CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Applying and Evaluating Large Language Models for Triage at a Paediatric Emergency Department in a Swedish Hospital

Master's Thesis in Computer science and engineering

Tindra Järgenstedt & Elin Nilsson

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's Thesis 2025

Applying and Evaluating Large Language Models for Triage at a Paediatric Emergency Department in a Swedish Hospital

Tindra Järgenstedt & Elin Nilsson

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Applying and Evaluating Large Language Models for Triage at a Paediatric Emergency Department in a Swedish Hospital
Tindra Järgenstedt & Elin Nilsson

Supervisor: Hans-Martin Heyn, Department of Computer Science and Engineering
Examiner: Miroslaw Staron, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Tindra Järgenstedt & Elin Nilsson
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis aims to explore and evaluate a software system using an LLM at the Paediatric Emergency Department (PED) at Sahlgrenska University Hospital. Approximately 60,000 patients visit the PED annually, while reports of decreasing staff availability and increases in burnout are observed. LLMs have shown potential in medical tasks, however, there is limited knowledge on how they would perform in a real setting. This thesis explored LLMs for streamlining the triage process to address this problem. Design Science Research was applied through three iterations involving interviews, prompt engineering, system-level simulations and human evaluations with nurses and voluntary patients.

15 functional and 10 non-functional requirements covering aspects such as accuracy, relevancy, usability and regulatory compliance were elicited from the stakeholders: nurses, the head of section, data scientists, and infrastructure providers. These were translated into a prototype using four instances of Llama 3.3 70B Instruct with Retrieval-Augmented Generation (RAG), each handling tasks such as generating follow-up questions, suggesting clinical controls and tests, or summarising information. The prototype demonstrated potential to support the triage process in 80% of the cases, showed particularly promising results in terms of accuracy when suggesting controls and generating relevant questions. However, it also exhibited certain limitations. Implementing LLM systems in a PED requires further research, especially on validating information completeness and how the RAG document structure and content affect accuracy.

Keywords: large language models, retrieval-augmented generation, artificial intelligence, healthcare, paediatric emergency department

Acknowledgements

This master's thesis was conducted during the spring semester of 2025 as part of our studies in Software Engineering and Technology/Management at Chalmers University of Technology and Gothenburg University. The study was performed in collaboration with Sahlgrenska University Hospital.

We are grateful to our academic supervisors, Hans-Martin Heyn and Hina Saeeda, for providing invaluable guidance and support throughout this project. We would also like to thank Isak Barbopoulos for generously sharing his expertise on Large Language Models and for always providing thoughtful advice, and a genuine willingness to help as this thesis took shape.

Our sincere appreciation goes to Hannah Sjöstedt for initiating an engaging and meaningful project and allowing our master's thesis to be a part of it, as well as for providing valuable guidance throughout the study.

We also extend our thanks to all of the research participants, the healthcare professionals at Sahlgrenska University Hospital, the voluntary patients and the experts within AI in healthcare. This project would not have been possible without your contributions.

Tindra Järgenstedt & Elin Nilsson, Gothenburg, June 2025

Contents

List of Figures

List of Tables

1

Introduction

Sahlgrenska University Hospital is a Swedish hospital with the largest Paediatric Emergency Department (PED) in Sweden. Approximately 60,000 children visit the PED at Sahlgrenska University Hospital each year. Almost 50% of these cases can be dismissed as non-urgent [1]. Due to the large number of visitors at the PED, much time and effort are spent doing triage. Triage is done to assess the severity of the patient's condition and to prioritise the patient correctly. At the same time, recent research suggests increasing patient volumes, decreasing staff availability, and an increase in burnout and overburden in emergency providers [2]. Therefore, accelerating the triage process to be able to handle the increased volumes of patients, while ensuring the accuracy of decisions, is crucial for ensuring the quality of emergency care [3].

Large Language Models (LLMs) can lead to reduced cognitive load on personnel and streamline processes, for example, in the medical sector [4], the judicial sector [5], and the banking sector [6]. Therefore, LLMs are a promising avenue to reduce the high workload of staff in PED. For example, parents waiting for an assessment by a nurse could chat about their child's condition with the LLM. The LLM could gather initial information, suggest treatments or tests and summarise the patient's condition and possibly facilitate the process for the nurse. Consequently, the use of LLMs for triage is a promising solution for improving the flow of patients at the PED and decreasing the workload for the nurses.

The technology behind LLMs is developing at a rapid pace. However, integrating LLMs into healthcare and specifically emergency departments is a new area of application, where research has been done to a lesser extent. Studies underline the potential of using an LLM in a medical environment: for example, when comparing different models to medical experts on early diagnosis retrospectively, research shows that GPT-3.5, GPT-4 and other variants of LLMs provide accurate results to a great extent [3]. However, risks remain, such as the LLM generating different answers for identical cases or hallucinations, meaning the LLM produces false information and presents it as if it were a fact [7]. Another issue is that almost all LLMs currently exhibit some degree of racial bias [8].

There are few studies focused on LLMs in actual clinical workflows or interacting with patients [9]. This implies that more research is needed to evaluate to what

extent LLMs can actually help reduce workloads on medical staff. There is a need to evaluate different applications of LLMs in the medical sector, and to validate the results of such applications, in order to incorporate this new technology successfully into the healthcare [9]. A validation strategy for the usage of LLMs in medical software is needed due to the wide variety of cases that can occur in healthcare, as well as the sensitive and critical nature of the operations of software in healthcare [10].

This thesis, therefore, aims to explore and evaluate a software application using an LLM to streamline information gathering in the triage workflow at the Sahlgrenska PED. By conducting interviews and observations, we intend to first identify the needs of the different stakeholders, such as nurses, doctors, patients' parents, and administrative staff, including software developers for the hospital, before designing the software. The elicited requirements are translated into an implementation of a software system using an LLM. An evaluation study is performed to evaluate to what extent the software can fulfil the stakeholders' needs.

This thesis aims to contribute to the field of software engineering in the following ways:

Contribution 1: A compilation of requirements to accommodate the special needs of the medical sector, specifically the needs present in PED, when implementing software using LLMs.

Contribution 2: The development of a prototype software system using LLMs that adheres to the elicited requirements.

Contribution 3: New test approaches that account for new quality aspects of the software, such as human alignment and safety in the context of the emotional safety of the patients.

1.1 Problem description

Due to the high demand of medical personnel at the PED at Sahlgrenska, there is a need to explore the use of new technologies to take over certain tasks and thereby relieve the medical staff of some work burden.

Currently, the use of software systems using LLMs has not been evaluated enough in a real clinical setting with patients. Therefore, important pieces are still missing before the LLMs can be used in a real context. One missing piece is the usability aspect. Such as, how should the LLM be used, and how should its assessments be displayed? To evaluate what is important in real settings, there is a need for requirements elicitation from a variety of different users with different backgrounds: e.g. nurses, doctors, and patients. This is something that will be more thoroughly explored in this study, more specifically at the PED at Sahlgrenska University Hospital.

Problem 1: Requirement elicitation : There is a lack of an overview of what

requirements are needed for a medical software system that entails the use of LLMs, both from patients and medical staff.

Problem 2: Implementing a prototype : There is a need to translate the requirements into a software implementation that performs according to them.

Problem 3: Evaluation of the prototype : It has been explained as hard to evaluate software systems using LLMs because of the complexity of the LLMs. Focuses lacking are on emotional safety as well as human alignment.

1.2 Purpose of the study

This study aims to investigate how a software system using an LLM for parents to chat about their child's medical condition can be implemented at the PED at Sahlgrenska University Hospital. Currently, the use of LLMs has not been evaluated in a real clinical setting or with real patients at the Sahlgrenska PED. Instead, previous research has analysed the LLM when assessing archived emergency cases to evaluate the accuracy of the LLM [11], [12]. However, even though the accuracy of the assessments can be high, important aspects of Software Engineering are missing before LLMs can become part of the operational software services in a hospital.

Besides determining features and common requirements such as, safety and usability, the purpose of this study is to explore and identify requirements needed for applying LLMs in the aforementioned medical setting. The prototype will be used to explore new validation techniques in a real clinical setting, in order to evaluate the success of the requirements elicitation.

1.3 Significance of the study

Developing software products that entail the use of LLMs in their operation in the context of a children's emergency waiting room is an area that has not been explored in previous research. This thesis distinguishes itself from earlier research by carrying out the study in the specific context, Sahlgrenska PED, as well as evaluating the software by having voluntary parents describing the condition of their child in the actual PED waiting room.

To the best of the authors' knowledge, previous studies on the use of LLMs in emergency cases, such as [12], [9], [13] have only been limitedly explored in the emergency room. In this thesis, we conducted a full software development cycle to explore all aspects from requirements elicitation, software development, and testing, with the possibility to interact with all stakeholders: Nurses, doctors, and patients. Further, the prior evaluations of LLMs mainly focus on accuracy, by comparing the diagnosis or answer to how a real doctor would assess the case. In contrast to these, this thesis will also highlight other validation metrics, such as usability, human alignment and emotional safety of the software system using an LLM, in the PED at Sahlgrenska University Hospital.

1.4 Research questions

The research questions in this thesis will be the following:

RQ1: What are the stakeholders' characteristics and requirements for software using an LLM in a PED at Sahlgrenska? The goal of this question is to collect requirements from the different stakeholders to get a detailed overview of the demand for a software product that uses an LLM in operation in the PED environment. The stakeholders that this thesis will consider are nurses, doctors, patients' parents, and administrative staff, including software developers. This will give an understanding of, e.g., what information the LLM needs to collect, desired features, usability requirements and other requirements that are crucial for being able to implement a suitable software system using an LLM.

RQ2: How can the requirements be translated into development processes for software using an LLM in terms of model selection and prompt engineering? The purpose of this question is to translate the requirements from RQ1 into processes applicable for the development of software that uses an LLM. The development processes we investigate include prompt engineering and model selection. Thus, no custom models will be trained. The focus will be on investigating prompt engineering approaches and Retrieval-Augmented Generation (RAG) as well as model selection based on the requirements. This research question will explore different configurations and prompts to identify the most accurate and appropriate solutions.

RQ3: To what extent can software using an LLM fulfil the requirements of different stakeholders in a PED environment? The goal of this question is to evaluate if the results from RQ2 meet the requirements identified in RQ1. One example of how the outcome of the LLM will be evaluated is by letting professionals assess the same case as the LLM and then evaluate the outcome of the LLM. Other tests will also be done, such as measuring the experience of user interactions.

