clever way. Then, the abstractive part is used to still be able to rewrite the text in some other structure/vocabulary.

First, TLM+Ext [33] takes a hybrid approach to the long text summarisation problem by first extracting significant parts of a text and thus not exceeding the number of character limits. The authors show that this approach delivers promising results for the summarisation of papers. However, it only uses the decoder mechanism of the transformer network, which could be optimised by also using the encoder part [38]. Furthermore, it is also difficult to easily apply this to patient experience documents as they could be structured in entirely different ways than scientific papers. Patient experience documents are not standardised by any means and each document can differ in length and structure. DANCER [34] uses a different hybrid approach in which the extractive part is implemented by calculating the similarity of the sentence and using the sections formed as input for the abstractive summarisation step implemented by a pointer-generator network (transformer). SEAL [35] dynamically extracts input phrases / fragments to summarise using sparse attention. LoBART uses an approach which is very closely related to TLM + Ext, where content selection is also done by training a transformer network instead of a more static approach to selecting the parts to use as input for the summarisation. Lastly, PG-Net [37] also utilises a pointer generator network for summarisation, but only uses an abstractive approach in which the pointer generator network is trained to pay special attention to the topics in the text. The issue here is that a lot of training data is required to build a sufficiently performing summarisation system, and topics need to be defined up front.

Many of the long-text summarisation approaches take a hybrid approach to circumvent the issue of having to deal with large texts in the actual text summarisation step. Splitting the larger problem into multiple smaller problems is a familiar divide-and-conquer approach, which is widely used to deal with long text summarisation. When dealing with LLMs to summarise the text, this can also be useful as it is very reminiscent of prompt chaining [39] which can be used to divide the problem into subtasks and circumvent the issues of hallucination [32] and context length restrictions. As can be seen in Table 4.1, the approaches are mainly not geared toward conversational data, which remains to be explored. Furthermore, they do not incorporate a human-in-the-loop approach, which might be necessary within the medical domain.

4.2. Automated text summarisation in healthcare

Within the field of text summarisation there has been research conducted specifically for medical notes and medical documents. Considering these approaches is relevant as they aim to perform the same set of tasks as the automated text summarisation will perform. However, some of these approaches do not provide any mechanisms to deal with long texts and thus make certain assumptions about the length and/or structure of the input data or limit them solely to certain summarisation tasks. For each approach, in addition to the previous table, the medical domain mechanism is also listed; the way they tailor the summarisation model towards medical text. An overview of the approaches is shown in Table 4.2.

MedicalDS [40] aims to tackle the summarisation problem by extracting triples based on matches with

Model (Architecture, Year)	Long Document Mechanism	Medical domain mechanism	Public data	Dialogue	Human -in-the -loop	Evaluation
MedicalDS [40] *(Ontology, 2020)*	Triple Extraction+Matching	Medical ontology combined with triple extraction + matching	✗	✓	✗	ROUGE + False Positives
Dialogue2Note [41] *(seq-to-seq, 2021)*	Section prediction	Annotation labels / pre-training medical notes data	✗	✓	✗	BLEU + ROUGE
Cluster2Sent [42] *(Transformer, 2021)*	Section clustering + section-conditioned network	Pre-training medical notes data	✓/✗	✓	✗	ROUGE + Human
GPT-3-ENS [43] *(LLM, 2021)*	Use of dialogue snippets	Use of medical entity recognizer + pre-train on synthetic medical notes data	✗	✓	✗	ROUGE, Medical Concept Coverage, Negation Correctness + Human
SentBERT [44] *(LLM, 2021)*	Multistage summarization + grouping similar snippets	Finetuning pre-trained model with medical notes data	✗	✓	✗	ROUGE + Human
Dr. Summarize [13] *(Transformer, 2020)*	-	One-hot vector encoding the presence of medical concepts	✗	✓	✗	ROUGE, Medical Concept Coverage, Negation Correctness + Human
BART-finet [45, 46] *(LLM, 2022)*	-	Finetuning pre-trained model with medical notes data	✓/✗	✓	✓/✗	Levenshtein distance, METEOR, BERTScore

Table 4.2: Comparison table of current approaches to document summarisation within the medical domain.

their ontology and using this to construct a summary and therefore is an extractive approach. As it does not use any language models, it does not have to deal with the long-text issue as other summarisation techniques would. Dialogue2Note [41] uses a seq-to-seq model in combination with a section prediction to create a structured summary. It relies heavily on structured training data to identify what information belongs to which section and can not be easily generalised to different systems which will use a different structure for the output. Cluster2Sent [42] takes a similar approach, using a section-conditioned network for section clustering, which is essentially very similar to Dialogue2Note. Dr. Summarize [13] improves on a basic pointer-generator network approach by making modifications specifically tailored to the problem of negations within the input text. BART-finet [45] explores how LLMs can be used for the generation of medical summary by fine-tuning the BART model on private medical note data. Furthermore, they provide a mock data set that can be used for future research [46]. Unfortunately, this data set is relatively small (57 transcripts) and contains only fake data instead of anonymised data. SentBERT [44] takes an interesting approach to the long text issues it aims to address as it uses an LLM (finetuned on the data of medical notes) for the underlying implementation. It groups similar snippets and feeds those to the summarisation pipeline. Afterwards, the individual summaries are combined by applying another summarisation step. Lastly, GPT-3-ENS [43] takes a different approach in which it uses dialogue snippets, created by a sliding window, summarises these, identifies medical phrases within the summaries, and converts them to a medically accurate description.

These approaches also confirm that summarisation moves towards LLMs and that these models show promising results for the medical domain. However, all of the approaches lack a human-in-the-loop approach and some of them do not deal with long text. Another big barrier is the need and use of large, private datasets for training the different approaches.

Important to note is that when the structured training data is available and some (hybrid) extractive summarisation approach is taken, pointer generator networks/transformers perform well. However, from the literature review it seems that the focus has shifted to LLMs as they can provide flexibility and more capabilities which are needed for summarisation of patient experience text. Furthermore, these models can be fine-tuned when necessary, but also already provide great results without training data required. In addition, defining the output structure/format/vocabulary, without having to define this explicitly but leveraging the underlying knowledge of an LLM is greatly beneficial to the use of these models. Based on these findings, LLMs will be used in the remainder of this research as the basis for text summarisation, although they will be extended to tackle problems specific to the research and domain context.

5
Interviews

Research questions 1 *(user)*, 2 and 3 aim to discover the requirements, barriers, and knowledge needed to build a text summarisation system for patient experiences. In the previous chapter Chapter 4, the knowledge base was researched to find current practices and state-of-the-art technologies related to the problem context. In this chapter, interviews will be used as a method to investigate the environment of information systems research. To be more specific, semi-structured interviews with the use of stimuli, a prototype, will be used.

The first section of this chapter will show how a structured method was used to create a semi-structured interview guide for thesis research. Then, Section 5.2, we will discuss the initial prototype that was created specifically for interviews. The last three sections of this chapter will describe how interviewees were selected, how interviews were conducted, and what data were the result.

5.1. Interview Guide

As discussed in Section 3.5, semi-structured interviews will be used to investigate the requirements, needs & barriers of healthcare professionals to use a text summarisation system. The basic requirements to conduct such research have also been briefly mentioned in those sections. Building on these requirements, this section will use a structured method (slightly altered) from [29] to construct an interview guide that will be used in this first phase of discovery interviews. The following subsections correspond to the different steps of this method and will discuss the actions taken at each step.

5.1.1. Identification & Retrieving Prior Knowledge

The first step of [29] consists in obtaining the necessary knowledge to be able to conduct the interviews. Furthermore, this stage also forces the researcher to validate whether interviews are an appropriate method for his research.

In this thesis, this step is taken by preliminary experience and a review of the literature on current practices and solutions (Chapter 4). The preliminary experiences were useful to obtain some domain knowledge and witness how work pressure and summarisation influence the healthcare professionals in the 'real world'. The latter, the literature review, was necessary to get a good understanding of the current state-of-the-art methods used within this field. In addition, relevant topics and themes were identified that provided a solid foundation for the interviews.

5.1.2. Creating Preliminary Interview Guide

The next step of [29] consists in creating an initial interview guide using the knowledge from the previous step. This is necessary to have a basis that can be improved (iteratively). As for this thesis, mainly the goals determined the questions that were in this first draft of the interview guide. This follows from the first three research questions that investigate the current state of summarisation, the needs & barriers of healthcare professionals, and the noise that can be present in patient interactions. Furthermore, the use of stimuli, the prototype, resulted in a section which aimed to explore as much as possible about the use of healthcare professionals and the requirements of such a system in order to be used for further improvements.

5.1.3. Testing Preliminary Interview Guide

After creating the preliminary interview guide, it needs to be tested to ensure its feasibility and reveal any hidden problems with the interview guide [29]. Therefore, in this stage, some validation is needed to ensure that the interview questions yield the desired results. When possible, it is recommended to conduct several test interviews with experts to test the initial guide. However, the limited time available for interviews with various healthcare professionals made this impossible. Therefore, the thesis supervisor has tested and reviewed the preliminary interview guide. This resulted in several changes/additions, which have been summarised:

- The additions of more introductionary questions about the healthcare professional (expertise, years of experience, what patient interactions). This is needed to ensure a good spread of interviews and allows for comparison.
- Focusing on the *why* of the answers to several questions. Rather than only knowing something, it is important to understand why this is the case for that specific question.
- Creating a tighter schedule due to the limited time and shifting the focus more to the prototype as this is important for the remainder of the research.

5.1.4. Finalizing Preliminary Interview Guide

After integrating the improvements into the preliminary interview guide, the final step of [29] could be carried out by creating the final interview guide that will be used during the interviews. This interview guide can be found in Table 5.1.

As can be seen, the interview guide is divided into several sections. First, the content of the research is introduced and the consent form is signed. The healthcare professional is then identified / classified by asking questions about his expertise, his years of experience, and the patient interactions he has during his work. Next, the current summarisation workflow that they perform during their patient interactions is explored. This also includes the time they need to spend on these tasks and where they see possibilities for improvement, which helps to further validate the problem context. Then, the questions about needs are used to discover what the system requirements to a solution are. These questions also touch on the way automated summarisation would be implemented and where the responsibility would lie. The barriers are explored next in order to allow healthcare professionals to share their concerns or mention any no-go's that would apply to a solution. Lastly, the prototype is introduced and several questions are used to get feedback on the prototype and to further stimulate participants to think about an automated flow and its requirements.

5.2. Prototype

As mentioned in Chapter 3, this thesis also uses a prototype during the course of the research. Section 3.5 argues why this is a useful addition to semi-structured interviews and Section 3.2 shows how this prototype is related to the different stages of the research. This section will dive deeper into how the prototype was initially developed and used in the interviews. Note that this is not the final/iteratively improved version of the prototype, the development and improvement of the prototype will be discussed in Chapter 7.

The aim of the initial prototype was to prompt the interviewees for more responses and to show them a tangible example that could be improved. Therefore, a basic implementation would suffice. However, an important finding from the preliminary experiences was the amount of noise that healthcare professionals have to filter through when performing summarisation. Given this fact, an additional layer to text summarisation was added as an initial approach to support this filtering. For this, topic modelling is used as a technique, which is further discussed in Subsection 5.2.1. Topic modelling is also used as a way to deal with the problem of summarisation of long texts mentioned in Chapter 4, as it provides a way to divide the text into smaller subparts.

5.2.1. Topic modelling: tackling noise and divide & conquer

Topic modelling is a technique that aims to find topics / groups within a text data set, to discover hidden semantic patterns [47]. There are numerous approaches to do this that use different methods of finding these relations within the text corpus, which can result in different output topics. Text, and especially conversational text, is characterised and structured by topics [48]. Depending on the length and type

Interview guide
Introduce goal of summarising patient experience to reduce workload on healthcare professionals. Mention that this summarisation is topic-based.
This interview will be used to identify the needs and barriers of healthcare professionals to using such a system. A prototype of this system will be used to further facilitate this.
Show and explain the consent form. After that, start recording.
Identify healthcare professional (2 min.)
1: What is your main field of expertise within EMC?
2: Could you describe the patient interactions you have? These include verbal and written communication.
3: How many years of experience do you have within this field with this level of patient interactions?
Current workflow (6 min.)
4: What kind of patient experience documents/interactions do you have? E.g., consultation sessions, phone calls, emails, patient history documents.
5: How much of your time goes towards administrative tasks for these patient interactions? E.g., preparing consultation sessions, going through patient emails. • Could you describe why this task is necessary? • Could you describe what this task implies?
6: For which of those tasks do you apply direct or indirect summarisation? • What makes summarisation good for this task? • How do you perform this summarisation for this task? • Does your way of summarisation for this task differ from other colleagues?
7: What processes regarding patient experiences would benefit greatest from automatic summarisation?
Identify needs (10 min.)
8: Do you think an AI tool will be able to fully automate the tasks you described in the current workflow? • What part do you think is irreplaceable and why?
9: Do you think an AI tool should aim to be a tool, rather than a fully automated approach?
10: What would an automated AI approach need to be capable of for you to use it? • What challenges do you think the model needs to overcome? • Is explainability an important property for you? • Why is that property absolutely necessary?
Identify barriers (7 min.)
11: Are there any barriers (absolute no-go's) when it comes to using an AI model for summarisation of patient experiences? • Do you think a model would be able to solve/circumvent this?
Prototype demo (10 min.)
Introduce the prototype and its different components.
12: Do you think the topic based approach is useful? • Do you frequently use topic based approaches in consultation sessions or when going through emails/patient documents?
13: What is the prototype lacking in your opinion? • How can the prototype include this? • What would you change?
14: If possible, would you use the prototype in its current state (ignoring the privacy/performance issues)? • What would need to improve in order for you to use the model? • Do you think this also holds colleagues and nurses?