1.5 Limitations and delimitations

This thesis will only investigate the potential implementation of a software using an LLM at the Sahlgrenska University Hospital's PED. As a result, the requirements and evaluation outcomes may not be directly transferable to other departments within the hospital or other PEDs, which limits the external validity of the thesis.

Additionally, the thesis does not aim to develop and train a new LLM but instead, will focus on evaluating a software system using one existing LLM in combination with different prompt engineering approaches. The selection of specific approaches and the LLM will be affected by the identified requirements, therefore, some approaches or LLMs will not be considered or evaluated, which narrows the scope of the study.

A key component of the evaluation process involves conducting user tests with nurses and voluntary patients. The study does not include all types of visitors to the PED since it is limited to including parents or patients with Swedish literacy skills. Due to ethical concerns, it was only possible to include voluntary patients with less urgent conditions. As the research project must not affect patients' care, only patients with a low priority for triage were included. This has the consequence that more urgent cases are not evaluated in this study.

2

Background

2.1 An overview of Swedish healthcare

A review of the Swedish health system was conducted in 2023 as a collaboration between the European Observatory on Health Systems and Policies and the Swedish Agency for Health and Care Services Analysis [14]. The review describes the overall quality of healthcare in Sweden as generally high, with a goal of ensuring high-quality and equal healthcare, with a primary focus on prioritising those with the greatest medical needs.

As an early adopter in introducing IT systems into the healthcare sector, Sweden achieved almost full digitalisation of medical documentation in 2022. According to the review, digitalisation is an increasing trend in Swedish healthcare. Correspondingly, a majority of respondents to a population survey, done in 2021, reported having used digital services when accessing medical information, scheduling healthcare visits or communicating with healthcare staff. However, digital systems are not fully integrated for use in medical practices such as diagnosing, treating or caring for patients. This requires additional investments for training healthcare professionals, adaptations of current systems, and ensuring patient safety.

Moreover, the review addresses some current challenges in Swedish healthcare. Despite an overall high quality of healthcare, the Swedish healthcare system faces challenges in areas such as staff turnover, sickness absence rates, and staff utilisation. According to the review, several regions have reported a shortage of general practitioners and registered nurses, with all regions reporting a lack of specialist nurses by 2020. Since 2015, the number of registered nurses per capita has decreased, and labour unions argue that insufficient and stressful working environments are barriers to recruiting and retaining healthcare professionals. A reported challenge for emergency departments is the lack of hospital beds. This can elongate the hospital stays, affecting both patient safety and the working environment. Additionally, a limited availability in primary care has led more patients to seek emergency care, further intensifying the problem of extended waiting times in emergency departments [14].

Sahlgrenska PED reports that approximately half of its 60,000 annual visitors are referred to other healthcare providers, as their conditions do not require urgent emergency care [1]. As reported in the review of the Swedish health system, in most

cases, patients have been referred to the emergency department after contacting another healthcare provider, most commonly through the 1177 helpline. A key policy goal highlighted in the review is to improve the overall efficiency of the healthcare system. This aligns with the challenges faced by Sahlgrenska PED, where patients seeking care in the wrong setting contribute to an increased workload for staff.

The definition of triage

Triage is defined as "the sorting of patients (as in an emergency room) according to the urgency of their need for care" [15]. To avoid risking the safety of patients, it is important to identify those with life-threatening injuries or illnesses quickly. However, the number of patients can vary a lot and predicting this number is difficult [16]. Therefore, different emergency departments use different triage systems to be able to prioritise patients effectively. A safe and effective triage system avoids over-triage, e.g. assigning higher urgency than necessary, while also ensuring that patients are not assessed as less urgent than they are.

There are several factors that could affect the result of triage. Haim et al. mention nurse experience, training and decision-making heuristics, as well as biases or initial impressions of the patient [13]. Additionally, other factors like patient volumes, the number of available staff or other distractions can also cause variance.

The WEST-P triage system: The West Coast System for Triage – Paediatrica (WEST-P) was developed in Gothenburg to reduce the risk of over-triage [17]. With a high inflow, over-triage can lead to allocating the resources less efficiently and risk the safety of patients. The system consists of objective parts, such as warning signs and a scoring system based on vital signs, as well as a subjective part, a triage nurse's clinical assessment [18]. Each of the parts results in a colour[1]. The part with the highest colour decides the final colour. Implementation of the system resulted in a reduction of an average of 21 minutes in waiting time [1].

Triage at Sahlgrenska PED: Sahlgrenska PED takes care of patients up to 16 years of age with serious conditions such as chest pain, severe abdominal pain, severe headaches, injuries from accidents or respiratory problems [19]. The PED has moved from using the popular triage system RETTS-p to WEST-P, which has improved the accuracy of triage [1].

As illustrated in Figure 2.1, when a patient is seeking emergency care at Sahlgrenska PED, they are issued a queue ticket and complete a registration form with standard questions for all patients. The standard questions can be found in the Appendix A.1.

When it is the patient's turn, a registration nurse conducts a brief assessment of the patient's current condition based on the answers to the registration form, cause

[1] **Red:** Needs to see a doctor immediately.
Orange: Needs to see a doctor within 10 minutes.
Yellow: Needs to see a doctor within 1 hour.
Green: The patient can wait.

of visit and their own clinical judgement. In some cases, vital parameters such as the percentage of oxygen in the blood (POX) are measured. The patient may also be prepared for further painkiller treatment or asked to give a urine sample. The outcomes of the short assessment are either that the patient will be referred home with information about their condition or to another care provider, such as a local emergency clinic that can treat urgent but less severe cases, or to a healthcare centre. If the patient requires further examination or urgent care, they will be admitted to triage. The registration nurse also does a quick prioritisation to determine the order in which the patients will be admitted.

A triage nurse is responsible for performing the triage according to the WEST-P triage system. Based on the cause of visit, the patient will be examined in three parts. These are vital parameters, warning symptoms and clinical vision, with each of these parts being assigned an individual triage colour. The vital parameters are different tests such as temperature, pulse, blood pressure and respiratory rate, and the values of these generate a score translated into the triage colour. Secondly, the nurse will investigate warning symptoms, defined in WEST-P compendium found in Appendix A.7 based on the cause of visit. For instance, an unconscious patient will get the triage colour red for the warning symptoms part. Lastly, the triage nurse is allowed to give their own personal judgement based on clinical vision and may increase the priority of the patient. The highest prioritised colour of the three triage assessments will be the final colour of the patient, and will decide when they will see the doctor.

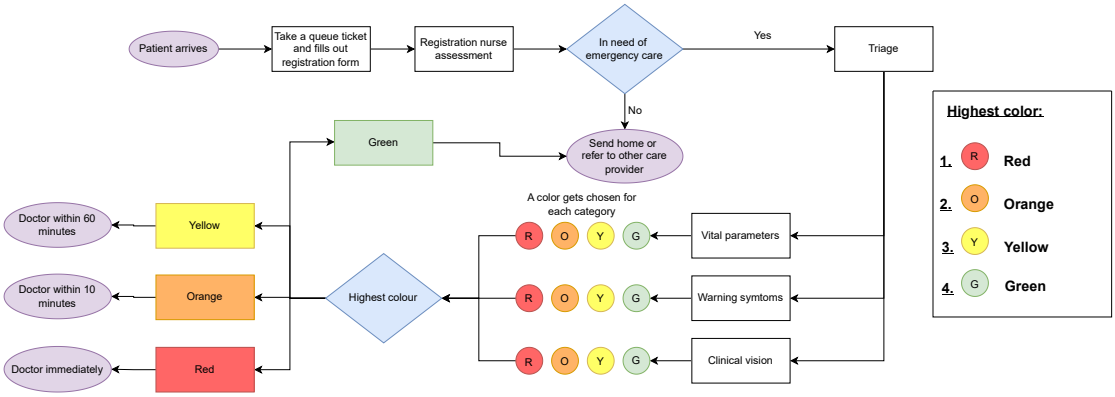

Figure 2.1: Flow chart of triage process at Sahlgrenska PED

2.2 Regulations for medical devices

The definition of a medical device is when a technical product used on humans has one or several purposes, defined in [20], such as *"diagnosis, prevention, monitoring, prediction, prognosis, treatment or alleviation of disease"*. Therefore, software, when used in healthcare, can be considered a medical device. Software used for digital decision support usually counts as a Medical Device Software (MDSW). In Sweden, an MDSW must follow EU regulations and Swedish national regulations, for instance Regulation (EU) 2017/745.

Regulation (EU) 2017/745 [21] covers requirements regarding areas such as development, design and documentation for MDSW. For instance, when designing a MDSW, the developers must ensure repeatability, reliability, performance and risk management. The development of MDSW should also follow the state of the art, taking into account the development life cycle and information security. There are also design requirements concerned with the environment in which the software will be used.

Additional regulations necessary to address when applying software using AI in Swedish healthcare are listed by the Swedish Medical Products Agency [22]. According to Article 22 of GDPR, individuals have the right not to be subject to decisions solely based on automated data processing. Further, there is a need to request documentation regarding data protection responsibilities related to the IT infrastructure where the model is run.

2.3 LLMs

An LLM is an AI model based on a deep learning neural network, trained on huge amounts of textual data [2]. This data can be derived from articles, books and other content found on the internet, and by analysing how words are used with each other, the LLM can apply these patterns to complete tasks [4]. Due to this, LLMs can handle unstructured input and respond to queries without being pre-trained on the specific task. Some common applications are chatbots, translating or generating text or text predictions.

Limitations and challenges of LLMs

Thirunavukarasu et al. [4] mentions some current limitations, such as accuracy, recency, coherence, transparency and ethical concerns. First, in terms of accuracy, if the models are trained on unverified or unvalidated data, the models can generate faulty answers. Also, two of the most common LLMs, GPT-3.5 and GPT-4 are only trained on data created before September 2021, meaning that the information might not be accurate or useful due to the lack of recency. Further, regarding coherence, the models are trained to learn probabilistic associations and patterns with words, rather than developing true understanding. As a result, both factual and fabricated information can be presented as if they were accurate. Another limitation is transparency, as the process by which models generate responses from input queries, architectural data, and algorithms remains unclear. This lack of transparency is due to the 'black-box' nature of the models, reducing the clarity of the generated outputs. Lastly, Thirunavukarasu et al. address the ethical concerns related to LLMs, such as the risk of privacy and security breaches, as well as the absence of established accountability for the consequences of their outputs.

The limitations of LLMs regarding the generation of nonsensical or incorrect answers are commonly known as hallucinations [23]. As mentioned, LLMs are not trained to develop a true understanding, but use patterns and probabilistic associations for

generating output. When faced with a knowledge gap, the model often attempts to fill it, using these patterns and associations, resulting in hallucinations. Incorrect or misleading outputs can be particularly problematic in critical contexts.

Evaluating LLMs

Chang et al. [24] state that there are several reasons why the evaluation of LLMs is of the utmost importance. Firstly, it gives insights into the strengths and weaknesses of LLMs. Another aspect is the fact that LLMs have extensive applicability, which means that it is important to ensure safe use in all of these areas of usage. This is even more important in sectors like healthcare, which are very safety-sensitive. Because of the broad abilities of the LLMs, current evaluation protocols quickly become insufficient. One of the reasons that they are becoming insufficient when it comes to LLMs is because of the LLM's multifacetedness. As of today, an extensive overview of all needed evaluations is missing.

Two common methods used to evaluate LLMs are presented by Chang et al. [24]. The first one is **automatic evaluation**, which is a method that uses standard metrics and evaluation tools to evaluate the performance. This is a method that is quite standardised since it does not require human participation. Which both minimises subjectiveness from humans and saves time. Most of the available benchmarks use automatic evaluation.

The second common method used to evaluate LLMs is through **human evaluation**. It means that evaluating the accuracy and quality of the model is done by humans. One example of when human evaluation is more suitable than automatic evaluation is on non-standard natural language tasks. Often in those situations, a human can more accurately assess the result than an evaluation metric can. When conducting this type of evaluation, experts, researchers or ordinary users are invited to evaluate the results from the model, which can generate more comprehensive feedback. To achieve a reliable human assessment, three evaluation criteria should be met. Firstly, a carefully chosen number of evaluators must be chosen to reach a nuanced and comprehensive result. Secondly, Chang et al. [24] also present six assessment criteria that should be considered. These criteria are as follows:

- **Accuracy** → assesses precision and correctness of a generated text.
- **Relevance** → assesses how well a text addresses the given context or query.
- **Fluency** → assesses grammar and readability through how well it chooses smooth and fitting expressions.
- **Transparency** → assesses how well the model communicates its thought process and decision-making process.
- **Safety** → assesses the ability to avoid generating harmful, offensive, or inappropriate text.
- **Human alignment** → assesses that the generated text respects and meets societal norms, user expectations and creates a positive LLM-Human interaction.

The third evaluation criterion is that the evaluators have relevant domain knowledge and are familiar with the tasks that are evaluated. By following these criteria, validity is added to the evaluation results [24].

2.4 Prompt engineering

Previously, to achieve task-specific performance from LLMs it has been necessary to retrain and fine-tune the model with more data. Prompt engineering has emerged as another way to achieve task-specific performance. A prompt is a strategically written instruction that is task-specific. It gives the possibility to alter the outcome of an LLM without altering the parameters or training it further. Consequently, prompt engineering is the activity in which these prompts are optimised for the specific task to perform. To achieve this optimisation, several Prompt engineering techniques have emerged [25].

Enhancing prompts with RAG

A way to complement prompt engineering techniques is by including Retrieval-Augmented Generation (RAG). RAG is a system that is used to reduce the hallucination of LLMs. When using RAG, you use a pre-built knowledge base to enrich the user-inputted prompt before it gets sent to the LLM. The user input gets analysed, and a custom-targeted query is made that is used to search through the knowledge base to find relevant resources. Snippets from the relevant resources are then incorporated into the user input to add more context to the prompt. The augmented prompt is then sent to the LLM, consequently creating more accurate answers [25].

3

Related Work

3.1 LLMs in healthcare

The usage of LLMs in general is a relatively new area proving high potential. Capabilities of answering questions, summarising, or paraphrasing text very similar to human performance have led to increased interest in the applications of LLMs across different domains, such as healthcare. Another indication of the potential is that ChatGPT has attained passing-level performance in the United States Medical Licensing, showing medical knowledge and reasoning [26, 4].

Clusmann et al. suggest that LLMs could be applied in areas within patient care like patient empowerment, translating and summarising, documentation and medical knowledge [26]. Assisting in the generation of standardised reports or organising unstructured notes could help medical staff with routine tasks. Although the potential, further development and extensive validation are essential to resolve current technological weaknesses.

Another recent study states that GPT-4 in combination with Chain-of-Thought (CoT) and few-shot prompting can achieve up to 96% accuracy in disease classification [27], indicating high potential of using LLMs in healthcare. The authors note that developing these models requires less effort than traditional machine learning. However, research on validating LLMs in healthcare is still limited, particularly regarding their appropriate roles in clinical workflows, and identifying use cases for both clinicians and patients [9].

AI Sweden [28] notes the remarkable lack of research into the practical implementation of LLMs and AI in healthcare, despite the significant attention it has received. Nor has there been a focus on examining whether the intentions of system developers match the expectations of those delivering care or the needs of patients. To create value from these applications, AI Sweden notes that competence from legal, ethical, health and implementation areas is needed. Detailed business knowledge och knowledge of the healthcare processes are also necessary to include in the development process.

The private usage of LLMs in medicine

Sandmann et. al [29] states that individuals likely use LLMs at home to seek medical advice. Therefore, they measured the accuracy of medical advice using the models GPT-3.5 and GPT-4 by extracting clinical cases without providing imaging or laboratory findings. Medical history and symptoms were provided in the ChatGPT session and prompted to the LLM with e.g. "What are my most likely diagnoses? Name up to five." Results were compared to Google searches for the same task and assessed independently by two physicians. The researchers tried to mimic a real patient by providing the information in first person and removing expert terminology. The results showed potential for answering medical questions and some weaknesses, mainly in relation to uncommon diagnoses [29]. Another study also showed that GPT-3.5 generated responses to actual patient queries were preferred in terms of quality and empathy compared to doctors [30].

3.2 LLMs in emergency care

Several studies have demonstrated the potential of using LLMs in healthcare overall, with some specifically exploring their applications in the specific environment of an emergency department. In a recent study, an LLM in a clinical simulation intended to simulate a parent asking for advice on their child showed good results in setting a diagnosis [9]. However, the recommended management was not always sufficient, risking the safety of the patient. In the scoping review of LLMs in Emergency Medicine (EM), the authors found that most of the previous research focuses on specific use cases, leaving a gap in exploring how LLMs can be integrated into the EM workflow [2]. It is mentioned that LLMs have been used or suggested in areas like Clinical Decision-Making and Support, and shown potential in assisting medical staff with identifying patients in need of urgent care.

To further explore the potential of LLMs, more specifically in an Emergency Department at a hospital, a study by Williams et al. was conducted [12]. The researchers found that while LLMs hold promise, their performance regarding accuracy needs to be improved before being used for actual clinical recommendations. The authors evaluated GPT-4-turbo and GPT-3.5-turbo performing three tasks – admission status, radiological investigation request status, and antibiotic prescription status. Both models performed worse than a resident physician in all cases except for one. When assessing if the patient needed antibiotics, GPT-4-turbo performed with higher accuracy than the physician. The study was carried out on clinical notes and is one of the few studies to have been carried out on real-world data.

Another study evaluated LLMs when generating discharge summaries, based on clinician notes at an Emergency Department [31]. The summaries generated by GPT-4 were entirely error-free in 33% of the cases and, in general, mostly accurate. However, 42% of the summaries exhibited hallucinations and 47% omitted clinically relevant information.

Williams et al. [12] highlight both the importance and difficulties with accessing real

clinical data to evaluate the LLMs on. While showing potential in previous studies, it is uncertain how well the LLMs will perform in a real-world setting. In the study, the clinical data used was a physician's first note of the patient's Presenting History and Physical Examination sections. Instead of a direct account from the patient, the gathered information has been interpreted, refined, and structured by the clinician. There is, however, a lack of research on the use of LLMs in contexts where such processing has not been performed.

3.3 LLMs in other safety-critical sectors

The adoption of LLMs is growing beyond healthcare, including in other safety-critical sectors like banking and the judicial system. These domains face similar challenges, and examining how they address them can offer valuable insights for healthcare applications.

Banking sector

Fan [6] and Kamalnath et al. [32] highlight key challenges in applying LLMs to the banking sector, including data privacy, lack of transparency, and bias.

Due to the sensitivity of banking data, privacy and security remain critical concerns. Both works note the risk of LLMs leaking personal or legally protected information if unintentionally included in training data. Fan proposes encryption, strict access control, and real-time monitoring to mitigate these risks.

Another shared concern is the opacity of LLM decision-making. To enhance transparency and trust, Fan advocates for Explainable AI methods and thorough model documentation. Kamalnath et al. similarly emphasise the difficulty of tracing model outputs and the limitations of relying solely on expert validation as usage scales.

Finally, both sources stress the importance of addressing bias in LLM outputs, particularly in contexts like credit approval. Fan recommends regular fairness audits across user groups, mirroring concerns in other high-stakes domains such as healthcare.

Judicial Sector

LLMs are increasingly applied in the judicial sector, which, like healthcare, involves high-stakes decisions, expert judgment, and strict regulations [33]. Applications include contract review, legal document summarisation, legal judgment prediction (LJP), and similar case retrieval (SCR) [5, 33, 34]. China's 'smart court' initiative has introduced AI-enabled judgments to reduce court backlogs [35], but concerns remain over undermining core values of justice, such as fairness, transparency, and reliability.

One study showing the potential of Legal AI is regarding LawLLM, a domain-specific model for the US legal system. It demonstrated promising results in LJP and SCR

tasks with accuracies of 79.4% and 81.6%, respectively [36]. Another study showed that an LLM trained on US Supreme Court decisions outperformed legal experts in predicting individual judges' decisions (71.9% vs. 66%) [5].

Despite these advances, ethical and technical concerns persist. Studies have documented racial and gender bias in AI legal judgments, particularly in areas like risk assessment and police violence cases [34]. Lai et al. [5] attribute this bias to the use of historical data that may embed systemic injustice. Data quality is another challenge, due to the success of Legal LLMs being heavily dependent on acquisition, organisation and deep learning of legal data. Therefore, inconsistent or flawed documents can reduce model effectiveness.