Table 5.1: Interview guide for interviewing healthcare professionals about patient experience summarisation.

of conversation, this could range up to any number of topics. This also holds for the conversations that healthcare professionals have during (web) consultation sessions, phone calls, etc. These could include medical topics, such as the medication the patient should use, and non-medical topics, such as the sports event the patient attended yesterday.

Within the medical domain and patient interactions there is a lot of noise [22, 49]. For example, the sports event talk that the healthcare professional has with the patient before the 'real' consultation session content could be considered noise, as it is not relevant for the healthcare professional from a medical perspective. In this thesis, we define noise as irrelevant parts of patient experience texts.

Alongside this definition of noise, there are two important notes. First, if some part of the text is noise, it does not imply that that part of the text was wasted time. It could be hugely important to patient satisfaction (doctor asks patient about his vacation) or be relevant in a different scenario/context (patient history mentioning toe surgery, but patient is now in cancer treatment). This immediately shows that noise does not have to be nonmedical information, but could also be medical information. Second, the notion of noise is subjective from the perspective of healthcare professionals. Preliminary experiences showed that one healthcare professional values the talk about the patient in the family situation much and will definitely include it in their summary, as they believe that this is relevant for patient treatment. However, another healthcare professional does not think that this should be included in their summary as they perceive it as noise. In addition to the differences between the healthcare professions, noise can also be perceived differently by different departments.

In the initial prototype, topic modelling is used as a technique to allow the healthcare professional to filter the topics that encapsulate noise out of their input before moving on to summarisation. Naturally, this approach contains a bit of an assumption and will be validated in the interviews (question 12 in Table 5.1). Topic modelling also tackles another problem; the long-text summarisation problem. As mentioned Chapter 4, the simple text summarisation models struggle with dealing with long texts. Furthermore, LLMs can also suffer from hallucination when the context window becomes increasingly long [32]. Using topics to group the text, the text can be divided into topics that relate to the underlying semantics of the text [48, 47]. These parts can then be used as a divide and conquer approach (such as [34]) to split the long text into multiple smaller parts, which can then be used as input for the summarisation step.

BERTopic [50, 51] for the topic modelling in the initial prototype, as it is a state-of-the-art method that utilises several embedding techniques in combination with c-TF-IDF. Furthermore, it can be greatly customised where necessary and allows for easy integration within a Python project. The BERTopic training step is unsupervised, eliminating the need for a large amount of labelled training data. A data set of patient forums posts is used focused on patients with pulmonary fibrisos that was created by previous research [52]. This data set contains posts in which patients discuss their illness and treatment with each other. We assume that these data are somewhat similar to patient interaction text documents, as they are both patient-centred and within the medical domain. After training the BERTopic model, it is able to output one topic per text input. We define the input as every five sentences of the entire input document, as this is the most similar to the average length of the post-data set.

The output of the training step of BERTopic does not produce human readable topics. Therefore, in order to obtain a good topic representation, the keywords of the topics (retrieved from BERTopic) are used as input to LLM along with the most representative documents:

Prompt 1 *(user)*

I have a medical topic that contains the following documents:
[DOCUMENTS]

The medical topic is described by the following keywords: [KEYWORDS]

Based on the information about the topic above, only reply with a short general label of
of this medical topic and nothing else.
Topic label:

5.2.2. LLM: text summarisation with no training data

Large Language Models (LLM) are great for numerous text-based tasks, including text summarisation. Even with limited or no training data at all, they can deliver promising results. This is a huge benefit within the context of this thesis research, as (Dutch) training data is extremely hard to come by or to generate within the time limits of the thesis. Unfortunately, such a data set was not available, also not by contacting researchers at Erasmus Medical Centre. Therefore, using LLMs with a zero-shot or a few-shot approach is a good option for this specific research scenario.

Chapter 4 argues that LLMs are the state-of-the-art method for medical text summarisation. Additionally, LLMs are especially suited for writing the output text, the summary, in a different perspective. This is an important property for the patient experience text summarisation system as it needs to transform conversational patient-doctor text towards a summary written for a healthcare professional specifically. Lastly, LLMs can also be easily finetuned based on training data which could be a huge advantage if it is necessary to train the summarisation for specific tasks, departments, or even users.

In the initial prototype, the LLM is used for text summarisation. As input, it takes the combined text of the input document per topic (divided by BERTopic) along with the topic representation to generate a summary per topic. Note that this will only be done for the topics that the user has selected in the intermediate topic modelling step.

Prompt 1 ***(user)***

I will give you some patient experience text and a topic that this text is about.
Respond with concise summary of 1 or 2 sentences written for a doctor.
Do not add any additional information which is not mentioned in the original text.

Patient experience text: [DOCUMENTS]
Topic: [TOPIC]
Summary for doctor:

After the result is retrieved, the different pieces are added to each other, resulting in the final summary. Vicuna [53, 54] was used for the initial prototype as the LLM implementation. Vincuna is an open source LLM which claims to perform almost as well as the ChatGPT models from OpenAI and can be privately hosted. Especially, the last step is important, as within the medical domain it is of paramount importance that data are transferred safely and not shared with non-related commercial parties.

Using the different parts discussed, BERTopic for topic modelling and Vincuna as an LLM implementation, the initial prototype and its flow are depicted in Fig. 5.1. The Streamlit framework (based on Python) [**streamlit**] is used to design the user interface and use the API of BERTopic (library) and Vincuna (hosted on a TU Delft machine). Naturally, this prototype can be refined and improved, which will be discussed in Chapter 7 after the discovery interviews have been analysed in Chapter 6. In Appendix B several screenshots can be found from the initial prototype.

5.3. Interviewee Selection

For this thesis research is important, especially the end users, healthcare professionals, of the text summarisation system. Therefore, these healthcare professionals within the medical domain are the main target for discovery interviews. Using the network of the thesis supervisor, it was possible to invite several healthcare professionals from Erasmus Medical Centre, which was a great opportunity as due to their extremely busy schedule it would be hard to plan interviews.

Several requirements were put in place before selecting healthcare professionals from the supervisors network to invite to participate in the interviews in order to ensure a good spread of interviewee attributes:

- In order for the interviews to be qualitatively be meaningful, at least 8 participants were required in total.
- Participants should be diverse in their level of experience and field of expertise (different disease). For both attributes there should be at least three different options, so at least three different expertise's and three levels of experience.

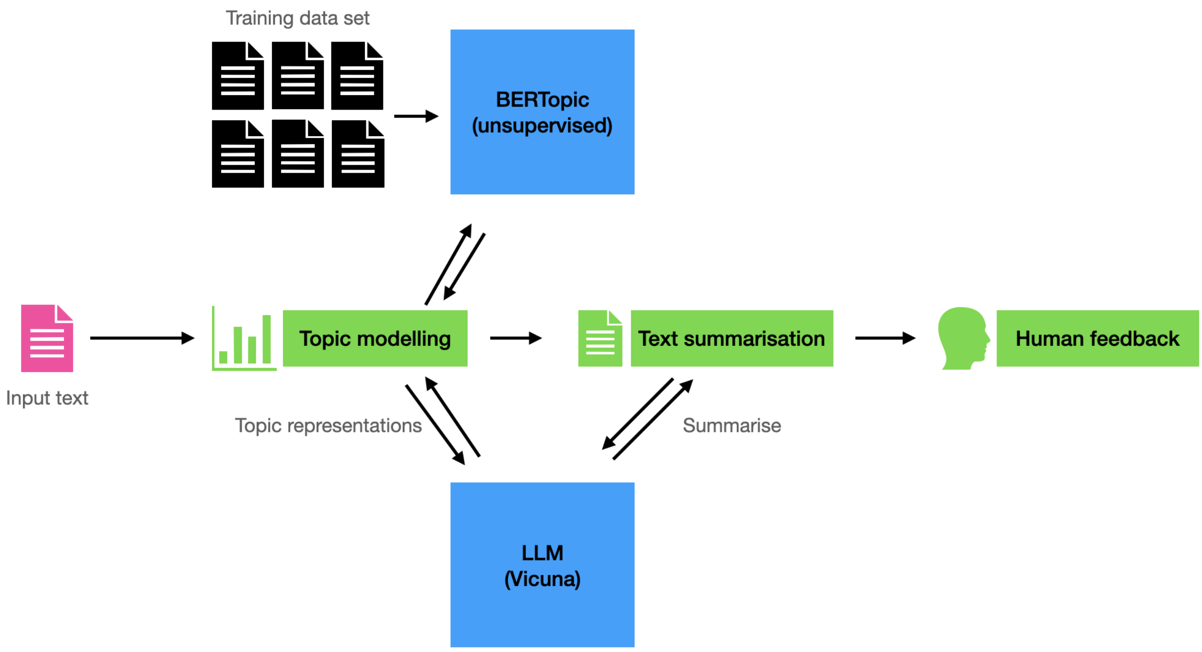

Figure 5.1: Different components of the initial prototype and its flow.

- Some of the fields of expertise of the participants should match the disease (pulmonary fibrosis) of the data set used to train the BERTopic model.

16 healthcare professionals were invited for a 45 minute interview based on the requirements listed above. Of those 16 participants, 11 accepted the invitation and a meeting was scheduled; either online or at the EMC. The overview of any identified participants can be found in Table 5.2 along with their role, expertise, years of experience, and attitude towards AI in healthcare (based on the perception of the supervisor from previous meetings). Note that all interviewee selection requirements are met.

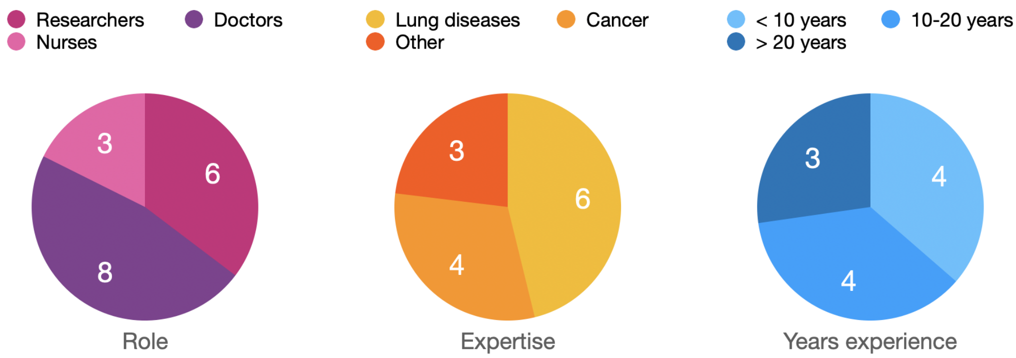

Figure 5.2: Spread of participants for first round of interviews.

5.4. Conducting Interviews

After sending the invitations, as mentioned in the previous section, 11 participants agreed on an interview date and time. A pair of healthcare professionals P6 & P7 wanted to have an interview with each other, which led to that only 10 meetings were scheduled. Most of the interviews took place at the EMC.

ID	Role(s)	Expertise	Experience	AI in healthcare attitude
P1	Doctor, researcher	Pulmonary medicine - interstitial and vascular pulmonary diseases	20+ years	Very positive
P2	Doctor, researcher	Oncological and gastrointestinal surgery	5 years	Positive
P3	Doctor	Oncological and gastrointestinal surgery - liver tumors, melanomas and sarcomas	12 years	Mildly positive
P4	Doctor, researcher	Oncological and gastrointestinal surgery - colon cancer, melanoma, sarcomas and liver cancer	25 years	Positive
P5	Doctor	Oncological and gastrointestinal surgery - thyroid gland, parathyroid gland, adrenal gland, neuroendocrine tumors, melanomas and sarcomas	7 years	Mildly positive
P6	Nurse specialist	Pulmonary medicine - interstitial pulmonary diseases	10+ years	Neutral
P7	Nurse specialist	Pulmonary medicine - interstitial pulmonary diseases	10+ years	Mildly positive
P8	Nurse specialist	Pancreas surgery, pancreatic cancer	9 years	Mildly positive
P9	Researcher, doctor	Pulmonary medicine - interstitial pulmonary diseases	11 years	Positive
P10	Researcher, doctor	Pulmonary medicine - interstitial pulmonary diseases	3 years	Positive
P11	Doctor, researcher	Pulmonary medicine - interstitial lung diseases, pulmonary fibrosis and sarcoidosis	20+ years	Positive

Table 5.2: Participants of the first round of interviews.

5.4.1. Interview Ethics & Privacy

The approval of the Human Research Ethics Committee (HREC) [55] was required to conduct the interviews as this research study deals with human subjects (a requirement of the TU Delft). Therefore, a HREC approval was requested by filling in all mandatory documents. A HREC checklist contained all the different risks related to this research study and a mitigation plan for each individual risk. Second, a Data Management Plan (DMP) was created and reviewed by the data wardens of TU Delft to ensure safe storage and processing of the data used. A consent form (see Appendix D for the interviewees, the last document) was also created to inform the participants about the ethics and privacy aspects of participating in the research study.

As the prototype used a dataset that contains data from human subjects (patients), another HREC approval was necessary to be able to use the data to train the BERTopic model. In this HREC request, the consent form was not necessary as this was already approved by the companies that developed the online patient forum. Both HREC letters of approval have been added in Appendix E.

5.4.2. Interview Procedure

Before the interview began, the interviewee was introduced to the content and scope of the investigation. If necessary, the interviewee was able to ask questions about the research. Then the consent form, from the HREC approval, was shown and the participant was asked if he or she agreed that the interview was recorded. In addition, the interviewee was told that their information, opinions, and views would be shared only anonymously in this thesis report.

After the participant signed the consent form, the interview started along with the recording. The rest of the interview procedure (and the questions) can be found in Table 5.1.

5.5. Interview Data

In total, 10 interviews with 11 interviewees were conducted to collect data on the first three research questions. As stated in the Data Management Plan, the data collected from these interviews is not stored forever, but deleted after the research is completed. The main outputs per interview are a .mp3 file of the recording, a .txt file of the raw generated transcript text, and a .pdf file of the edited, finalised, and reviewed transcript. This data is used as input for the following chapter in which the interview transcripts will be analysed by means of coding. All transcripts have been added (in Dutch) in Appendix F

6 Analysis

In the previous chapter, the way to conduct semi-structured interviews with stimuli has been discussed. The output of that chapter serves as the input for this chapter, where the interview data will be analysed using the coding methodology. The results of this coding process will then be used to create several system requirements which aim to provide an answer to the research question 1 *(user)* and 2.

The first section of this chapter will explain the methodology used for the coding process. Then, the coding process itself and the code groups and codes that were created will be described. Finally, the results will be discussed in Section 6.3.

6.1. Methodology

In order to gain meaningful insights from the interviews conducted, an analysis methodology is required. As there have been only 10 interviews with 11 interviewees, it makes sense to choose a qualitative methodology, as a quantitative approach would not be statistically significant. Therefore, for the analysis of the semistructured interviews, thematic coding analysis has been chosen. This qualitative technique aims to find common themes in the input data by taking a close look at the words and sentences appearing in those data [56]. After the thematic coding is done, the output will be a list of identified themes (or groups), each with corresponding quotes from the original data.

There are two main approaches to thematic coding; the deductive approach and the inductive approach [57, 56]. These approaches have a different starting point when it comes to defining the codes or code groups (themes). The first approach starts with a set of predefined codes and / or code groups that are assigned to the input data [58]. These codes can be found in the literature or originate from previous studies. The second approach creates (new) codes and code groups based on the input data. In addition, these codes could evolve over time and can be merged, split, created, and deleted over the course of the analysis [59].

In this thesis research, the inductive approach will be used as there are no pre-defined codes that can be applied to this research. Furthermore, by using inductive coding, the analysis can be done in a flexible way, specifically tuned toward the research questions that the semi-structured interviews were conducted for to answer.

6.2. Coding Process

Building on the decision to use inductive coding, as mentioned in Section 6.1, this section will describe the process and show the results of the coding process, consisting of code groups and individual codes.

As discussed in Chapter 5, each interview resulted in a transcript that accurately captured everything said in the interview. These transcripts were imported into the ATLAS.ti software application, which provides great tools to perform coding analysis. Furthermore, it also allows one to easily present or group the results in whatever way necessary. In this way, it was easier to create codes and code groups dynamically and merge them when necessary.

ID	Name	Description
CG1	(Un)willingness to use AI	Contains codes describing the sentiment of the medical professionals and patients towards the use of AI within the medical domain when it comes to patient interactions/documents.
CG2	Applications of AI summarisation	Includes the codes which apply to quotes where participants indicate they see use cases for AI summarisation when it comes to patient interactions/documents.
CG3	Differences in summarisation	Contains codes related to the differences between the way medical professionals create summaries in the current state and the differences in the content of summaries.
CG4	Goal of AI summarisation & fears	Codes which refer to the goals that participants think will be and/or should be achieved by the use of AI summarisation with regard to patient interactions/-documents.
CG5	Human validation	Codes which describe the view of participants on human validation and which role it should have in AI summarisation for patient interactions/documents.
CG6	Medium	Contains codes which refer to the different mediums the participants use for patient interactions.
CG7	Noise	Codes related to the amount or content of noise in patient interactions/documents.
CG8	Problem	Codes which describe the current problems medical professionals experience with regard to summarisation without a proper solution.
CG9	Properties of summary	Codes related to the properties that participants indicate a good-enough summary should have, mostly relates to the current state of summaries that they use.
CG10	Reinforcement learning	Codes which describe medical professionals talking about their views on how AI should be able to learn from the way they work.
CG11	Requirements	Includes codes referring to different requirements an AI summarisation tool should have. These are 'raw' requirements from the participants' perspective.
CG12	Task specific	Participants talk about the different tasks they use/need summarisation for. This group encapsulates codes that relate to this task specificness.
CG13	Topic flow	Codes related to how participants respond to the initial idea of having topics as a means of content selection. Also includes codes on whether and how they see use for this topic flow.
CG14	Use of summarisation	Codes which what medical professionals use summarisation for and how they apply it in these tasks/domains.

Table 6.1: The code groups derived from the clustering of codes.

6.2.1. Code Groups

After each transcript was analysed and the codes assigned, the process of creating code groups began. Codes that touched on the same topic were grouped together, although a code could belong to multiple code groups. The code groups aim to encapsulate important themes that were discussed during the interviews and group responses that belong to the same topic together. A good example of this is the first code group, CG1, where responses to the willingness to use an automated text summarisation system are grouped. Although they may be almost opposite (for example, C7 & C10, see Appendix A) they contain responses related to the same topic, which makes comparing these codes within this group interesting. All created code groups are listed in Table 6.1.

6.2.2. Individual Codes

As mentioned, the previously defined code groups are built on the first-identified codes. Defining individual codes was an iterative process in which a transcript was scanned multiple times and reviewed after finishing all transcripts. Often, new codes were created or codes were merged as they were extremely similar to other codes. Furthermore, the code names have been refined to be on a similar level of detail. All individual codes belonging to code groups can be found in Appendix A.

6.3. Analysis Results

Following the codes and code groups identified during the coding process, this section will dive into the actual analysis results that should provide answers to the first three research questions. First, the research question 3 will be answered using the results of the interviews in Subsection 6.3.1. Then Subsection 6.3.2 continues by defining the system requirements for a text summarisation system for patient interactions based on the interview findings. This will be used as the main input for Chapter 7.

From now on, the quotes of the participants will be used in this format: “Some quote here” (P1), where P1 refers to the participant and how he said this. These quotes will be used to illustrate the interview findings and highlight particular views and opinions. Note that all interviews were conducted in Dutch, so the quotes are translated to the best of our ability.

6.3.1. Noise

First of all, the interviews showed that noise is definitely present in patient interactions; “patient interactions are full of noise” (P11) and “a lot of noise, a lot of irrelevant information” (P8) are only a few of the numerous responses the interviewees gave. The codes C39, C40, and C41 show that noise is present through different communication channels of patient interactions and even in patient history documents. Furthermore, noise can be medical and non-medical information. As an interviewee pointed out, “a big range of information is shared ... I have talked about my own children with the patient... a toe surgery might not be relevant ” (P9), which highlights that noise is not limited to just non-medical information. It also immediately shows that the relevance of information is subjective and context-dependent. What is noise in one situation might not be noise in a different patient situation. In addition, what is noise for one healthcare professional might not be noise to another; “for me it is important to also write about the emotional well-being of the patient ... I know that others do not always write this down” (P2). This further emphasises that noise in the medical domain differs from noise in different domains, where it can be defined more objectively.

In addition to revealing that there is noise present, the interviews also showed that it increases the amount of time healthcare professionals have to spend finding / selecting the relevant pieces of information for summarisation; “substantial amount of time to process all the information and select the parts I need” (P11). This only complicates the time-consuming task of summarisation and further increases the work load. When asked about the percentage of noise present in the patient experience texts they have to summarise, the healthcare professionals interviewed answered ranged from 20% to 50%.

The participants agree on the fact that the summary should only include relevant information: “the main challenge is to include all and only relevant information” (P1). Furthermore, a large part of the summarisation is “identifying what is relevant and what is not” (P8).

6.3.2. System Requirements

This section looks at the eight main requirements of the system that were derived from the coding process. Each system requirement relates to one or multiple code groups/themes identified. In addition, the system requirements are based on a concrete finding from the interviews, which encapsulates the individual codes within the related code groups. For every system requirement, relevant or exemplary quotes have been selected and added to illustrate why this is important. An overview of all system requirements can be found in Table 6.2.

The system requirements have also been divided into functional system requirements, requirements that determine what the system should be capable of, and contextual system requirements, which dictate the context of the system in order to be useful. The first category includes six system requirements, and the latter contains two system requirements.

Summarisation tool must include ways to select relevant information and remove noise. Building on the previous section about noise in patient experience text, it follows that automated text summarisation would need to be able to remove noise from the data and select the relevant information. This is challenging given the fact that the noise can not be objectively identified and can differ between scenarios, healthcare professionals, or departments. The prototype part of the interviews showed that the topic selection intermediary step is liked by the participants, but should be evaluated further (which will be done in Chapter 8). However, this topic selection mechanism matches with the current way of summarisation; “it

ID	System requirement	Finding	Code group(s)	Quotes
Functional system requirements				
S1	Summarisation tool must include ways to select relevant information and remove noise	participants confirm that there is noise in all patient documents/interactions and that they always use selection / relevance when creating a summary. Noise (ir)relevancy is not objective.	Noise Topic flow Requirements Properties of summary	"Patients consultation sessions start with a lot of noise" "And the step I like is that you can filter it yourself. In my summary, I want this and that."
S2	Summarisation must include human validation	Participants indicate that they want to be able to edit/correct summary. They also express that they view AI as a tool that should be checked.	Human validation Properties of summary (Un)willingness to use AI Requirements	"Yes, I think human checkpoints are necessary." "That can still be adjusted by the doctor?"
S3	Summary must be somewhat traceable to information	Finding: Participants do not think the tool should be fully explainable, but indicate that it would be good to know how the intermediary steps map back onto the input data.	Requirements Topic flow Task specific	"That it is verifiable where the data comes from" "That you click on a kind of hyperlink and that you can then see, oh yes, he said there that he was in pain"
S4	Summarisation must include reinforcement learning	Participants handle summarisation in their own way. Structure is more or less the same, but length-/content can differ (per task).	Reinforcement learning Task specific	"That differs per colleague because you never write everything down." "Well, not so much in the essence of what is written, but in how extensive."
S5	Summary must be numerically factual correct	Participants indicate the importance of numeric facts (such as lab results) being correct in their summary.	Requirements Properties of summary	"...if a lung function is 2.3 liters instead of 3.2 liters, that is a problem." "But correctness is of course the most important."
S6	Summarisation must be task specific	Participants talk about a wide range of tasks they use summarisation for. These tasks handle different kinds of data and have different purposes.	Task specific Applications of AI summarisation	"The summary should have some kind of purpose. And the purpose determines what needs to be summarized."
Contextual system requirements				
S7	Summarisation tool must be embedded in current system	Participants have the feeling that tools which live outside of the current IT-ecosystem will not be used.	Requirements (Un)willingness to use AI Applications of AI summarisation	"every hospital has its own system" "What you should totally avoid is having to start new applications"
S8	Summarisation tool must be intuitive, time efficient and easy to use	Participants express that if the tool is too complicated or does not yield enough time benefits, it will not be used.	Requirements Goals of AI summarisation Problem	"That it was made for stupid people? ... it should be extremely intuitive." "...an absolute must have is that it saves time."

Table 6.2: System requirements based on the coding analysis of the interviews.

fits within our current work process " (P6) and " it is logical, this fits the way of selecting relevant information" (P1). As shown in previous chapters it also captures the underlying structure of text data.

Summarisation must include human validation. During the interviews it became clear that text summarisation should not be fully automated without human interaction. Healthcare professionals indicate that there are several reasons for this. First, lack of trust in AI solutions can cause (certain) healthcare professionals to not want to fully let go of control; “we might be not able to fully understand what is happening anymore” (P1). Another big reason is the need for responsibility, both from a legal perspective and from an ethical perspective. Healthcare professionals think it is only logical that they should validate the generated summary and be fully responsible for the contents; “I will always be responsible” (P4). Lastly, most interviewees indicate that it is only expected that a solution will not perform optimally at the start of the adoption and therefore human validation is needed especially at the start of use of the system.

Summary must be somewhat traceable to information. Even though the participants have quite some trust in the use of an automated approach, most of them indicate that it is necessary to have some level of understanding where the system is basing its output of. This does not have to be fully transparent, “it does not need to be fully explainable, but it would be good to be able to trace the data back when needed” (P5) , but should give some insight to the process and data the system takes into account.

Summarisation must include reinforcement learning. This system requirement is mainly based on the finding that participants have different ways of summarising. These differences do not only include the length of the summaries (C17), but also some minor structure differences (C19). “One doctor can write long stories, whereas the other keeps it very short” (P1). “Some limit themselves to the bare minimum of relevant information” (P1). Therefore, learning by user or department would be a huge advantage to gain the “learning effect” (P1). For healthcare professionals it would be necessary to see some “development of efficiency” (P8). As the summarisation is done by AI models which can be fine-tuned and learn over time, this system requirement has been formulated as reinforcement learning.

Summary must be numerically factual correct. This is one of the more obvious system requirements, but nonetheless crucial to the success of such a summarisation system. Healthcare professionals indicate that the summaries must be factual correct and express even more importance to the numbers in the summary; “...if a lung function is 2.3 liters instead of 3.2 liters, that is a problem.” (P11). As the initial prototype relies on LLMs, which can suffer from hallucination [32], it is of paramount importance to include this system requirement.

Summarisation must be task specific. The interviews showed that summarisation is used for a wide range of patient experience text, including patient history, consultation sessions, and phone calls. However, the summarisation differs in the level of detail and size & type of input data greatly between these tasks. Therefore, interviewees argue that the summarisation tool must not be a ’one solution fits all’-system, but rather allow for the user to apply summarisation given a specific task (context); “the summary has a certain goal, and the process of summarisation depends on the goal” (P5).