Given these risks, researchers emphasise that Legal LLMs should support, not replace, human professionals to maintain trust and integrity in the justice system [34].

The role of Software Engineering in building software using LLMs in critical applications

AI-based systems require interdisciplinary, collaborative teams with diverse expertise in areas such as data science, software engineering, social science, human-machine interaction and user experience [37]. This is due to the differences between developing traditional software systems and AI-based ones. For instance, other aspects need to be considered, such as ethics and equity requirements engineering. Failing to recognise these differences can result in the development of substandard AI-based systems [38],[39].

Lu et al. [37] highlight the importance of classifying ethical principles into carefully formulated requirements in a way that makes them quantifiable or measurable, while avoiding vague or unverifiable specifications. However, current principles defined by governments, research institutions, and enterprises are high-level and lack practical guidance on how to design and develop responsible AI-based systems.

Martínez-Fernández et al. [38] also highlight the challenge of specifying requirements, particularly due to difficulties in making them testable and measurable. The authors also state that it is unclear how non-functional requirements should be measured. Furthermore, there is a need to develop new types of non-functional requirements that consider explainability and freedom from discrimination. Lu et al. also mention *privacy*, as a non-standard software quality, increasingly important in AI-based systems and essential as a non-functional requirement for realising regulatory requirements, such as the GDPR.

The need for explainability is further explained by Menon et al. [40]. They highlight explainability as key to improving human-AI interactions by building trust in model outputs. This can be supported through features that explain predictions or decisions, as well as providing access to system artefacts [37]. Other ethical aspects, such as fairness and human-centred values, can be supported by involving stakeholders throughout the lifecycle of the system. An additional recommendation is to

limit automatic decision-making. Alternatively, AI-based systems can offer suggestions and seek human consent. There should also be a process that allows people to challenge the use or output of the system.

Verification and validation are used to ensure that systems comply with specified requirements and responsibly achieve their intended purpose. For AI-based systems, one activity suggested by Lu et al. is usability testing. They also mention system-level simulations to understand the behaviour of AI systems and evaluate the ethical quality attributes before deployment into a real-world setting.

Martínez-Fernández et al. state that there is a need to further explore software engineering practices for the development, maintenance and evolution of AI-based systems. They also state that AI-based systems have been widely used in the automotive domain, but much less so in the healthcare domain.

4

Methodology

4.1 Iterative design of the research

This study aimed at developing and evaluating a software using an LLM in a PED at Sahlgrenska University Hospital. To achieve this, a case study was conducted. According to Stol et al. [41], a case study is a type of field study and a research method that focuses on a part of or an entire organisation. They are done to deeply understand a real-life phenomenon without actively changing or controlling any environmental variables. This maximises the potential for a realistic context while remaining unobtrusive. This will contribute to knowledge about the possibility of LLM applications in a Swedish hospital.

The thesis was conducted following the Design Science Research framework. Wieringa [42] explains Design Science as "the design and investigation of artefacts in context". The artefact is designed to improve a problem in a certain context with which it interacts.

There are two main parts of Design Science, which are **design problems** and **knowledge questions**. A design problem requires a real solution. To design this solution, there is a need to understand the context and stakeholders. Knowledge questions, on the other hand, do not seek a change by designing a solution, but instead seek knowledge about the current world or context as it is. In the case of this study, a knowledge question could be:

- What requirements do a software system that uses an LLM in the Sahlgrenska PED need to fulfil?

In contrast, a design problem could be formulated as follows.

- Design a software system using an LLM that can be used in the Sahlgrenska PED according to the identified requirements.

This study is going to answer one knowledge question, RQ1, as well as solve a design problem, by answering RQ2 and RQ3.

This thesis has followed guidelines presented by Knauss [43] for applying Design Science Research in a Master's thesis, which is strongly influenced by Wieringa's

regulative cycle [43]. Below, the guidelines followed will be described in more detail.

The first guideline proposed by Knauss is to define the **artefact** early. Both researchers and the company, Sahlgrenska, should agree on it, even though the artefact can still evolve. By creating the artefact, knowledge questions can be answered. According to Knauss, the artefact must guide the knowledge questions, not vice versa. The second guideline suggests working in iterations. Each iteration should focus on gaining new knowledge for each research question, and the artefact should be further improved in each iteration. Knauss states that a Master's thesis usually achieves three full cycles. The third guideline refers to how to formulate the research questions. As the guidelines proposed by Knauss follow Wieringa's regulative cycle, they suggest formulating three research questions that each correspond to one phase in the regulative cycle. One question should be related to the problem, one to the possible solutions, and one to the evaluation of the solution. Guideline four proposes to have regular meetings with the supervisor. The fifth guideline proposes that each iteration should have a stronger focus on its corresponding research question while contributing knowledge to the other ones as well. For example, the second cycle's main focus should be on the solution. The seventh guideline states that the thesis document should be written during the study and restructured upon submission.

In the following sections, we will describe the methods applied in each of the three iterations. An outline of the methods and the research question affected is illustrated in Table 4.1.

4.2 Iteration 1: Elicitation phase

In the initial iteration, the emphasis was on RQ1, i.e., the process of requirement elicitation. The requirements were collected by conducting observational studies, reading applicable material, and conducting interviews with relevant stakeholders. During this iteration, initial prompt engineering approaches were also investigated, thus touching upon RQ2 as well. We also had a workshop to set up the infrastructure and explored different structures of LLM-based systems, touching upon RQ3. At the end of this iteration, evaluation was done as well, primarily by verifying elicited requirements with healthcare specialists through interviews. This was done to get a first assessment of their opinion on the requirements as well as their opinion on the possible prototype.

Document analysis

Relevant documents were studied to understand the domain and current situation at the hospital. Documents can provide insight into the context in which research participants function and provide valuable background information [44]. In addition, a document analysis can bring about questions that need to be asked and situations that are necessary to observe. The document analysis contributed to an objective view of the environment of the paediatric emergency department. The head of section provided us with documents specifying the triage system and work-

Table 4.1: An outline of the methods applied in the thesis and the affected RQ.

Iteration	Method	RQ affected
1	Document analysis	1
1	Interviews	1
1	Data analysis	1
1	Observation	1
1	Requirements specification	1
1	Requirements prioritisation and validation	3
1	Criteria-based filtering	3
1	Prototype workshop	2
2	Model selection	2
2	RAG implementation	2
2	Prompt Engineering	2
2	Requirements refinement	1
2	System-level simulations	3
3	iPad-based system deployment	2
3	Evaluation study	3
3	Requirements refinement	1

ing methods. By gathering relevant documents, relevant requirements for the LLM were elicited. The documents analysed were the WEST-P compendium and the triage guide, which can be found in Appendices A.7 and A.8.

Interviews

To further confirm the knowledge gained from the documents and to get an understanding of relevant stakeholders' wants and needs for an LLM application at the PED, interviews were held.

An interview is a research method used when conducting Field Studies. According to Lausen [45], interviews are good to conduct when there is a need for knowledge about the present situation and problems. When conducting interviews, it is important to have participants who are in the situation the researcher is interested in. Often, a manager might be interviewed even though they do not have real experience in the situation, thus giving a false picture of the setting.

Different types of interviews suit different situations depending on what information you want. There are cases where the interview could resemble a friendly conversation where the interviewer has a set of topics to be discussed. Other cases call for strictly

formulated questions that must be asked in the same way to all interviewees [46]. The interviews in this study aim to get a holistic view of the triage process. Therefore, this study used a combination of these two types, which is referred to as a semi-structured interview.

In this case, that means that interview guides with questions were formulated, but there was also room for asking any other questions that came to mind during the interview. The questions were open-ended, which allowed for different types of answers, even though all participants were asked the same set of questions. There has been a focus on keeping the questions neutral and clearly formulated. Judgmental language, for example, when asking about the reasons for a certain action ("why did you perform this action?"), was avoided in order not to influence the answers. [47].

The interviews started by having the interviewees sign a consent form, which allowed us to record and use their answers anonymously. They received information about where the data would be stored and that they were allowed to terminate their participation at any time. Afterwards, semi-structured interviews were carried out according to our interview guides, which can be found in Appendix A.2.

Selection of interviewees: The participants of the interview groups were chosen based on Patton's description of purposeful sampling. Patton [46] describes purposeful sampling as choosing information-rich participants whose interviews will shed much light on the research questions. He states that one might learn much more about the problem by carefully choosing a smaller number of participants that fit the study than when selecting a very large number of participants. Chain sampling is another purposeful sampling technique mentioned by the author. Therefore, interviewees were also chosen by having prior interviewees suggest additional participants they considered could provide useful information for this study.

The first group consisted of nurses who were chosen using a combination of these two sampling techniques. The first contact at the PED was the Head of Section, who suggested who could be interviewed. This was beneficial since she already knew which people could provide useful information for our study. This way, six interviews with nurses were conducted.

The second group consisted of two interviews: one with a data scientist and one with a manager at the AI platform at Sahlgrenska University Hospital. These interviewees were also found by purposeful sampling. The data scientist is one of the initiators for this study and an employee at the AI Competence Centre at VGR. The AI platform manager was suggested to us by our supervisor, who is also an employee at the AI Competence Centre at VGR.

The third and last group consisted of one interview with the Head of Section, who is also a doctor at the hospital.

The reason this study used three groups was to gain different kinds of information. It was crucial to interview nurses to gain knowledge about the working conditions,

how things are done in practice and what their opinion is on using AI. Technicians were interviewed since we needed to gain an understanding of what technical obstacles might arise, how an LLM-based system should be implemented, and to gain a deeper understanding of the technical environment at Sahlgrenska University Hospital. Lastly, we wanted to gain a broader understanding. Therefore, we interviewed the Head of Section. During this interview, the goal was to confirm that we had grasped the entirety of the problem.

Data Analysis

This study used two cycles of coding to extract themes and analyse the data from the conducted interviews. The first cycle includes *In Vivo coding* and *Descriptive coding*. The second cycle consisted of Pattern coding to extract patterns and themes from the first cycle of coding. According to Saldana [48], coding is to categorise something into a systematic order. By doing this, passages in text are captured by a summative phrase or word.