Summarisation tool must be embedded in current system. In order for the tool to be used by healthcare professionals, it should be easy to use the tool within the current software system they are using. Otherwise, it will be a big barrier to start using the tool as it will require additional training and resources. This is something all of the participants agreed on and it is key to the use and survival of a summarisation system; “it must be one integrated system in order for it to be used” (P3). This is a challenging requirement as different hospitals use different software systems. Furthermore, development within these various systems is extremely costly and slow.

Summarisation tool must be intuitive, time efficient, and easy to use: The last requirement touches on very basic properties that the system should possess; ease of use and the ability to save time. “The one thing it should absolutely do is save time” (P10), clearly illustrates the finding that this is the core reason that healthcare professionals will want to use the system. Of course, there are other reasons such as the dislike of doing manual less-intelligent work as indicated by interviewees, but being more time efficient is the number one priority. If this is not achieved it blocks the use of such a system; “if it does not save us enough time, I think we will not be using it” (P11).

7

Design

This chapter uses the output of the previous chapters to design an automated text summarisation system that adheres to the system requirements listed in Table 6.2. First, a selection of the system requirements will be made to implement in the prototype. Then, these selected requirements will be converted into actual goals; how the prototype should achieve these requirements. Then, the development of the prototype and its different components will be discussed.

7.1. Design Input: Selecting System Requirements

The requirements listed in Table 6.2 are the main findings of the interviews and are specifically related to the research question 1 *(user)*. These requirements are also the main input to improve the initial prototype and adapt it to the needs of healthcare professionals (research question 4.

Although all requirements were obtained from the interviews, not all of them are feasible/achievable within the scope of this thesis research. Therefore, this section will discuss the choices that have been made in regard to selecting certain requirements to be considered and implemented in the remainder of the research. The first and most important selection criterion was to select on the basis of the feasibility to implement the requirement with the available resources and time. For example, S7 is very important for the success of a text summarisation system in the 'real world'. However, it would be impossible to implement it within an EMC software system within the time frame of the thesis given that permission would be required from both the software provider and the hospital. Second, the importance and priority of the requirements were considered by reviewing the interview transcripts and identifying the requirements the participants felt most strongly about. This also relates to the relevance of the requirement to text summarisation. Lastly, the requirements were also selected that were absolutely necessary for a working system. For example, S4 was not required for a working system but would be a great extension of a working system. For each requirement, the list below indicates whether it has been selected or not. In addition, a small description has been written to explain the choice.

- ✓**S1:** As participants indicated, this is a very important requirement. This will be included in the development of the prototype.
- ✓**S2:** Considering the importance of responsibility and privacy in the medical domain, this will be included in the development of the prototype.
- ✓**S3:** Explainability is an important concept in AI model implementations. Although full explainability is not needed according to the interviews, this system required was selected as it also improves the trust users have in a system.
- ✗**S4:** Reinforcement learning would be a great addition to the prototype. However, integrating this did not fit within the timeline of the thesis. This requirement can easily be added as an extension later. However, preparation for this will be performed by already saving all necessary data to setup reinforcement learning.
- ✓**S5:** This touches closely to the issue of hallucination of LLMs. Therefore, a goal of the system will be to address this.
- ✗/✓**S6:** The interviews showed that the summaries can be very task dependent. To address all the different tasks is not feasible within the scope of this research. However, to still acknowledge the task specificity of the summarisation, consultation sessions will be the main focus of the prototype.

- ✗**S7:** This requirement is not feasible as it would require cooperation from both software provider and the hospital within the time frame of the feasible.
- ✓**S8:** This requirement is not a simple yes or no. The prototype will be tried to be designed in the most intuitive way possible and focus especially on the time efficiency.

Although certain system requirements have not been selected for further development in this thesis, the remainder of this research still keeps those requirements in mind when making design choices (no choices will be made which lock requirements out). Therefore, all requirements would still be able to be achieved given more time and resources.

7.2. Design Goals

Now that a subset of the system requirements of Table 6.2 have been chosen, this section will discuss how these requirements will be met in the final prototype. This basically defines the blueprint of the improvements that need to be made to the initial prototype. For each selected system requirements, a small description of the desired implementation is listed below.

- **S1:** The topic selection intermediary step will facilitate the removal of noise/irrelevant input to the summarisation step. As the topic selection follows the mental model from the healthcare professionals, this will be an intuitive way of removing the subjective noise from the input data.
- **S2:** The summary will be easily editable after it is generated. This will allow the healthcare professional to make whatever changes are necessary. In the future, this should result in less and less changes as the reinforcement learning property of the system will adjust the model to fit the style and structure of the user(s).
- **S3:** The topics can be traced to the input data; each topic can be inspected to show what text is present in this topic. This creates transparency to what data is inputted into the summarisation step of the system. Furthermore, it allows the healthcare professional to check whether certain information belongs to a certain topic.
- **S5:** Numerical sentences will be extracted from the input text (either written or actual numbers) and can be added directly to the summary and can be used to validate information in the generated summary. This creates an easy validation mechanism as the healthcare professionals can check certain information summary and add any important missing details as they wish.
- **S6:** Different tasks will be listed in the prototype where summarisation is used for patient interactions. The only implementation (for now) will be consultation session summarisation, which will be the focus for the prototype as it is the most illustrative example scenario where summarisation is needed.
- **S8:** The prototype will be designed in the most intuitive way possible and focus especially on the time efficiency.

7.3. Design Implementation

Using the selected system requirements and the concrete design goals that have been defined based on these requirements, this section will dive into the actual implementation of the improved prototype. Although this section is written in a linear way, the development of the prototype was not linear at all as it took many iterations to get the desired results.

This section has been divided into three different main components that make up the improved prototype. First, the used Large Language Model (LLM) used for the development will be discussed. Then, further development of the intermediary topic modelling step will be described. As the initial prototype did not perform well enough for this step, the underlying structure for this step was drastically changed. Next, since summarisation and topic modelling (see Subsection 7.3.2) both depend on the LLM and prompts to access the model, the prompt engineering process of the development will be discussed. Finally, the overall system architecture of the prototype and its different components will be shown.

7.3.1. Large Language Model

As mentioned in Section 5.2, Vincuna was used in the initial prototype as the LLM component. This was mainly to show that the system could be run with a privately hosted, open source LLM that could be finetuned in the future with medical data specific to a hospital. However, since the LLM was hosted on a shared machine, the performance related to speed was a bit disappointing, and sometimes only the 7B parameter model could be used instead of the 33B parameter model.

Although swapping out the LLM for any available (open source) model is still possible, for the improved prototype, we have chosen to use the newest OpenAI LLM model, ChatGPT-4 [60]. Using this model instead of Vincuna provided faster responses and resulted in better responses due to the fact that it has more parameters.

7.3.2. Topic Modelling

The interviews confirmed that the intermediate step of topic modelling made sense and followed the mental model of healthcare professionals and therefore will be continued to use in the development of the prototype. This step is used to remove irrelevant information/noise from the input data. As this notion of noise is subjective and can easily change within the medical domain, it is necessary that healthcare professionals make a conscious decision about what topics they want to include and which to remove. The initial prototype used a BERTopic implementation to train a topic model that could extract topics from the input data. The topic representations were generated by the LLM model based on the input from the topic model. However, the original BERTopic implementation resulted in some problems:

- **Too many topics for new input data.** By letting BERTopic train on a large dataset it can accurately cluster documents that are about the same topic. It will detect many of such clusters (250-300 for the patient forum datasets), which is an accurate representation when looking at the topics identified. However, when a consultation transcript is fed to the topic model it will assign different topics to most of the chunks of text which results in 20-30 topics per input document. This is undesirable, as also indicated by the interviewees as the topics should capture the essence of the transcript and not be a long list.
- **Undesired topic level, too general or very specific depending on training data + parameters.** This issue closely relates to the first issue. The topics that were identified for a transcript were too specific and did not capture the essence of the transcript. This is quite logical as the model does not consider the context of the transcript but the context of the entire dataset when assigning topics (i.e., the transcript is related to the known dataset). This results in a lot of mediocre topics. Tweaking the training parameters (specifically max number of topics) did limit the number of topics per transcript but resulted in very high level topics which are not useful when filtering noise.
- **New topics will not be recognised by the topic modeling, retraining necessary.** As the topics identified in the input data are based on the training dataset, it will not recognise/wrongly classify information that talks about a new topic. For example, if COVID related text is not in the training data, this topic will not be present in the output of the model for a new transcript which does include this topic. At best, the topic model might be able to classify it to a similar topic. To solve this, the training data would have to be kept up to date and the model should retrain every now and then. This poses an issue as it might costly to acquire training data over and over again and to retrain the model.

In order to develop a sufficient prototype that adheres to the system requirements and which users can actually use, these problems had to be addressed. After several advice talks with experts (from TU Delft) on topic modelling and the issues described, the use of LLMs was advised as an option to consider. This approach would require little to no training data and could easily be tweaked for the purpose of the research by performing prompt engineering. To support the choice of a topic modelling technique, four different approaches were considered: LDA [61] (a traditional topic modelling technique), BERTopic [50, 51] (a state-of-the-art topic modelling technique), Vincuna [54] (privately hosted open source LLM implementation) and ChatGPT-4 [60] (the latest / largest LLM model by OpenAI).

For the choice of the topic modelling technique, several aspects were considered to be relevant for the choice of the topic modelling technique. These aspects have been summarised in Table 7.1. Some of them have already been touched upon in this section, such as the need for training data, the number of topics, and the ability to recognise new topics. In addition, the aspects have been divided into three main sections; input, output, and process. The input contains aspects that are relevant to the input & setup of intermediary topic modelling step. Output relates to how the output of the topic model is defined and what additional actions need to be taken to present the topics to the end user. Lastly, the process encapsulates aspects that relate to the process of the topic-modelling technique. In Table 7.1 the reason why these aspects are relevant for the choice of topic is also described.

Aspect	Description	Relevancy for task
	Input	
Training data	Whether or not the approach requires training data and in what format/volume.	Training data is difficult & costly to obtain within the medical domain. Privacy restrictions play a great role.
Context	Whether the approach can take the context of a single new document into account. Does it recognise topics within their context?	Each patient/experience is different, the approach would need to account for this.
	Output	
Nr. of topics	How many topics does the approach output? Is this number static or dynamic?	Different tasks require different levels of topics and thus different number of topics. This is also required within a specific task.
New topics	If and how the approach recognizes/deals with new topics.	Topics might be different over time and new noise might enter the input. It is important that the approach can handle new topics.
Topic representations	How the approach generates topic representations for the identified topics. Can it be easily tuned towards users?	The topic representations are important as they should match the expectations of the user.
	Process	
Finetuning	How the approach can be finetuned towards better performance for the particular task and context.	The approach might need to be fine-tuned towards the medical domain with the aim of being an intermediary step for summarisation.
Speed	How long does it take for the approach to come up with topics? How long does it take to setup this approach?	Although it is partially a background process, the topics should be generated in a timely manner as the tool is aimed at being time-efficient.

Table 7.1: Different aspects considered relevant to evaluate different topic modelling approaches by.

For each of the aspects in Table 7.1, different topic modelling techniques were investigated. This led to the comparison shown in Table 7.2. This resulted in the choice to use ChatGPT-4 as the model to perform topic modelling. This approach allows to build a dynamic approach which can easily be reused in a different context and adjusted to the needs of the end user. It addresses both the issues of the number of outputted topics and the identification of new topics. However, when the system is implemented in a real-world scenario, an open-source model such as Vincuna should be used, as it can be privately hosted and fine-tuned to a better performance than the default ChatGPT-4 model. The new structure/architecture of the different components of the summarisation pipeline is depicted in Fig. 7.1.

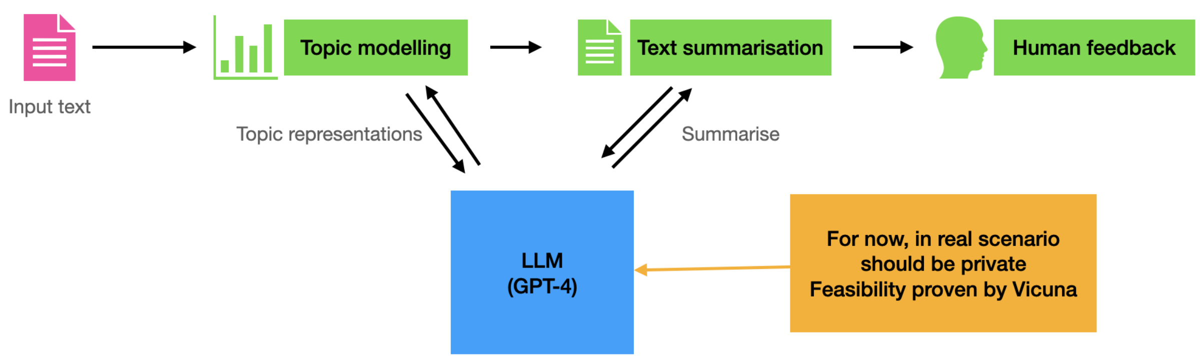

Figure 7.1: Structure and components of the improved prototype.

Approach	**LDA [61]**	**BERTopic [50, 51]**	**Vincuna (LLM) [54, 53]**	**ChatGPT-4 (LLM) [60]**
Description	Latent Dirichlet allocation, a probabilistic approach to assigning a topic to a collection of words.	Approach that uses transformers and c-TF-IDF to create topic clusters from a set of documents.	Open-source LLM based on Llama and further trained with examples from ChatGPT-4, has a 7B- and 33B-parameter model.	The latest and biggest LLM of OpenAI, not open source. More than a trillion parameters.
		Input		
Training data	Requires set of documents which will be used to create probabilities	Requires a set of documents which are used to construct clusters	Can be finetuned with training data, but can also performs well with zero-, one- or few-shot prompts	Same as Vincuna
Context	Considers overall context of documents, which defines specificity	Same as LDA	Has the option to consider only one document, which gives great context specificity	Same as Vincuna
		Output		
Nr. of topics	Requires a number of topics as input before training	Can be dynamically determined by the model, maximum number can be set	Can be determined by prompt (chaining)	Same as Vincuna
New topics	Will not identify new topics unless trained again with new data	Same as LDA, not suited to train on only few documents	Will identify new topics as it can easily consider only one document easily	Same as Vincuna
Topic representations	Interpreted/labeled by human or LLM	Interpreted/labeled by human or LLM	Output is already representation, can be fine-tuned by prompt engineering	Same as Vincuna
		Process		
Fine-tuning	Parameters can be tuned (number of topics and learning parameters)	Parameters of components can be fine-tuned & components can be swapped for different strategies/approaches	Can be done by prompt engineering or fine-tuning model for this specific task	Same as Vincuna
Speed	Only training step is time consuming, fetching topic for new document is fast	Only training step is time consuming, fetching topic for new document is fast	No training required, fetching topic rather slow on local machine	No training required, fetching topic is fast (but slower than LDA and BERTopic) due to the use of OpenAI API (*chatgpt-4-turbo*)

Table 7.2: Different approaches considered for topic modelling.

7.3.3. Prompt engineering

This section will discuss the prompt engineering process that has resulted in the final prompts, displayed later in this section. For prompt engineering, prompt engineering techniques have been considered listed in Table 7.3.

From Table 7.3, four prompt techniques have been used during the development of the improved prototype. This section will now further discuss the prompt engineering process done for both the topic modelling part and the summarisation part of the improved prototype. The evaluation of this prompt

Technique	Description	Goals
Zero-shot Prompting [62, 63]	Prompt contains only instruction without any demonstration or example(s). Instruction tuning is the main strategy that is utilised in this technique.	Clearly instruct the model what to do, keep prompt simple, rely on performance/-knowledge of model
Few-shot Prompting [64, 65, 66]	Add examples of desired outputs to prompt to stimulate in-context learning. Number of examples can vary and can be added as much as necessary.	Helps to solve more complex tasks, define the format of the output, define labels/output, add context to prompt
Chain-of-Thought Prompting [67, 68, 69]	Add intermediate reasoning steps/thoughts in the examples provided or ask the model to explain its reasoning in a zero-shot approach.	Helps to solve more complex tasks, especially arithmetic tasks where intermediary are necessary
Self Consistency [70]	Sample multiple reasoning paths (chain-of-thought) and find the most consistent response out of the set of answers.	Improves arithmetic and commonsense reasoning (more complex)
Generated Knowledge Prompting [71, 72]	First generate knowledge (by using the LLM) which can be used in the final prompt to infuse the model with knowledge.	Improves commonsense reasoning where more knowledge about the world is required
Prompt Chaining [39]	Split a (complex) task into multiple sub tasks (prompts). By simplifying the bigger prompt the goal is to improve performance	Aims for better results by splitting the bigger task into multiple sub tasks
Tree of Thoughts [73, 74, 75]	Let the LLM self-evaluate the intermediate steps/thoughts it generates and trace down the tree for the optimal response.	Improves performance for complex tasks that require exploration or strategic look ahead
Retrieval-Augmented Generation [76, 77]	Add additional (dynamic) knowledge to the LLM by accessing a set of external, relevant knowledge documents	Improves performance for more complex and knowledge-intensive tasks, mitigates hallucination

Table 7.3: Prompt engineering tasks considered for this research.

engineering is relatively basic and will be conducted qualitatively to some extent in Chapter 8. This is mainly due to the fact that the requirements for the output of the prompts are neither objective nor binary. The 'correctness' of the output is context dependent and can not be measured numerically easily, especially when lacking labelled evaluation data. Therefore, the prompts described in this section could be further refined in future research, but this research focusses on the important decisions made in the prompt engineering process based on the requirements and findings of the first round of interviews.

To test the different prompts, a realistic transcript of an imaginary lung disease consultation session has been created (and validated by healthcare professionals; see Chapter 8). This transcript was used to change the prompts and observe the difference after improvements based on prompt engineering techniques. The results were also used to find the appropriate prompting technique for the problem at hand. After applying these techniques and being satisfied with the effect on the realistic transcript, the change was further validated by testing several transcripts from the diverse data set from the PriMock57 mock consultation sessions [45].

Topic modelling prompting
As the topic modeling of the initial prototype relied on BERTopic, there was no initial prompt yet for the topic modeling. Therefore, the development for this part of the system was started with a simple zero-shot prompt to obtain the topics from the input data:

Prompt 1 (v1) ***(user)***

I will give you some patient experience text from a consultation session.
Respond with a comma separated list of topics that this consultation session is about.

Patient experience text: [TRANSCRIPT]

Topics:

Example result: "pulmonary fibrosis, symptoms management, trip to Spain, quality time with granddaughter, lab test results, lung function decline, medication regimen, new inhaler, long-acting bronchodilator, inhaled corticosteroid, inhaler instructions, inhaler side effects, exercise routine, diet, cold weather challenges, self-care strategies, follow-up appointment"

This prompt resulted in too many topics, although the topics were identified accurately (but too detailed). Therefore, instruction tuning was used to further specify the output requirements.

Prompt 1 (v2) ***(user)***

I will give you some patient experience text from a consultation session.
Respond with a comma-separated list of topics related to this consultation session. The maximum number of topics you can return is 10, try to stay below this number as much as possible.

Patient experience text: [TRANSCRIPT]

Topics:

Example result: "pulmonary fibrosis symptoms, travel experiences, family bonding, symptom fluctuation, medication adjustment, new inhaler, side effects, exercise routine, cold weather impact, follow-up appointment"

Although this prompt performed well and created good results, there were several issues/requirements that could not be addressed by using this prompt. As the system requirement (S3) demanded that the intermediary step of the summarisation should be traceable to the input data, each piece of text should be assigned to a topic. This fact, in combination with the possible limits on the context size of the LLM (with Prompt 1 we assumed that this would not exceed this limit), resulted in the task of topic modelling being split into several subtasks. Therefore, prompt chaining was used to facilitate this, as this is a technique that is used to split up a problem into multiple sub-problems that can be handled by one prompt. In this way, the 'bigger' prompt was split up into multiple smaller prompts that each resulted in one piece of text being assigned one topic. Naturally, this also meant that the input data had to be split. Because the input data that were focused on were transcripts, every time a different person said something was used as the main divider (e.g. Doctor: "...", Patient: "..."). These parts were bundled together in a chunk of three, which was then used as input for this prompt:

Prompt 2 (v1) ***(user)***

I will give you some piece of patient experience text from a consultation session.
Respond with a topic that this piece of transcript is about.

Patient experience text: [PART OF TRANSCRIPT]

Topic:

Example result: "Medication side effects and expected time for symptom improvement"

By simply linking these prompts together, a list of topics could be constructed that captured the different parts of the transcripts. However, if the transcript became longer, the topic list did, as well as it was not merged. Furthermore, the same topic could be described in different ways, resulting in unnecessary long lists of topics. Therefore, context was added to the chain of prompts and the results of the previous prompts. This was done to instruct the LLM to reuse topics if possible and to be able to get some of the context of the entire transcript.

Prompt 2 (v2) ***(user)***

I will give you some piece of patient experience text from a consultation session, along with already identified topics.
You will reply with one of the existing topic labels OR a new topic label if the input does not match any of the given topics.
Always try to choose one of the existing, only respond with a new topic label if necessary

Patient experience text: [PART OF TRANSCRIPT]

Topics: [ALREADY FOUND TOPICS]

Example result: "New topic label: [Medication Side Effects and Symptom Improvement]"

Using the results of these prompts, it was possible to build a list of topics that the LLM reused when necessary. However, as can be observed in the above example, the output structure is a bit messy because the output format is not clearly defined. Furthermore, the topic level was inconsistent; sometimes very detailed topic labels were generated when other times the topics were very general. Lastly, sometimes the topics were not described in a medical fashion to suit the end user (i.e., not feeling well instead of 'symptoms'). To combat the first issue, the prompt instruction was further fine-tuned with the output format. In addition, a few-shot prompting was used by providing some examples. This also specifies the output format and influences the topic level and the vocabulary used in the responses. This resulted in these prompts being chained together, where the last prompt is used again and again to retrieve all the topics. As only a piece of the input data is used for every topic retrieval, the topics can be easily mapped to the input data, adhering to the design goal of S3. The qualitative evaluation of these prompts will be done in Chapter 8.

Prompt 3.1 ***(system)***

You are a topic modelling tool used by doctors after they concluded a consultation session.consultation session produced a transcript of dialogue between the patient and the doctors.

Prompt 3.2 ***(system)***

You will be fed parts of the transcript as input along with a list of already identified topic labels. For each input, you will reply with one of the existing topic labels OR a new topic label (plain text, no prefix) if the input does not match any of the given topics. Always try to choose one of the existing, only respond with a new topic label if necessary.

Prompt 3.3 *(system)*

Example input:
Doctor: I understand. It's not uncommon for pulmonary fibrosis symptoms to fluctuate. Before we dive into your current health status, tell me about your trip to Spain. How was it?
Patient: Spain was amazing, Doctor. The weather was perfect, and I had a chance to relax and enjoy some time away from the usual stressors. It was truly rejuvenating. Well, you see, it was her first time traveling abroad, and I wanted to make it special. We explored the streets of Barcelona, marveled at the intricate architecture of Gaudí's masterpieces, and indulged in delicious tapas at quaint little cafes tucked away in hidden alleys.

Example output:
Holiday

Prompt 3.4 *(system)*

Example input:
Patient: Understood. Are there any common side effects I should be aware of?
Doctor: Some common side effects may include headaches, throat irritation, or a dry mouth. These side effects are usually mild and tend to improve with continued use. However, if you experience any severe or persistent side effects, please let me know right away.

Example output:
Medication side effects

Prompt 3.5 *(user)*

Input:
[INPUT]

Topics:
[ALREADY FOUND TOPICS]

7.3.4. Summarisation prompting

As mentioned in Section 5.2, for the initial prototype, zero-shot prompting was used to generate the summary. The summaries generated by this prompt did not meet the needs of healthcare professionals, which was noted by almost all participants. The main issue, according to the interviewees, was the general nature of these summaries; the summaries were not written for healthcare professionals. Therefore, some context was added to inform the LLM about the goal and input of the task. Although this is still not specified on a, for example, hospital department level, this prompt can easily be tweaked to add department/illness related information to further instruct the model.

Prompt 1.1 *(system)*

You are a summarisation tool used by doctors after they concluded a consultation session.
The consultation session produced a transcript of dialogue between the patient and the doctors.

Prompt 1.2 *(system)*

The transcript has been divided into pieces and clustered by topic. You will be given as input one of these topics along with the pieces of the transcript that correspond to this topic. As output, respond only with a short summary which the doctor can use as notes and administration.

Prompt 1.3 ***(user)***

Topic:
[TOPIC OF TEXT]

Input:
[TEXT]

Example result: "The doctor prescribed a new combination inhaler for Mr. Anderson, containing a long-acting bronchodilator and an inhaled corticosteroid, to improve symptom control by relaxing airway muscles and reducing inflammation. Instructions include deep inhalation once or twice daily and rinsing the mouth after use to prevent throat irritation or oral thrush. Common side effects include headaches, throat irritation, and dry mouth, with a recommendation to report any severe or persistent side effects."

The summaries produced by this prompt were still relatively long, and thus some extra instruction tuning about the length of a subsummary was used.

Prompt 2.1 ***(system)***

You are a summarisation tool used by doctors after they concluded a consultation session.
The consultation session produced a transcript of dialogue between the patient and the doctors.

Prompt 2.2 ***(system)***

The transcript has been divided into pieces and clustered by topic. You will be given as input one of these topics along with the pieces of the transcript that correspond to this topic. As output, respond only with a very concise summary (of maximum 2 sentences) which the doctor can use as notes and administration.

Prompt 2.3 ***(user)***

Topic:
[TOPIC OF TEXT]

Input:
[TEXT]

Example result: "Inhaler efficacy decreasing, prompting prescription of new combination inhaler for better symptom control. Discussed proper inhaler usage and potential side effects."

Furthermore, some examples (few-shot approach) were added to the prompts, as this further ensured consistency in the summaries, along with the style of writing that healthcare professionals use for their summaries.

Prompt 3.1 ***(system)***

You are a summarisation tool used by doctors after they concluded a consultation session.
The consultation session produced a transcript of dialogue between the patient and the doctors.

Prompt 3.2 *(system)*

The transcript has been divided into pieces and clustered by topic. You will be given as input one of these topics along with the pieces of the transcript that correspond to this topic. As output, respond only with a very concise summary (of maximum 2 sentences) which the doctor can use as notes and administration.

Prompt 3.3 *(system)*

Example input:
Topic: Medication side effects

Input:
Patient: Understood. Are there any common side effects I should be aware of?
Doctor: Some common side effects may include headaches, throat irritation, or a dry mouth. These side effects are usually mild and tend to improve with continued use. However, if you experience any severe or persistent side effects, please let me know right away.
Patient: Will do, Doctor. How long will it take for me to notice any improvements in my symptoms?
Doctor: It may take a few days, I think three days, for you to start noticing the full benefits of the medication. Everyone responds differently, so it's essential to be patient and give it some time to work. In the meantime, continue to monitor your symptoms and let me know if you have any concerns.
Patient: Thank you for clarifying, Doctor. I appreciate your thorough explanation. I'm feeling more confident about trying out this new inhaler.