When doing In Vivo coding, each code refers to an actual phrase or word in the interview. Saldana states that In Vivo coding is particularly applicable when doing practitioner research since it captures the terms and language used in the field of study. Since this study examines practitioners at a PED, In Vivo coding was found to apply to the first cycle of coding. Descriptive coding is when each code is a word or phrase that summarises the topic talked about in a short section. Saldana also states that this creates a "categorised inventory" that is crucial preparation for the second cycle of coding. These two types of coding were used interchangeably in the interviews.

During the second cycle of coding, this study used Pattern coding, which is a coding technique applicable in second-cycle coding. Saldana describes pattern codes as codes that identify a meaningful theme by pulling together a lot of material. They group previous codes into a smaller number of sets or themes. In this study, this was done by going through all coded sections and grouping them into bigger themes.

When the second cycle of coding had been done, we created general themes into which the patterns/themes from the second cycle were grouped. This was then displayed using fishbone diagrams to get a comprehensive picture of the data analysis.

Observations

The information gathered from the document analysis and interviews was verified by conducting observations at the PED. An observational study, like the interview, is also a method that can be applied in field studies. Observations allow researchers to explore questions regarding how things work or what is going on within a given context [41]. By using observations as a technique for elicitation of requirements for a system, it is possible to either confirm or find mistakes in how a user has explained, for instance, their daily tasks [45].

Patton [46] states that the observation and interview complement each other. An observation allows for confirming what was said in the interview. The interview, in turn, gives an insight into the thoughts behind the observed actions. For a person to explain exactly how a routine is done, total awareness needs to be obtained from that person. More often than not, people are not aware of everything they do, and thus, observations allow the researchers to observe the things that escape awareness. Further, the present work, how the user is currently performing tasks, and present problems, problems that arise when performing tasks, in the context, are two areas where observations contribute the most to an elicitation.

In this study, we participated in the observations by being onlookers. This means that we were merely observing the situation and not participating in any way, as stated by Patton. The observations were performed in two different ways. Firstly, observations were done when the patients first encountered the registration nurse. The desired outcome of this observation was to see what questions were asked to the patients and in what order. Did the questions differ based on the cause of the visit, and were any other actions performed on the patient? It was also important to note how the patients were acting and how the nurses met the different behaviours. We also kept in mind whether any of the actions could be replaced by an LLM.

Secondly, observations were made inside the triage room when the patients were waiting for their turn. The desired outcome of this observation was to observe if the information gathered from the interviews aligned with reality. It was also important to note if any other actions took place that the interviewees had missed mentioning. During this observation, we also thought about whether any of the actions performed during triage could have been done by an LLM beforehand.

Methods for elicitation of requirements

In this thesis, after analysing the interviews with all stakeholder groups, requirements were extracted [1]. Statements from the interviews referring to functionalities were coded as "functional requirements" during the data analysis. The interviewees were explicitly asked about desired functionalities, but statements from other questions regarding the present situation and goals were also coded as "functional requirement". These statements were then formulated into actual functional requirements for the system. Similar statements were grouped into one functional requirement. The first iteration generated 20 functional requirements.

The functional requirements for the system were formulated as feature requirements. This is a straightforward way, and the requirements can later be directly translated into functions in the system [45]. The functional requirements were also formulated as user stories to understand the purpose of the requirement for the concerned stakeholders. An example of a functional requirement and traceability to interview statements and the corresponding user story is illustrated in Figure 4.1.

[1] An explanation on the importance of requirements elicitation can be found in Appendix A.3

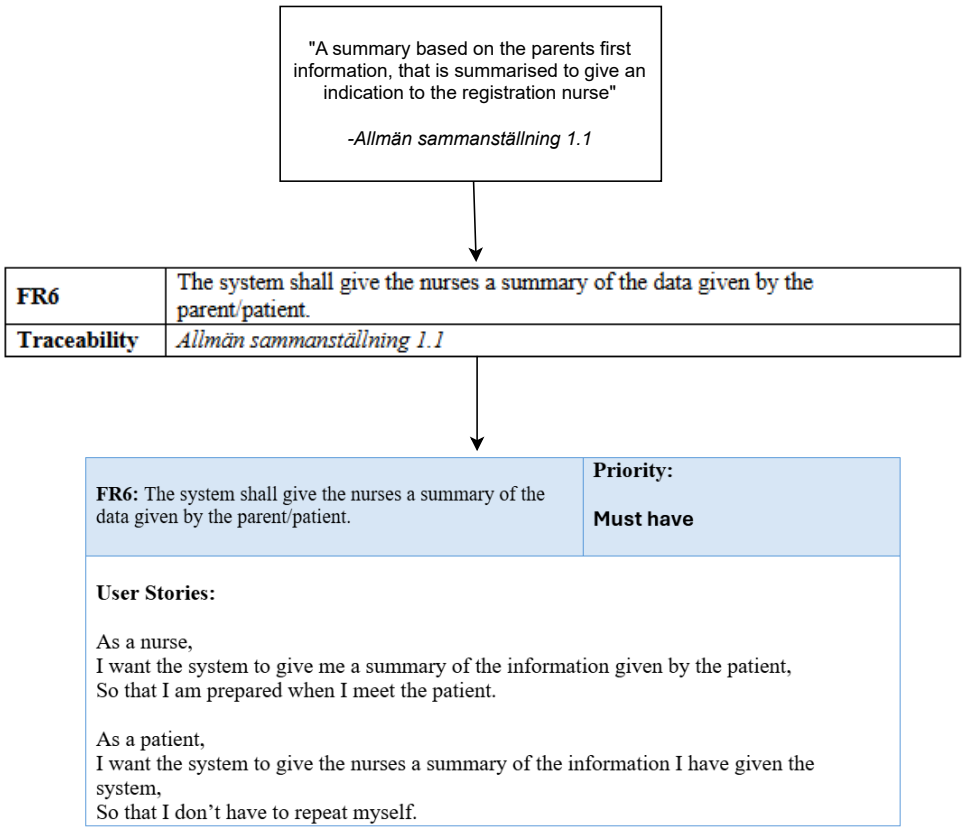

Figure 4.1: An example of a functional requirement FR6 with the corresponding user story and interview statement.

Requirements validation and prioritisation: As part of addressing RQ3 in this iteration, the elicited requirements were evaluated as part of interviews with interview groups two and three. After formulating requirements and user stories, it is important to validate that the requirements correctly reflect the stakeholder needs [45]. This was done by allowing the relevant stakeholders to read through the requirements and voice their opinions. For the stakeholders to be able to do this thoroughly, they must understand the requirements. To ensure this understanding, we as researchers read through the requirements and explained their purpose, as well as answered any questions that came up. The stakeholder then told us their opinions.

We had two cycles of requirement validations with two of the stakeholder groups. The first cycle was conducted as part of interviews with interview group 2. The purpose of this was to verify if the requirements were feasible both technically and ethically. During this cycle, four new requirements were formulated to ensure that the suggestions from the LLM are traceable. One example of such a requirement can be found in Figure 4.2. This requirement ensures that the nurses can look at the chat to see if what the LLM says or proposes is accurate, ensuring a safer and ethical use.

FR10: The system shall allow the nurse to look at the chat between the patient and LLM.	**Priority:** **Must have**
User Stories: As a nurse, I want to be able to look at the chat between the patient and the LLM, So that I can make sure that I don't miss any information.	

Figure 4.2: FR10: The system shall allow the nurse to look at the chat between the patient and LLM

The second cycle was conducted as part of the interview with interview group 3. This cycle aimed to confirm that the needs we had identified accurately reflected actual requirements at the PED. This cycle consisted of us explaining each requirement and then asking the stakeholder to voice their opinion, as well as ranking the requirements using the MoSCoW method. The MoSCoW method [49] is a method for ranking requirements using four levels. *Must have*, which means that the requirement must be satisfied for the solution to be considered a success. *Should have*, meaning that the requirement has a high priority and should be included if possible. *Could have* is a requirement that is desirable but not necessary. The last one is *Won't*, which means that the stakeholder thinks that the requirement should not be included in the solution. This allowed us to get an understanding of what requirements to focus on and what requirements should be removed from the project.

Inclusion Criteria

The input from the two validation cycles was formulated into three inclusion criteria (**IC1**, **IC2** and **IC3**) to determine which requirements should be dismissed in the first iteration, before starting to implement the software system. For a requirement to be considered, it had to meet all the criteria. All functional requirements can be found in Appendix A.3. The application of the first three criteria, **IC1, IC2** and **IC3**, resulted in not including 7 of the elicited functional requirements. **IC4**, **implementation scope**, was applied during iteration 2 to decide what requirements were technically feasible and prioritised to implement and evaluate in the prototype. After all of the inclusion criteria had been applied, the functional requirements were 8. This implies that, during iteration 2, 8 requirements were implemented into the prototype of the software system. The inclusion criteria table can be found in Table 4.2.

IC1 - Area of application: The purpose of the system is to provide a chatbot, used by a parent or patient in the waiting room after their first encounter with a registration nurse. After this, the system will provide the nurse with information and suggestions to support the triage process. The definition of

the area of application was agreed upon with the Head of Section.

IC2 - Isolated system: The software system should not interact with any other system within the hospital. No data should be fetched or sent outside of the implemented software system.

IC3 - No clinical prioritisation: The software system should not make any decisions regarding the prioritisation of patients. Thus, in no way influence their possibility to receive care.

IC4 - Implementation scope: The functional requirements that were technically feasible and prioritised to implement in the prototype. This criterion was added in iteration 2.

Table 4.2: Inclusion criteria table for elicited functional requirements from all three iterations of the study.

FR	IC1	IC2	IC3	IC4 (Iteration 2)
FR1	X	X	X	X
FR2	X	X	X	X
FR3	X	X	X	
FR4	X	X	X	X
FR5	X	X	X	X
FR6	X	X	X	X
FR7	X	X	X	
FR8	X	X	X	X
FR9	X	X	X	X
FR10	X	X	X	X
FR11			X	
FR12			X	
FR13			X	
FR14				
FR15		X		
FR16			X	
FR17				
FR18	X	X	X	
FR19	X	X	X	
FR20	X	X	X	
FR21	X	X	X	
FR22	X	X	X	

Prototype workshop: As part of addressing RQ2 in this iteration, a workshop was held to set up the development environment and to explore different ways to implement the prototype. The first part of the workshop was held by representatives from the AI Platform. During this part, the development environment within their infrastructure was configured to ensure secure data handling and access to their computing. Afterwards, different approaches to structuring LLM-based systems were discussed and tried out together with a data scientist from the AI Competence Centre at VGR.