Example output:
Described medication side effects; headaches, throat irritation, or a dry mouth. Likely three days until improvement in symptoms.

Prompt 3.4 *(system)*

Example input:
Topic: Medication adjustment

Input:
Doctor: OK. Let's focus on your cough first. Can you tell me a bit more about your cough?
Patient: Uh, it's just so continuous. It's really, really frustrating. It's quite dry. And it's just all day, it like keeps me up at night as well, which is really annoying.
Doctor: OK. So it's a dry cough. So you're not bringing up any , you're not bringing up any mucky, ohh I can hear that, yeah. Um you're not bringing up any uh, mucky phlegm or anything like that.
Doctor: No, OK. So it keeps you up at night. Um you mentioned uh the cold symptoms as well, the runny nose. Do you have a do you have a sore throat at all?
Patient: I mean, apart from, I mean the coughing is giving me a sore throat. But it's not like, you know not more than, than that. You know just, just my throat is irritated from coughing. But it's not like, it's not that bad when I swallow or anything.

Example output:
Continuous, dry, cough. No phlegm, but sore throat and disturbs patients sleep.

Prompt 3.5 *(user)*

Topic: [TOPIC OF TEXT]

Input:
[TEXT]

7.3.5. System Architecture

This section will briefly discuss the system architecture that was used for the improved prototype. The initial prototype only used the Python-based framework Streamlit, as it could encapsulate all necessary libraries and perform the communication with the LLM (Vincuna). Furthermore, it provided a time-efficient way to build the user interface. Although this would also be great for the improved prototype, some of the requirements (i.e., data persistence) and design wishes (i.e., stepper functionality) could not be met by Streamlit as it was not built for this purpose. Therefore, several changes were made to the system architecture to be able to implement the selected system requirements.

Flask + SQLite

First of all, the system needed to be able to persist data as a start for reinforcement learning. Furthermore, it would be a great addition to transparency and human validation. To this end, SQLite was used as a simple database for persistence of the information stored in the system. As a database would need to be accessed, it would be easy to have a backend that would support this and perform business logic. Flask was used for the backend of the prototype, as it is easy to set up and can make use of all the needed Python packages.

Angular

Angular was used as the frontend framework of the prototype. This was mainly due to the knowledge that I already had about this framework, which allowed for fast development without the need to learn a new framework. Using this frontend framework allowed for more complex user interface functionality and also a good way of connecting to the Flask backend.

8 Evaluation

This chapter discusses the evaluation used to validate the findings and development of the prototype/system. First, the methodology used for the evaluation is explained. The results of this evaluation are then presented.

8.1. Methodology

To evaluate the results of the interviews, the development of prototype expert interviews with healthcare professionals have been used. Due to the qualitative nature of the previous round of interviews and the availability of the healthcare professional network, another round of semi-structured interviews was chosen. Furthermore, as the participants have seen the initial prototype, they can actually validate whether their feedback and views on the system requirements have correctly been taken into account. The main goals of these interviews are to:

- validate the topic level/mechanism; does the topic level match the expectations/needs of the user and are they able to filter noise from the input data?
- validate the usefulness of the summary; does the quality suffice for it to be used?
- validate the intuitiveness of the summarisation pipeline; does the design facilitate a time-efficient workflow?
- validate the system requirements; are they accurate and not redundant?

8.2. Interview Guide & Participants

To conduct semi-structured expert interviews, another interview guide was created and is illustrated in Table 8.1. This interview guide has been created using the same structured approach as in Section 5.1 [29]. First, the research is briefly introduced again and the goal of this evaluation interview is made clear. Then a short set of questions is used to identify the healthcare professional. Subsequently, the new improved prototype is introduced and the different components are shown. Then, the participants are asked to use the prototype themselves (with mock consultation session transcripts) and answer several questions about the flow and performance of the different steps in the system. In addition, they are asked to test the prototype with a self-written transcript. After this part is completed, several questions about the system requirements of Table 6.2 are asked to validate those requirements and obtain additional feedback. This last section has been refined to be optional as healthcare professionals might not have enough time to stay around for this part of the evaluation interview. Lastly, the interview ends with the participant thanking them and asking if they have any last remarks or questions.

This interview guide was used for a subset of the healthcare professional interviewed in the first round of interviews. In total, six healthcare professionals were interviewed; P1, P2, P3, P5, P10, and P11. A brief description of these participants can be found in Table 5.2. The transcripts produced by the evaluation interviews can be found in Appendix G.

8.3. Results

This section will discuss the results of the evaluation interviews. Like in the first round of interviews, thematic coding analysis was applied to the transcripts of the evaluation interviews to obtain meaningful

Interview guide
Introduce goal of summarising patient experience to reduce the workload on healthcare professionals. Mention that this summarisation is topic-based.
Mention that there already has been a round of interviews earlier in this study that investigated the needs and barriers to using an AI system for patient experience summarisation.
This interview will be used to validate the findings from the first round of interviews. Furthermore, the prototype will be evaluated by letting the participant use and test it in comibination with asking questions about the prototype to the participant.
Show and explain the consent form. After that, start recording.
Identify healthcare professional (1 min.)
1: What is your main field of expertise?
2: How many years of experience do you have within this field?
Introduce prototype (4 min.)
Introduce the prototype to the participant to be able to discuss it later. Mention with what goals in mind it was built and that the first round of interviews have been used to improve the prototype. Explain the flow of the summarisation process (input, topics, summary).
Discuss prototype (20 min.)
Introduce the prepared example; what kind of transcript it is and which topics are prevalent in this transcript. Demonstrate the prepared example. Clarify where necessary.
Let the participant write a small transcript themselves. They can then proceed to use the transcript as input for the prototype and try out the entire flow.
3: Do you think the topic level suits the purpose of summarisation and matches your expectations? • Why do you think this is the case or should be different? • Were you able to filter out the noise of your transcript?
4: Do you think the design of prototype is intuitive to use? • Why do you think this is the case or should be different? • What could be improved to make the prototype more user friendly?
5: Do you think the quality of the summarisation is sufficient to be actually used? • Why do you think this is the case or what should be improved to be sufficient? • Does this also hold for other colleagues, in your opinion?
6: Do you think there is something missing for the prototype to be able to use it? • How can this be improved? • Why is that important in order to be able to use it?
System requirements (5-10 min. (only if possible))
Go through each system requirement (S1 up to S8).
7: Do you strongly disagree, disagree, neither disagree or agree, agree or strongly disagree with this system requirement? • Why do you think this is necessary? • Why do you think this is not necessary? • Is this requirements too specific? • In what way does this requirement match (or does not match) within the domain you work in?
8: Do you think there are requirements missing?
Closing remarks (4 min.)
Wrap up the interview, ask for any last remarks. Thank the participant for their time.

Table 8.1: Interview guide for evaluating system requirements & prototype.

ID	Name	Description
ECG1	Topic evaluation	Contains codes describing the response of participants towards the intermediary topic step. Includes codes about the quality of the topics and the ability to filter out noise.
ECG2	Summary evaluation	Includes the codes which apply to the quality of the summary outputted and its structure, and suggestions to improve summary generation.
ECG3	Intuitive design evaluation	Contains codes related to the ease of use of the system and the flow of the summarisation steps.
ECG4	Transcript validation	Codes which refer to validity of the mock consultation transcript for the evaluation interviews.
ECG5	Requirements evaluation	Codes which describe the view of participants on the created system requirements derived from the first round of interviews.
ECG6	Contextual aspects	Contains codes which encapsulate views of the participants on (the use of) the system in its environment.
ECG7	Reinforcement	Codes which describe the views of the participants on possible use of reinforcement in the summarisation system.
ECG8	Valuable additions	Codes describing suggestions of the participants for further improvements/additions to the prototype or different use cases of the system.

Table 8.2: The evaluation code groups derived from clustering the codes.

results. The code groups will first be discussed in this chapter. Then, the key findings will be highlighted that were derived with the help of the code groups.

8.3.1. Code Groups

The code groups found in the outputted transcripts of the interviews roughly correspond to the sections of the interview guide inTable 8.1. However, some code groups encapsulate some specific information that was discussed during the interviews, such as reinforcement learning and contextual aspects of the prototype. Therefore, these have also been added. A complete overview of the code groups used can be found in Table 8.2, and the individual codes can be found in Appendix A.

8.3.2. Key Findings

First of all, it is important to note that the participants did not have time / motivation to write a transcript of their own, which was asked of them in the third section of the interview guide. Often they were too busy during the week and did not feel the need as they could see that the system was working with the mock consultation session transcript. Therefore, to compensate for the lack of transcripts provided from the healthcare professionals' side, the mock consultation session was also briefly validated by the participants to confirm its validity and make sure it was a realistic scenario. The results from this have been captured in the code group Transcript Validation (CG?). All participants agreed that the mock consultation session transcript is realistic enough compared to what they encountered to be able to evaluate a summarisation tool. Some participants identified some minor discrepancies such as the mention of a normally not prescribed medicine, but expressed that this is not relevant to the summarisation tool evaluation.

The key findings of the interviews have been summarised in Table 8.3. These are the findings that stood out when conducting the evaluation interviews and were present among at least four of the six participants. These mainly touch on further exploring the problem of automated summarisation in the medical domain and provide valuable lessons for further development and research in this domain. They have been divided into three categories: evaluation findings, requirement findings, and implementation & context findings. The first touches upon the general evaluation of the prototype in its problem context. The second focusses on the evaluation of the system requirements derived from the first set of interviews. The last category contains findings on how the system would need to be implemented and possible use cases. Each of the findings will be discussed in more detail.

Ability to remove noise with topic selection. The participants all indicate that by using the intermediary topic selection step with the generated topics they are able to remove the irrelevant noise from the transcript. Furthermore, they indicate that this step is intuitive and easy to use. The topic descriptions are good enough, but could be more tailored towards the medical domain (e.g., use more medical vocabulary).

Summary captures essence. The participants agree that the summary includes the important information of the input in the output. This further shows that the topic selection step is sufficient to not only filter out irrelevant information but also ensure that important information is not lost.

Intuitive summarisation steps. The flow of the summarisation tool is intuitive and easy to use according to participants. It should not be made more complex but also none of the components is redundant. The intermediary topic selection step allows participants to deal with the subjective noise in the input. The numerical sentences steps is perceived as an important addition to the creation and validation of the summary.

Summary structure. The summary does indeed encapsulate the essence of the input but structure of the summary can be improved according to participants. Firstly, the summary can be much shorter as this allows healthcare professionals to scan the summaries quicker. In addition, the summary would benefit from being formatted as bullet points as this format is preferred by the interviewed healthcare professionals. These structure changes could easily be incorporated into the prompt engineering process for the system.

Traceability of topics; *nice to have*. Although the first round of interviews showed that traceability is important to some extent, the participants indicate that this is more of a *nice to have* instead of a hard requirement for the system. Many find this extra feature helpful and not complicating the system, but indicate that it would not be a disaster if the system were to launch without this feature. Furthermore, they argue that this feature will be used less over time as users trust the system more.

Reinforcement learning; collective/individual. Although all interviewees agree on the need for reinforcement learning in the system, they express different views on the nature of reinforcement learning. Collective reinforcement learning creates structured, uniform data and enforces a standard summarisation process for all healthcare professionals. However, this could complicate the adoption process as the time investment at the start will be bigger. Individual reinforcement learning is the most time efficient, but will not create overarching structured data. Most interviewees believe collective reinforcement learning is the best option but do worry about the adoption process. This finding is not limited to the scope of summarisation but to the use of AI systems in organisations where people have different ways of working in general.

Summary usage. Although participants are mostly interested in the system replacing their manual summary writing process, they see a lot of possible use cases for the summarisation tooling. For example, a slightly modified version of the summary could be used to send in a letter to the GP or to hand over to the patient which they could use to share with their loved ones. In addition, the summarisation tool could be used as a form of agreement between the doctor and patient on what has been discussed and to assure the patient that the doctor has not missed anything.

Embedded not strictly required in pilot. Naturally all participants would like to see summarisation tooling implemented in their current system. However, during a pilot it is necessary for success to implement it in the current system due to the regulations and costs. Furthermore, not having it in the system during the pilot could convince the software provider to integrate it quicker and more easy.

Besides the validation of topics, summary, intuitive design of the system, and the requirements, additional feedback was collected about the (idea of) the system. This resulted in several suggestions for future research or improvements to the system. These ideas have mainly been captured in the code group Valuable Additions (ECG8), but have not been touched upon in Table 8.3. The reason for this is that each of these ideas/improvements was not suggested by more than two participants. Each of these points will be briefly discussed.

- **Non-binary selection of topics.** The intermediary topic step is a good way to filter noise from the input according to participants. However, in some cases they might want to include only two words about a topic as it is not very important but still worth mentioning in the summary. For example, the healthcare professional would like to note down that the patient will go on holiday to ask about the holiday the next time. However, they do not want a summary of everything about the topic holiday, just a very brief mention in the summary. Currently this would only be possible by indicating that it is a relevant topic and thus getting a bigger summary. Therefore, the choice could be non-binary meaning that there would be more than two choices (e.g., no/a little/yes rather than yes/no).