4.3 Iteration 2: Solution phase

During the second iteration, the main focus was on RQ2, i.e., the solution phase, including prompt engineering and implementation of a working prototype. We continued the initial investigations of the first iteration on available LLMs and prompt engineering to find a combination that we could develop further in this iteration. This was done in addition to adding the inclusion criteria **IC4**. Thus, more of the functional requirements were excluded. The prompts used were further developed to reflect the requirements more accurately. As a result of this iteration, the artefact emerged, consisting of an LLM with a simple frontend. Workshops were held with stakeholders to get feedback on whether the model was working as specified in the requirements or not.

Infrastructure

Due to the project being in collaboration with VGR, the system is running on their Run:ai platform. Run:ai is a GPU orchestration and optimisation platform. It is maintained by the AI Platform team at VGR, and the project is thus kept within VGR's own network, which enables privacy protection and the utilisation of their computing resources. The development environment consists of Jupyter Notebooks within Visual Studio Code, using Python 3.11.0rc1 as the programming language. All other versions of frameworks and libraries used in the project are listed in Appendix A.1.

Model selection

For projects in the medical sector, it is important to ensure data security and patient integrity. Therefore, the LLM was implemented and ran only locally within the firewalls of VGR, which are classified for storing sensitive data.

We decided to use Llama 3.3 70B Instruct[2], as it was a model already available at VGR and approved to handle sensitive data. An alternative LLM could have been DeepSeek, but we disregarded it because we could not allocate additional resources within the VGR environment for a second model.

[2]For brevity, we shorten billion with B

Llama 3.3 70B Instruct: Llama 3.3. 70B Instruct is an instruction-tuned model, which means that it is trained with reinforcement learning from human feedback (RLHF). OpenAI used RLHF to train instruction-tuned models, which showed significantly improved instruction-following performance compared to standard models [50]. The software system that was developed was prompted with instructions to fulfil the different functional requirements. Therefore, an instruction-tuned model was considered a good choice.

The official model card of the Llama 3.3 model states that the instruction-tuned model is optimised for multilingual dialogue use cases and is specifically intended for assistant-like chat use. A central feature of our software system using an LLM is its chatbot functionality, implemented in Swedish. Thus, it seemed suitable to use a model trained for multilingual purposes. Llama 3.3 70B Instruct has also been shown to outperform many of the available open-source models on common industry benchmarks [51]. Therefore, the Llama 3.3 70B Instruct model was considered applicable for this study.

Prompt engineering technique selection - RAG

During the first iteration of prioritising the requirements, the data scientist pointed to the importance of not letting the LLM come up with suggestions on its own. Instead, it should base its suggestions and follow-up questions on existing medical documents. This is reflected in the functional requirement FR1, found in Figure A.1 in Appendix A.3. To achieve this behaviour, we decided to implement and use RAG. The purpose of this is to let an LLM search through the material provided in the RAG and use relevant sources to base its answers and follow-up questions on.

Structuring and sorting of data

To be able to implement a RAG, there was a need to restructure the medical documents in a way that was readable for the LLM. The relevant documents that were to be included were the WEST-P compendium and the triage guide, which can be found in Appendices A.7 and A.8. They were in the format of PowerPoint, containing tables and pictures, which cannot be processed by the LLM without modifications.

We chose to restructure the documents into text files in a structured HTML format, which is illustrated in Figure 4.3. There are several reasons why we chose this method. LLMs encountered many HTML documents during their training on large amounts of data. This leads to them possessing the ability to process such information in a good way [52].

Tan et al. [52] demonstrated that removing unnecessary content, such as CSS and empty tags, improves the LLM's understanding of the document's structure compared to plain text. The authors also state that research has shown that data in HTML format contains richer information compared to when it is in plain text. By converting the PowerPoints into this format, we allow the LLM to easily understand the structure of the content.

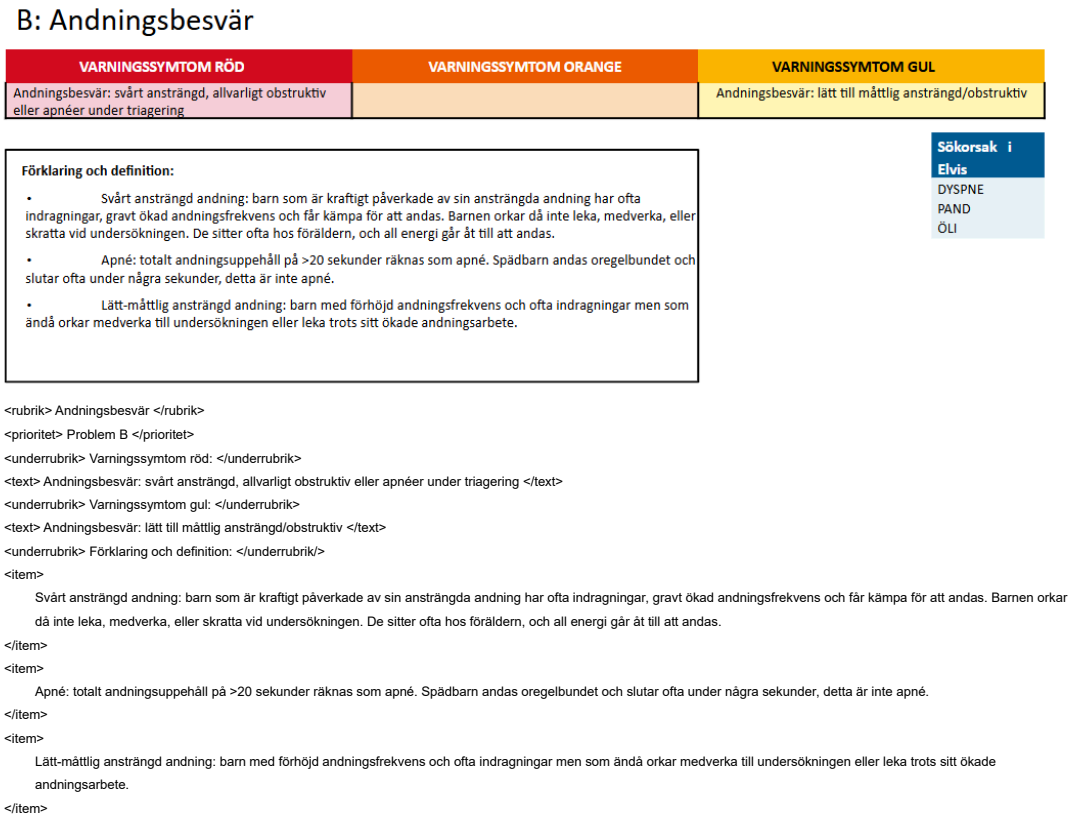

Figure 4.3: Conversion from PowerPoint to a structured HTML file.

Implementation of RAG

As part of the RAG development, the WEST-P compendium and the triage guide used by the staff at the Sahlgrenska PED were restructured into HTML format. The documents were split into chunks, where each chunk consisted of information about exactly one symptom or cause of visit. The embeddings model "KBLab/sentence-bert-swedish-cased" [53] from HuggingFaceEmbeddings, specifically trained on similarity searches on Swedish language, was used to map the chunks into a vector space, as illustrated in Figure 4.4. The numerical representation of the data was stored in a Facebook AI Similarity Search (FAISS) [54] vector store that holds embeddings to enable similarity searches.

The vector store is used in the RAG to provide relevant documents to enrich prompts. The process is illustrated in Figure 4.5. First, an LLM is prompted with the chat history and instructions to create a search query consisting of at most five words. After this, a retriever is invoked and uses the search query to find similar documents in the vector store. The two most similar document chunks are returned and used to enrich prompts.