ID	Name	Finding	Code group(s)	Quotes
Evaluation findings				
KF1	Ability to remove noise with topic selection	Participants indicate that they are able to remove noise from the transcripts and that this works intuitively.	Topic evaluation	"I can apply a filter of what is relevant and what is not"
KF2	Summary captures essence	The summary generated by the system captures the important information from the input according to participants.	Summary evaluation	"the summary is good ... it captures the important information"
KF3	Intuitive summarisation steps	The participants like the summarisation steps that need to be taken as it gives them control and addresses the subjective nature of noise in the medical domain.	Intuitive design evaluation	"it is good to have this intermediary step", "the summarisation flow is logical"
Requirement findings				
KF4	Summary structure	The summary generated by the system needs to be shorter and would benefit from the use of bullet points.	Summary evaluation	"bullet points help you see the important information at a glance", "more short sentences, like a telegram"
KF5	Traceability of topics; *nice to have*	The participants express that mapping the topics back to the input text is helpful but not a requirement when the system is used in a real scenario.	Topic evaluation Requirements evaluation	"I think it is a great extra tool, but not a must", "after people trust the system they will likely not use it anymore"
Implementation & context findings				
KF6	Reinforcement learning; collective/individual	Participants express different views on the nature of reinforcement learning. Collective creates structured uniform data, but is a risk to adoption. Individual is the most time efficient, but will not create overarching structured data.	Reinforcement learning	"It must learn, but collectively and not per individual", "collective learning ... it will be difficult for some colleagues"
KF7	Summary usage	The summary generated by the system might be long for the healthcare professional but could be used for communication to the GP and/or patient.	Summary evaluation Valuable additions	"a summary that can be shown to the patient", "a longer, more contextual 'story' summary for the letter to the GP"
KF8	Embedded not strictly required in pilot	The participants emphasise that the survivability depends on the integration in their current system, however, this is not needed in a possible pilot.	Contextual aspects	"it is good to have this intermediary step", "the summarisation flow is logical"

Table 8.3: The key findings from the evaluation interviews.

- **Summarisation tool could identify non-validated medical information.** This will help healthcare professionals as medical disinformation keeps growing among patients with the rise of social media platforms. For example, doctors encounter medical 'fake news' more often nowadays about alternative medicine. This could be identified by the summarisation tool when certain topics are addressed in the input. Implementing this would require the presence of medical knowledge base which is maintained by experts.

- **Only numerical sentences from selected topics should be displayed.** The system displayed all numerical sentences, no matter where they originated from in the input. Some participants indicate that it might be desirable that only numerical sentences are shown from parts of the input that are included in a selected (relevant) topic.
- **Generated summary should trigger actions.** Besides summarisation, a healthcare professional has to perform many other tasks after a consultation session. These tasks often originate from the summary. Some participants mentioned that connecting actions to the summary (based on information in the summary) would be a great addition. For example, new medicine prescribed by the doctor could automatically trigger an order for this patient to get this medicine. Another example would be the scheduling of a next appointment.
- **Generated summary to show to patient.** The generated summary could act as an agreement between the patient and doctor after a consultation session. In this scenario the patient could read the generated summary to make sure the doctor has not forgotten anything or changed any of the original information shared by the patient. This could even be given a legal status if some dispute arises between the patient and doctor later on.
- **Topic hierarchy.** The generated topics could benefit from a (predefined) hierarchy, such that topics fall into certain topic categories. For example, all medical topics could be clustered together and all non-medical topics. This would make the topics more understandable in the case of multiple topics.

9
Discussion

This chapter will discuss the findings of this thesis research and will also discuss the limitations that were met during the course of the research. Those limitations are important to address and will be discussed in the first section of this chapter. Then, a personal reflection will be used to discuss surprising findings or other things that are worth mentioning. In the last section of this chapter, possible future research directions will be mentioned. Some of those future research suggestions will also relate to the limitations mentioned in the first section.

9.1. Limitations

This section will discuss the limitations that were encountered during the course of this thesis research. Note that some of these limitations are related to the scope and time of a master's thesis. Addressing these limitations is important to recognise the boundaries of the research findings. Furthermore, some of the limitations mentioned in this section will serve as the basis for suggestions for future research in Section 9.3.

- **Number of interviews:** Although it was quite a surprise that a lot of healthcare professionals accepted the invitation to join the interview round, the number of interviews conducted is still relatively small for research study. However, to address this and acknowledge this, the interviews were conducted in a qualitative manner and the results were mostly not based on numerical values like number of occurrences. To further validate the findings from this research in a broader setting more interviews and ideation sessions could be used.
- **Lack of training data:** This limitation has been mentioned multiple times throughout this thesis report and mostly hindered the improved development of the prototype. Especially for finetuning the LLM for both topic modelling and summarisation more (labelled) training data is needed. This research has focused primarily on the problem context but in future research were the 'perfectness' of the system is necessary, the availability of training data is a requirement .
- **Resources for prototype:** In addition to the performance of the prototype being lowered by the lack of training data, also the lack of resources for the prototype has an effect on the performance. Naturally, bigger, faster machines would improve the speed and capabilities of the prototype (especially in the case of a privately hosted LLM). For this research a (although powerful) shared machine was used from TU Delft.
- **Evaluation of prompt engineering:** The evaluation of the prompt engineering was quite simple and qualitative for this research. This was caused by mainly two reasons. Firstly, the lack of labelled training data made it hard to quantatitively evaulate the performance of the designed prompts. In addition, the subjective nature of the performance and context dependency of the prompts also made it hard to evaluate the prompts. Future research could find ways to evaluate the performance in an implementation & adoption setting.
- **Evaluation of usage:** A big limitation of this research, although out of the scope of this thesis but nonetheless important, is the lack of evaluation within a use case scenario. The ideas and findings of this thesis have been analysed and evaluated with experts, but have not been evaluated within a real life scenario. This is necessary to prove the feasibility and applied usefulness of the summarisation system.

9.2. Reflection

The content of the thesis research was not only interesting and relevant on a scientific and social level, but also on a personal level. Coming from a theoretical computer science background, it was very refreshing and interesting to apply computer science and research principles in a practical setting with a very concrete and relevant problem that affects real people. This thesis showed me that computer science can be used to tackle problems that people have to deal with on a day-to-day basis and this can have huge impact, especially in a domain that is crucial to everyone.

It was surprising to see the amount of enthusiasm with which healthcare professionals want to start working with smart technologies that will change the healthcare domain. I was happy to see that so many of the invitations that were sent to participate in the interviews were accepted and that healthcare professionals could make time in their busy schedule to help me with my thesis research. I also experienced that using a network such as my supervisors is extremely helpful (almost necessary) to conduct such research and could be a great resource to future research.

A challenge that might occur when conducting research about these topics in a broader setting is the differences between hospitals, countries, and healthcare professionals when it comes to summarisation. Even within one hospital, there are different ways of handling this. Between hospitals, there are differences in the systems and methodologies used. Comparing these processes between countries or regions might reveal even greater differences. It remains to be seen how widely applicable a generalisable system may be, but this is definitely something to take into account.

9.3. Future Work

In addition to the limitations and personal reflections discussed in this chapter, this section presents several future research directions that build on the contents of this thesis. These suggestions aim to remove the limitations discussed in Section 9.1 and to solve the remaining unknowns.

- **Validate findings in different hospital/region:** The research of this masters thesis focused and took place within the Erasmus Medical Center in Rotterdam. Future research should also focus on different hospitals and regions/countries to validate the generalisability of the findings of this research. In addition, more interviews and evaluation sessions with experts could further strengthen and improve the system design proposed in this research. This addresses the limited number of interviews of this research and the sole focus on one hospital in the Netherlands.
- **Implement remaining system requirements:** Not all of the system requirements stated in Chapter 6 and further improved in Chapter 8 have been (fully) implemented due to feasiblity or other constraints. Therefore, future research could further design, implement and evaluate these requirements within a working system. In addition, if necessary, more system requirements could be added with improved knowledge of the problem setting.
- **Implementation & adoption study:** This research has mainly focused on discovering the problem setting and its context/requirements. Future research could and should focus more on the implementation and adoption process within the medical domain of the summarisation system. Studies on the adoption process (how to get everyone on board for this system) and the effect on stakeholders (patients, doctors, etc.) are of paramount importance to understand and further develop the system.
- **Create shareable data set:** As mentioned in the limitations of this research, the development of the system (especially the finetuning parts) are heavily reliant on the availability of training data. Future research could be used to create such shareable data sets that can be used in many patient - healthcare professionals interactions centred AI applications.

10 Conclusion

This thesis began by describing the problem that was the center of the research conducted; the manual summarisation process of healthcare professionals before, during, and after patient interactions. This summarisation process, which is part of the documentation tasks of the professional, impacts the healthcare professional by taking up a lot of time which results in work pressure. In addition, the patient is also affected by this, as the healthcare professional is forced to spend less time (of the already limited time per patient) on the actual interaction. In addition, the professional has to interact with the computer during the consultation session, which negatively impacts the patient. The problem is also characterised by the large amount of data that the healthcare professional has to consider and the noise that can be present, which distinguishes it from a normal text summarisation task.

To address the problem, the primary goal of this thesis was to explore the setting of the problem and learn about the requirements and objectives of a solution. Furthermore, this research aims to provide a solid patient experience text summarisation system design that adheres to requirements of healthcare professionals and can be used as a basis for real-world applications. Therefore, the primary research question of this thesis was: "How can the summarisation of patient experience documents for healthcare professionals be automated by AI to reduce their workload?". This research question was divided into five subquestions which aimed to learn about the requirements, barriers, noise, design, and evaluation of such a text summarisation system. These research questions were translated into a research process using the design science research method [14, 15], which provided rigour and structure to the research.

To answer the first three sub-questions, a literature review has been used in combination with semistructured interviews. The literature review was conducted on the current state of (long) text summarisation and also with a specific focus on the medical domain. This provided a solid knowledge base for this research. The semi-structured interviews were used to learn and gain knowledge from the environment part of the problem setting and to understand the business needs. This first phase helped discover the problem and understand the needs and barriers of healthcare professionals to using an automated summarisation system. The main output of this step are the system requirements (Table 6.2) which were derived from the coding analysis of semi-structured interviews with stimuli. For these interviews, an initial prototype was created to stimulate participants to think about the use of the system. The last two sub-research questions were built on top of the created system requirements by implementing a selected sub-set in the prototype (research question 4) and evaluating this prototype and the requirements with experts in another round of semi-structured interviews (research question 5). The output of these research questions are the key findings summarised in Table 8.3, which are also derived based on a thematic coding analysis of the transcripts. In addition, several other findings were included in Chapter 8, which provide interesting and important insights on the use and requirements of a summarisation system.

The system requirements derived in the first phase of this thesis (Table 6.2) summarise the most important findings about the requirements & barriers to a summarisation system and the noise present in patient experience documents. These findings showed that noise in patient experience documents has a subjective nature as certain information can be noise or not depending on the context and the perspective of healthcare professionals. Furthermore, research shows that for the summarisation system to work properly, a mechanism is needed to filter out the noise of the input. In addition, human validation, reinforcement learning, and numerical factual correctness are important requirements to using a summarisation system in the medical domain. The interviews also revealed contextual requirements to the system; it being embedded in the current system used and the paramount importance of the time efficiency attribute.

This research further extends on these findings by building a prototype that implements most of those requirements to further conduct research on the problem setting and objectives and requirements to a solution. The prototypes includes a newly proposed method of combining topic modelling and LLM-based text summarisation to tackle the subjective noise problem and the long text problem at the same time. Furthermore, topic modelling and text summarisation were engineered and improved, resulting in using an LLM to perform the topic modelling task. Then, this improved prototype was used to conduct a second round of interviews focussing on the evaluation of the system and its requirements. The findings of this phase of the investigation are best summarised in Table 8.3 where the most important results of the coding analysis are described. This shows that the system, and most importantly the flow and approach of the system, are positively validated by experts. Furthermore, it states the interesting finding about the dilemma of reinforcement learning; implementing this in a collective or individual way. Collective reinforcement learning yields great benefits when it comes to data quality and standards, but poses some risks for the adoption of an automated summarisation system. Lastly, the embedding of a summarisation system in the currently used system is not strictly necessary during a pilot period according to experts. The findings of this research are intended to provide answers to the research questions that were first mentioned in Chapter 3. Table 10.1 shows how the findings relate to individual research questions.

How can the summarisation of patient experience documents forhealthcare professionals be automated by AI to reduce the	
Question	Answer
1: What are the needs of healthcare professionals to trust and use an AI tool for summarising patient experience documents?	The requirements are best captured in Table 6.2. The most important needs are the ability to remove noise from the input data, reinforcement learning, and the numerical factual correctness of the summary.
2: What are the current barriers to use an AI tool for summarising patient experience documents?	This question is also related to the system requirements constructed. The main barriers are the need to embed the summarisation tool in the currently used system and possible unwillingness of healthcare professionals to use the system if it does not provide immediate time efficiency. Another risk is the use of collective reinforcement learning as professionals will be forced to adhere to standardised summaries.
3: What amount and types of noise are present in patient experience documents?	According to healthcare professionals there is a lot of noise present during patient interactions. This noise is subjective of nature; which means that information is not always noise or not noise. The noise is also not limited but nonmedical information, but can also include medical information.
4: How can an AI tool be designed to fit the needs of healthcare professionals?	This question was explored by the building of a prototype based on the system requirements. The prototype shows a start of how the tool can be designed with the needs of the healthcare professionals in mind. Topic modeling combined with summarisation (both based on LLMs) was proposed as a solution to subjective noise & long-text summarisation.
5: How can the designed AI tool be evaluated and improved upon?	The system requirements & prototype was evaluated with experts to gain more insights about the problem settings and objective of a solution. The key findings are the main result Table 8.3. The most important findings are the need for a different structure of the summary and the dilemma of chosing between collective or individual reinforcement learning.

Table 10.1: The five research questions of this thesis research and their answers.