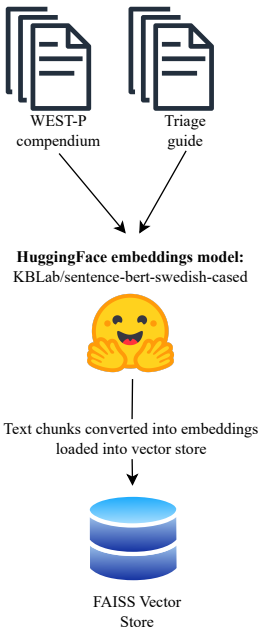

Figure 4.4: The process of loading documents into the FAISS Vector store

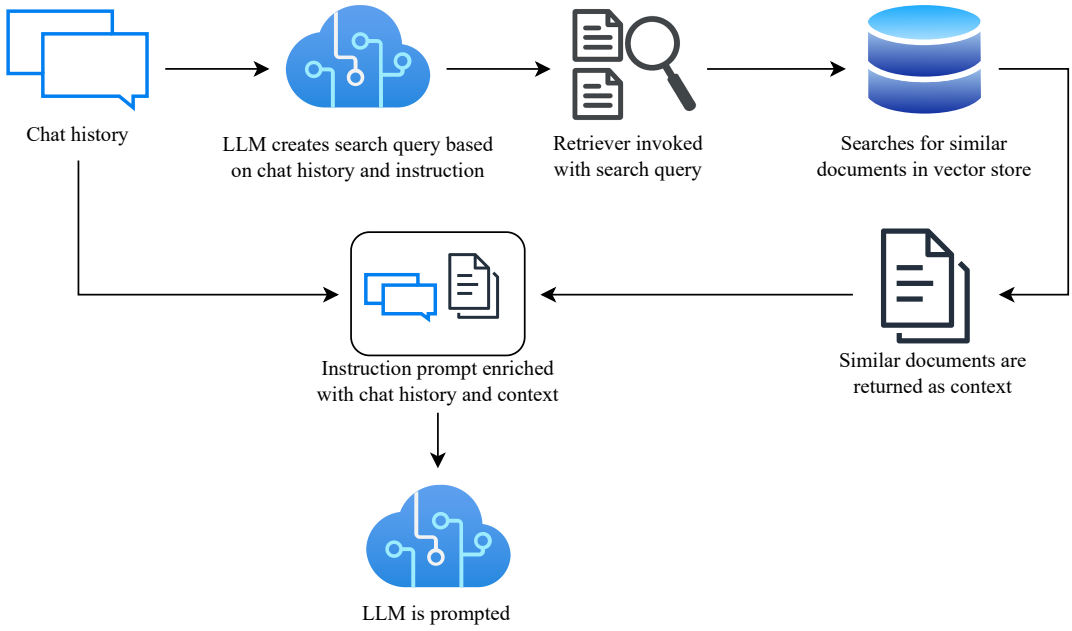

Figure 4.5: The RAG process

Prompt engineering approach

The initial prompt engineering approach adopted was to create a single system prompt comprising all the instructions that the LLM was required to follow. A context prompt, describing the role and purpose of the LLM, was also formulated and given to the LLM once upon its creation. The initial context prompt can be found in Listing 4.1, and the initial system prompt can be found in Listing 4.2. The original Swedish prompts can be found in Appendices A.2 and A.1. On each occasion that the LLM was prompted, the system prompt was provided in conjunction with the message history. The approach appeared to be inadequate due to the LLM's difficulty in adhering to all the instructions. It was frequently the case that the LLM disregarded some parts of the instructions in the system prompt whilst adhering to others. On occasion, it also failed to recall instructions during the chat session.

```
contextualize_q_prompt = ('''
    You are a chatbot at the Paediatric Emergency Department,
   gathering information from parents.
    You are not allowed to ask more than one question at a time.
    You are not allowed to speak any language other than Swedish.
    You are not allowed to say that something sounds concerning or
   serious.
    Continue asking questions until you have a clear understanding
   of the problem.
    When you have enough information, thank the user and end the
   chat.
    ''')
```

Listing 4.1: Prompt used for contextual question generation

```
system_prompt = ('''
    <searchresult>
    A search in the vector database resulted in the following
   matches:
    """
    {context}
    """
    - Use the information from the document extracts if the user
   seems to be asking about the document or if the information is
   relevant to answer the question.
    - You can ignore the document extracts if they do not seem to
   be about the user's question.
    - If the extracts do not contain relevant information, give an
   answer based on your general knowledge.

    *** Your role ***:
    - You are a chatbot at the Paediatric Emergency Department,
   helping nurses with gathering information from parents about
   their child's cause of visit.
    - You shall never say that you are going to contact a doctor or
    a nurse.
    - You are part of a web-based application, and therefore cannot
    bring the patient anywhere.


    *** Your task ***:
```

```
- You are ONLY ALLOWED to ask one question at a time.
- Ask only the question, without telling how you are thinking.
- Do not express any feelings or opinions.

''')
```

Listing 4.2: System prompt

At this point, the same LLM had several responsibilities. As shown in the example prompts, the LLM should determine when to stop, what documents were relevant as context and what the following question should be. In order to investigate whether or not the LLM found it difficult to follow the instructions due to a high level of responsibility, the instructions were broken down into smaller parts. Different classes were created with one responsibility each, using one LLM each. The first class had to formulate a search query based on the chat history. The second class was responsible for formulating a follow-up question based on the chat history and documents from RAG. The third class had to decide whether enough questions had been asked. Each of these classes had its own set of instructions, and they had no interaction or knowledge of each other. The final prompt engineering approach is presented in Chapter 5, the Results section.

Refinement of requirements

RQ1 was touched upon as well in the second iteration through refinements of some requirements, as well as the exclusion of some requirements.

Together with stakeholders, the requirement FR18, found in Appendix A.11, regarding the gathering of standard information through multiple-choice questions was moved outside of scope. This is due to the focus being on implementing the actual LLM interaction and information from the LLM to the nurses. However, FR18 was prioritised as "Should have", which is the second highest prioritisation. Therefore, a new requirement was created, FR8, found in Appendix A.6, where the standard information should be gathered but *not* through multiple choice questions. In the current implementation, the user is prompted to answer a few selected questions in normal chat format at the beginning of the chat before the LLM generates its questions.

IC4 was also applied to the requirements during this iteration. This inclusion criterion relates to the implementation scope. In order for a requirement to be implemented, it must fulfil the three earlier inclusion criteria, as well as this new one. Requirements that were discarded based on this criterion were either technically unfeasible or not prioritised within the time frame. FR21, found in Appendix A.14, is an example of a requirement that was considered both technically infeasible and deprioritised for this project. This was due to the amount of time required to implement the graphical interface not being justified by the value it would have added to the software. Another example is FR3, found in Appendix A.9, which was deprioritised. Due to not being allowed to intervene in the patient care in any way, we would not have been able to test this feature in the actual triage flow. Since we

could not evaluate it, it was not a priority to implement.

System-level simulations

RQ3 was addressed in this iteration by continuously evaluating the prototype through system-level simulations. Before testing the prototype in the real context of the Sahlgrenska PED, basic test cases were simulated to assess the implementation. This process was particularly important when formulating prompts to ensure that the system behaved as intended. By pretending to be a parent of a child visiting the PED, we were able to simulate various scenarios and evaluate how effectively the system fulfilled its intended tasks and requirements. One significant issue identified during these simulations was that the LLM occasionally included advice and interpretations in the answer. This behaviour was considered potentially harmful, particularly if the chatbot suggested to the patient that a condition was severe or alarming. To mitigate this risk, the instructions "You are NOT allowed to give advice or interpretations about conditions, symptoms or treatments" and "ONLY respond with a new follow-up question, nothing else" were added. These instructions can be seen in Listing 5.2.

4.4 Iteration 3: Evaluation phase

The third and final iteration focused on testing and evaluation, i.e., RQ3. A larger evaluation study was conducted in the real workflow with patients and nurses. Conclusions were drawn from this about how the chosen approach works. During this iteration, minor adjustments were made to the model in order to comply with iPads, thus touching upon RQ2 as well. After performing the evaluation study, requirements were added based on feedback from participants.

Ethical approval and consent

This study has received approval from the Head of Operations at Children's Medicine, Sahlgrenska University Hospital. The approval gives this study the possibility to evaluate the software application with patients at the PED who volunteered to participate in the trial. The approval can be found in Appendix A.26. All participants in this evaluation study have also signed a consent form, which can be found in Appendices A.24 and A.25.

Execution of user tests

During the third iteration, a human evaluation study was carried out. As presented in Section 2.3, human evaluation is suitable when evaluating non-standard natural language tasks. In such an evaluation, both experts and ordinary users should be included. The study took place at Sahlgrenska PED and involved both patients and nurses, patients being the ordinary users and nurses the experts. After an initial assessment by a registration nurse, the nurse told us which patients to ask about taking part in the study. The conditions for participation were that they had to be

able to read and write in Swedish, and that they had to have a low priority for triage, so that they had to wait for the nurse to further assess them. Participation was voluntary, and participants provided informed consent after receiving information both verbally and in writing. If they agreed to participate, the parent, together with the child if they wanted, would use the software application to chat with the LLM and then complete a short form about the experience. A nurse would then triage the patient as usual. After which, the nurse would look at the LLM-generated summary and suggestions on the summary page, as well as the chat and complete another form. To conduct the study ethically by not interfering with the patient care in any way, the LLM-generated summary and suggestions were presented to nurses strictly after the triage had been completed.

Forms

To evaluate the prototype, two separate forms were created. One for parents or patients, and one for nurses. These were partly designed to evaluate different aspects of the LLM: **Accuracy**, **Relevance**, **Fluency**, **Safety**, and **Human alignment**. They were also designed to evaluate the system as a whole through: **Usability**, **User experience** and **Provided value**. These evaluation aspects of LLMs and the software system as a whole are further described in Section 2.3. When deciding on what questions to ask, each aspect was taken into consideration and questions concerning it were formulated. For instance, questions regarding accuracy were formulated for the nurse evaluation form. Nurses, being clinical experts at the PED, are well-equipped to assess whether the information gathered or generated by the LLM aligns with the actual patient condition and clinical expectations.

Likert scale questions: All questions, except the last question of each form, were answered on a Likert scale. The respondents got six alternatives, whereof one was used to indicate that the respondent did not want to answer this question, and one to indicate a neutral answer. The answer "Do not want to answer" was incorporated to avoid the usage of the neutral answer in situations where the respondent simply did not want to answer or did not have an opinion on the question.

Parent and patient form: The parent and patient form addressed LLM aspects of **human alignment** by including questions about the overall experience and evaluating whether the chatbot-generated questions maintained a professional standard. Another aspect evaluated was **fluency**, by assessing the clarity of the chatbot's questions. The **safety** aspect was assessed by asking whether the questions were perceived as respectful. The parents also evaluated the **relevance** of the questions from their perspective.

Additional aspects of the software system were evaluated. The overall **user experience** was assessed by examining whether the user would feel safe if the application were part of the actual clinical process. The form also examined whether participants felt they had the opportunity to communicate everything they wished to express to medical staff. **Usability** was evaluated through a question addressing

how easy the chat interface was to use. Table 4.3 contains the questions from the parent form and their mapping to quality attributes.

Table 4.3: The questions asked in the form for parents are mapped to the corresponding quality attribute.

ID	Question	Attribute
Q1	It was easy to understand the questions asked by the chatbot.	**Fluency**
Q2	It was easy to understand how to use the chat page.	**Usability**
Q3	The chatbot asked relevant questions based on the cause of visit.	**Relevance**
Q4	I felt like I got the chance to mention everything that I would have wanted the nurse to know about my child.	**User experience**
Q5	I would feel safe if this chat process was a real part of the process at the Paediatric Emergency Department.	**User experience**
Q6	My experience with using the chatbot today was positive.	**Human alignment**
Q7	I felt that the chatbot asked professional and respectful questions.	**Safety**
Q8	Do you want to add anything else regarding the experience?	

Nurse form: In the nurses' evaluation, one of the main focuses was on assessing the **accuracy** of the LLM in all the tasks assigned. One task was to gather information from the patient by asking questions and then generate a summary for the nurse. To assess this, the nurse compared the summary generated by the LLM with the information they had collected independently, allowing an assessment of how accurately the LLM had captured and summarised the patient's condition. Similar assessments of **accuracy** were made for suggested next steps, controls and tests.