The research carried out in this master's thesis is relevant in multiple ways. First, the research is scientifically relevant, as it outputs new knowledge about text summarisation in the medical domain. The contributions on this aspect are best summarised in the system requirements and key findings of the two rounds of interviews. An important finding here is the notion of subjective noise in the healthcare domain. In addition, this research proposes a novel method of combining topic modelling and text summarisation to tackle subjective noise and deal with long texts in a divide-and-conquer manner. Furthermore, this research has social relevance, as it is closely related to the work pressure and burnout issues present in the healthcare domain. This domain has significant challenges for the coming years that have a huge impact on both healthcare professionals and patients. Addressing issues such as the time spent on manual text summarisation is of paramount importance for a healthy future of the healthcare domain and its stakeholders. This societal relevance is best captured in interviews with healthcare professionals and the findings that directly relate to the needs and barriers of healthcare professionals to using an automated summarisation system.

Naturally, this research has its limitations. These limitations have been discussed and addressed in Chapter 9. As mentioned, one of the most important limitations is the limited number of interviews and the focus on the Erasmus Medical Centre. Future research should broaden to include multiple regions / locations to validate the generalisability of the findings of this research. Furthermore, future research could build on the requirements of the system and the key findings of the evaluation interviews to build a system that can be studied in the implementation and adoption phase in the medical domain. This system should also implement the requirements that could not be implemented in this thesis research. Another important addition for research addressing AI solutions related to patient interactions is the creation of a shareable data set that can be used as training data for these kinds of research projects.

References

[1] Centraal Bureau voor de Statistiek. “Helft zorgwerknemers vindt werkdruk te hoog”. In: *Centraal Bureau Voor De Statistiek* (Nov. 2022). url: https://www.cbs.nl/nl-nl/nieuws/2022/46/helft-zorgwerknemers-vindt-werkdruk-te-hoog.

[2] Colin P West et al. “Physician burnout: contributors, consequences and solutions”. In: *Journal of internal medicine* 283.6 (2018), pp. 516–529.

[3] Maria Panagioti et al. “Association between physician burnout and patient safety, professionalism, and patient satisfaction: a systematic review and meta-analysis”. In: *JAMA internal medicine* 178.10 (2018), pp. 1317–1331.

[4] Daniel S Tawfik et al. “Physician burnout, well-being, and work unit safety grades in relationship to reported medical errors”. In: *Mayo Clinic Proceedings*. Vol. 93. 11. Elsevier. 2018, pp. 1571–1580.

[5] Aug. 2023. url: https://www.rtl.nl/rtl-home/artikel/5403019/zorg-gezondheidszorg-verpleegkundige-uitstroom-instroom-cbs-hoge-werkdruk.

[6] Rachel Willard-Grace et al. “Burnout and health care workforce turnover”. In: *The Annals of Family Medicine* 17.1 (2019), pp. 36–41.

[7] Jeffrey Budd. “Burnout Related to Electronic Health Record Use in Primary Care”. In: *Journal of Primary Care & Community Health* 14 (2023), p. 21501319231166921.

[8] Emily Gesner et al. “Documentation burden in nursing and its role in clinician burnout syndrome”. In: *Applied clinical informatics* 13.05 (2022), pp. 983–990.

[9] Ming Tai-Seale et al. “Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine”. In: *Health affairs* 36.4 (2017), pp. 655–662.

[10] David Luna et al. “Burnout and Its Relationship with Work Engagement in Healthcare Professionals: A Latent Profile Analysis Approach”. In: *Healthcare*. Vol. 11. 23. MDPI. 2023, p. 3042.

[11] Frans Raaijmakers. *Healthcare is under pressure, but technology can help solve its challenges*. en. Sept. 2023. url: https://www.tue.nl/en/news-and-events/news-overview/27-09-2023-healthcare-is-under-pressure-but-technology-can-help-solve-its-challenges.

[12] Hans Moen et al. “Comparison of automatic summarisation methods for clinical free text notes”. In: *Artificial intelligence in medicine* 67 (2016), pp. 25–37.

[13] Anirudh Joshi et al. “Dr. Summarize: Global Summarization of Medical Dialogue by Exploiting Local Structures.” In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 3755–3763.

[14] Hevner et al. “Design science in information systems research”. In: *MIS Quarterly* 28.1 (2004), p. 75. doi: 10.2307/25148625.

[15] Ken Peffers et al. “A design science research methodology for information systems research”. In: *Journal of management information systems* 24.3 (2007), pp. 45–77.

[16] Bert Bukman. *Lengte consult Nederlandse huisartsen in internationale middenmoot - Zorgvisie*. nl-NL. May 2019. url: https://www.zorgvisie.nl/lengte-consult-nederlandse-huisartsen-middenmoot/.

[17] Ian Scot Watt. “Consultation, medical history and record taking”. In: *Clinical examination*. Elsevier, 2008, pp. 1–19.

[18] Peter J Edwards. “GPs spend 14% of their session time documenting consultation notes and updating electronic health records”. In: *British Journal of General Practice Open* 74.742 (2024), p. 202.

[19] Susan Cameron et al. “Learning to write case notes using the SOAP format”. In: *Journal of Counseling and Development* 80.3 (July 2002), pp. 286–292. doi: 10.1002/j.1556-6678.2002.tb00193.x. url: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1556-6678.2002.tb00193.x.

[20] Maria Alcocer Alkureishi et al. “Impact of electronic medical record use on the patient–doctor relationship and communication: a systematic review”. In: *Journal of general internal medicine* 31 (2016), pp. 548–560.

[21] Cheryl Rathert et al. “Patient-centered communication in the era of electronic health records: What does the evidence say?” In: *Patient education and counseling* 100.1 (2017), pp. 50–64.

[22] Hoang Nguyen et al. “Text mining in clinical domain: Dealing with noise”. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 549–558.

[23] M Clare Taylor. “Interviewing”. In: *Qualitative research in health care* (2005), pp. 39–55.

[24] John W Creswell et al. *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications, 2016.

[25] Jill Vincent et al. “Using multimedia stimulus materials in an electronic interview to gain in-depth responses from professionals: a case study”. In: *Advancing Research Methods with New Technologies*. IGI Global, 2013, pp. 316–328.

[26] K Louise Barriball et al. “Collecting data using a semi-structured interview: a discussion paper”. In: *Journal of Advanced Nursing-Institutional Subscription* 19.2 (1994), pp. 328–335.

[27] Barbara DiCicco-Bloom et al. “The qualitative research interview”. In: *Medical education* 40.4 (2006), pp. 314–321.

[28] Tegan George. *Semi-structured interview: Definition, Guide & Examples*. Nov. 2022. url: https://www.scribbr.com/methodology/semi-structured-interview/.

[29] Hanna Kallio et al. “Systematic methodological review: developing a framework for a qualitative semi-structured interview guide”. In: *Journal of advanced nursing* 72.12 (2016), pp. 2954–2965.

[30] Steven Eric Krauss et al. “Preliminary investigation and interview guide development for studying how Malaysian farmers’ form their mental models of farming”. In: *The Qualitative Report* 14.2 (2009), p. 245.

[31] Femke WM Damen et al. “Added value of physical food products as a stimulus during interviewing”. In: *Appetite* 169 (2022), p. 105819.

[32] Razvan Azamfirei et al. “Large language models and the perils of their hallucinations”. In: *Critical Care* 27.1 (2023), p. 120.

[33] Jonathan Pilault et al. “On extractive and abstractive neural document summarization with transformer language models”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020, pp. 9308–9319.

[34] Alexios Gidiotis et al. “A divide-and-conquer approach to the summarization of long documents”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 3029–3040.

[35] Yao Zhao et al. “Seal: Segment-wise extractive-abstractive long-form text summarization”. In: *arXiv preprint arXiv:2006.10213* (2020).

[36] Potsawee Manakul et al. “Long-Span Summarization via Local Attention and Content Selection”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 6026–6041.

[37] Zhengyuan Liu et al. “Topic-aware pointer-generator networks for summarizing spoken conversations”. In: *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2019, pp. 814–821.

[38] Huan Yee Koh et al. “An empirical survey on long document summarization: Datasets, models, and metrics”. In: *ACM computing surveys* 55.8 (2022), pp. 1–35.

[39] DAIR.AI [DAIR.AI]. *Prompt Chaining*. Apr. 2024. url: https://www.promptingguide.ai/techniques/prompt_chaining.

[40] Sabine Molenaar et al. "Medical dialogue summarization for automated reporting in healthcare". In: *Advanced Information Systems Engineering Workshops: CAiSE 2020 International Workshops, Grenoble, France, June 8–12, 2020, Proceedings 32*. Springer. 2020, pp. 76–88.

[41] Wen-wai Yim et al. "Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization". In: *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*. 2021, pp. 10–20.

[42] Kundan Krishna et al. "Generating SOAP Notes from Doctor-Patient Conversations Using Modular Summarization Techniques". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2021, pp. 4958–4972.

[43] Bharath Chintagunta et al. "Medically aware GPT-3 as a data generator for medical dialogue summarization". In: *Machine Learning for Healthcare Conference*. PMLR. 2021, pp. 354–372.

[44] Longxiang Zhang et al. "Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations". In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2021, pp. 3693–3712.

[45] Alex Papadopoulos Korfiatis et al. "PriMock57: A Dataset Of Primary Care Mock Consultations". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2022, pp. 588–598.

[46] Francesco Moramarco et al. "Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation". In: *ACL 2022: 60th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2022, pp. 5739–5754.

[47] Kurtis Pykes. *What is Topic Modeling? An Introduction With Examples*. Oct. 2023. url: https://www.datacamp.com/tutorial/what-is-topic-modeling.

[48] Douglas W Maynard. "Placement of topic changes in conversation". In: (1980).

[49] Gianpaolo Manalastas et al. "What does the structure of a medical consultation look like? A new method for visualising doctor-patient communication". In: *Patient education and counseling* 104.6 (2021), pp. 1387–1397.

[50] Maarten Grootendorst. "BERTopic: Neural topic modeling with a class-based TF-IDF procedure". In: *arXiv preprint arXiv:2203.05794* (2022).

[51] Maarten P. Grootendorst. *BERTopic*. url: https://maartengr.github.io/BERTopic/index.html.

[52] Rembrandt Oltmans. "Clearing the Air: An Exploration of Pulmonologists' Needs and Intents in XAI Solutions for Respiratory Medicine". In: (2023).

[53] Lm-Sys. *GitHub - lm-sys/FastChat: An open platform for training, serving, and evaluating large language models. Release repo for Vicuna and Chatbot Arena.* url: https://github.com/lm-sys/FastChat.

[54] Lianmin Zheng et al. *Judging LLM-as-a-judge with MT-Bench and Chatbot Arena*. 2023. arXiv: 2306.05685 [cs.CL].

[55] *HREC Approval Application*. url: https://www.tudelft.nl/over-tu-delft/strategie/integriteitsbeleid/human-research-ethics/hrec-approval-1-application.

[56] Alyona Medelyan PhD. *Coding qualitative data: How to code qualitative research (2023)*. Jan. 2023. url: https://getthematic.com/insights/coding-qualitative-data/.

[57] Benjamin F Crabtree et al. "A template approach to text analysis: developing and using codebooks." In: (1992).

[58] Richard E Boyatzis. *Transforming qualitative information: Thematic analysis and code development*. sage, 1998.

[59] Pranee Liamputtong Rice et al. “Qualitative research methods: A health focus”. In: *Melbourne, Australia* (1999).

[60] Josh Achiam et al. “Gpt-4 technical report”. In: *arXiv preprint arXiv:2303.08774* (2023).

[61] David M Blei et al. “Latent dirichlet allocation”. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[62] DAIR.AI [DAIR.AI]. *Zero-Shot Prompting*. Apr. 2024. url: https://www.promptingguide.ai/techniques/zeroshot.

[63] Jason Wei et al. “Finetuned Language Models are Zero-Shot Learners”. In: *International Conference on Learning Representations*. 2021.

[64] DAIR.AI [DAIR.AI]. *Few-Shot prompting*. Apr. 2024. url: https://www.promptingguide.ai/techniques/fewshot.

[65] Tom Brown et al. “Language models are few-shot learners”. In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[66] Sewon Min et al. “Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?” In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 11048–11064.

[67] DAIR.AI [DAIR.AI]. *Chain-of-Thought Prompting*. Apr. 2024. url: https://www.promptingguide.ai/techniques/cot.

[68] Jason Wei et al. “Chain-of-thought prompting elicits reasoning in large language models”. In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

[69] Takeshi Kojima et al. “Large language models are zero-shot reasoners”. In: *Advances in neural information processing systems* 35 (2022), pp. 22199–22213.

[70] Xuezhi Wang et al. “Self-Consistency Improves Chain of Thought Reasoning in Language Models”. In: *The Eleventh International Conference on Learning Representations*. 2022.

[71] DAIR.AI [DAIR.AI]. *Generated Knowledge prompting*. Apr. 2024. url: https://www.promptingguide.ai/techniques/knowledge.

[72] Jiacheng Liu et al. “Generated Knowledge Prompting for Commonsense Reasoning”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 3154–3169.

[73] Shunyu Yao et al. “Tree of thoughts: Deliberate problem solving with large language models”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[74] Jieyi Long. “Large language model guided tree-of-thought”. In: *arXiv preprint arXiv:2305.08291* (2023).

[75] DAIR.AI [DAIR.AI]. *Tree of Thoughts*. Apr. 2024. url: https://www.promptingguide.ai/techniques/tot.

[76] Sebastian Riedel. *Retrieval Augmented Generation: Streamlining the creation of intelligent natural language processing models*. Sept. 2020. url: https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/.

[77] Patrick Lewis et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.