Relevance was another main focus of the nurses' evaluation. This aspect was assessed by examining whether the LLM-generated questions were appropriate in relation to the patient's reason for visit, whether a sufficient number of questions were asked, and whether the summary contained all the important information. Finally, a question was asked to evaluate the system as a whole. The question was whether the summary would have facilitated the triage process if it had been received in advance. By asking this question, it is possible to analyse whether or not the software system, as it is, provides any value to the nurses. Table 4.4 contains the questions from the nurse form and their mapping to corresponding quality attributes.

Elicitation of additional requirements

After the evaluation study, we had shorter discussions with the nurses who were participants in the study. The discussions were held for them to voice their opinions

about the functionalities of the software application. Some nurses suggested that the LLM should follow the same approach that they do when generating questions for the patient. The approach is to structure the triage to follow the SAMPLE [55] and OPQRST [56] guidelines. These are mnemonic rules, used to remember all important aspects during triage. Therefore, FR22 A.23 was added to reflect this wish. Another refinement was regarding NF10, where SAMPLE and OPQRST were also included as relevant frameworks for generating follow-up questions.

Table 4.4: The questions asked in the form for nurses are mapped to the corresponding quality attribute

ID	Question	Attribute
Q1	The suggested next steps in the treatment are consistent with the steps I carried out.	**Accuracy**
Q2	The suggested controls and tests are consistent with the ones I carried out.	**Accuracy**
Q3	The summary is consistent with what I summarised about the patient.	**Accuracy**
Q4	The questions the chatbot asked were relevant based on the patient's cause of visit.	**Relevance**
Q5	The chatbot asked the patient enough questions.	**Relevance**
Q6	The summary contains all important information from the chat conversation.	**Accuracy, Relevance**
Q7	It was easy to take in the information from the summary.	**Fluency**
Q8	Receiving this information before performing triage would have facilitated the process.	**Contributed value**

5

Results

5.1 RQ1: Stakeholders' characteristics and requirements

RQ1 is concerned with what the characteristics and requirements the stakeholders, at the PED at Sahlgrenska University Hospital, have for a software using an LLM. For the software application using an LLM at the Sahlgrenska PED, five main stakeholders have been identified. These are presented in Table 5.1. Two of these stakeholders were categorised as being highly impacted by the software, as both are potential users of the system at the Sahlgrenska PED, while the remaining three stakeholders are not using the software on a daily basis [1]. The analysis focuses on the characteristics of these two stakeholders.

Indications for areas for improvement

The stakeholder interviews included questions about the current situation and what areas can be improved in order to identify how a software application using an LLM can improve the workflow at the PED. In the fish bone diagram illustrated in Figure 5.1, the main areas for improvement in the PED are categorised into four themes, as described by the stakeholders and reflecting their perspectives and characteristics. These are areas where improvements can be made to enhance both the work environment of the personnel and the experience of patients.

Nurse performance: One of the improvement areas, *Nurse performance* describes characteristics of the stakeholder group, nurses, that may influence their ability to currently perform effectively. One aspect in this area is *subjectiveness*. One nurse explained that human factors can cause subjective judgements of patients based on previous interactions between the patient and the nurse. As a consequence, nurses can be inconsistent in assessing patients.

> *"That I can be touched by a patient due to other circumstances. Which means that I may be a little more vague in my judgement"* - Nurse 2

[1] One is a developer, one is a manager at the organisation that provides the infrastructure and one is head of section.

Table 5.1: Stakeholder map for the project. Impact level refers to how much the system will affect the stakeholder, and influence level refers to their ability to affect the system's design.

Name (Stakeholder Type)	**Topics of Interest**	**Impact Level**	**Influence Level**
Head of Section (Project owner)	Patient safety and quality of care. Alignment with department goals. Regulatory compliance and ethics.	Low	High
Nurses (Potential users)	Accuracy and clarity of patient-provided information. Workflow integration and time-saving potential.	High	Medium
Patients/Parents (Potential users)	Accessibility and ease of use. Clearly stated purpose. Trust and safety.	High	Low
AI Platform (Infrastructure provider)	Provide a secure platform to deploy software within VGR.	Low	Medium
Competence Centre AI (Technical and regulatory support)	Support the development and adoption of AI in clinical practice.	Low	Medium

Another characteristic of the nurses is their vulnerability to being affected by *external factors*. One nurse mentioned that even though external factors should not influence their judgement, it sometimes does. If there is a very high number of patients on a certain day and it is a stressful environment, that might affect the assessments that nurses make.

> *"But that you might unconsciously downplay or exaggerate symptoms depending on what a working day looks like. Depending on whether there are 200 applicants a day, or whether it is freezing cold in the rooms, or whether you are having a bad day."* - Nurse 2

The nurses are also characterised by their exposure to consistently high workloads in the PED. Something which they identified as contributing to *fatigue* among nurses and physicians. This is another human factor that can lead to mistakes during assessments and work in general.

> *"And the fact that after seven hours, if it's a high flow, you get a bit. Yes, it's called decision fatigue. I can't make another decision. I can't make another decision because I don't know, I can't. I don't know what I'm doing any more. And then you somehow get a worse triage process because I've heard too much. I can't think any more."* - Nurse 2

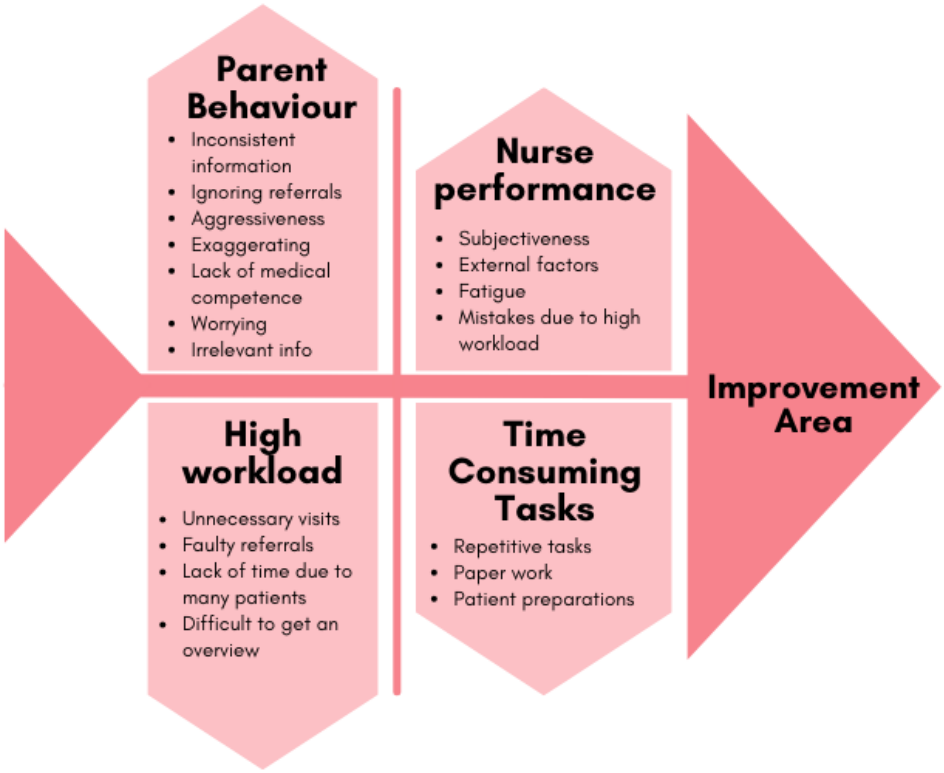

Figure 5.1: Fish bone diagram of the potential areas of improvement at the PED.

Parent behaviour: Characteristics of parents as stakeholders include strong emotional involvement, together with limited medical knowledge. This can lead to an overemphasis on symptom severity when communicating with the nurses. This highlights a potential area where improvements could be made.

> *"For example, parents will say that their child is tired and lethargic. But when the parents here say their children are tired and lethargic, they are not tired and lethargic in our eyes. They are tired and lethargic to the parents. Absolutely, they are, because they are more tired than they usually are. But that's not the same thing as lethargic in the eyes of the health care."* - Nurse 5

> *"Because what you notice quite often when you talk to the parents is that what they say is not true. Some of them exaggerate a bit, so that can also be a risk, for example."* - Nurse 1

An additional characteristic of parents, as described by nurses, is that the information they provide is often inconsistent. The registration nurse may receive one version of the story, while another version is told to the triage nurse and yet another to the physician, causing confusion. Nurses also note that some of the information can be irrelevant. One example given by a nurse is that a parent might talk about what happened when their 14-year-old child was born, which can be completely

irrelevant to the reason for the current visit.

Time-consuming tasks: Several nurses described spending a significant portion of their time on tasks that do not require their full professional competence. Three recurring categories emerged:

Repetitive information gathering: Nurses frequently need to ask patients and caregivers for background information to complete assessments. This was described as time-consuming and could potentially be streamlined if such information were collected in advance, during patient waiting time, and made available digitally, allowing nurses to review rather than repeatedly asking for the same details.

Administrative work: Paperwork was identified as another source of inefficiency. Nurses noted that much of the documentation is still handled manually, resulting in time lost to managing physical forms and deciphering unclear handwriting. Digital solutions were suggested as a clear opportunity for improvement.

Patient preparation: Depending on the patient's condition, certain actions should be taken before triage. For instance, patients needing surgery should fast, while others may benefit from eating or drinking. These instructions are often not provided until the nurse interaction, leading to missed opportunities. A system, including an LLM, that could provide personalised, condition-specific advice while patients wait could enhance efficiency and improve clinical outcomes.

Potential risks when adopting an LLM at the PED

The fish bone diagram in Figure 5.2 illustrates the potential risks of using an LLM at the PED based on the conducted interviews. The risks are divided into four main areas.

LLM limitations: Some of the risks are traced to the limitations of LLMs in comparison to clinical staff, particularly in the area referred to as *lacks clinical vision*. This contrasts with a key characteristic of nurses, who consistently rely on their clinical vision when assessing patients. The absence of this capability in an LLM is considered a potential risk when integrating such systems into the PED workflow. Clinical vision includes interpreting non-verbal cues such as body language, facial expressions, and general appearance. These are factors that are inaccessible to an LLM. As a result, the inability of an LLM to "see" the patient is perceived as a significant risk, as it may lead to less accurate assessments. In line with this, the LLM may also fail to detect exaggerations or misconceptions.

Depending on the implementation, a risk mentioned was *over-triage*, which is not a risk for the patient concerned, however, this could increase the workload and other patients in need of care might be affected with longer waiting times.

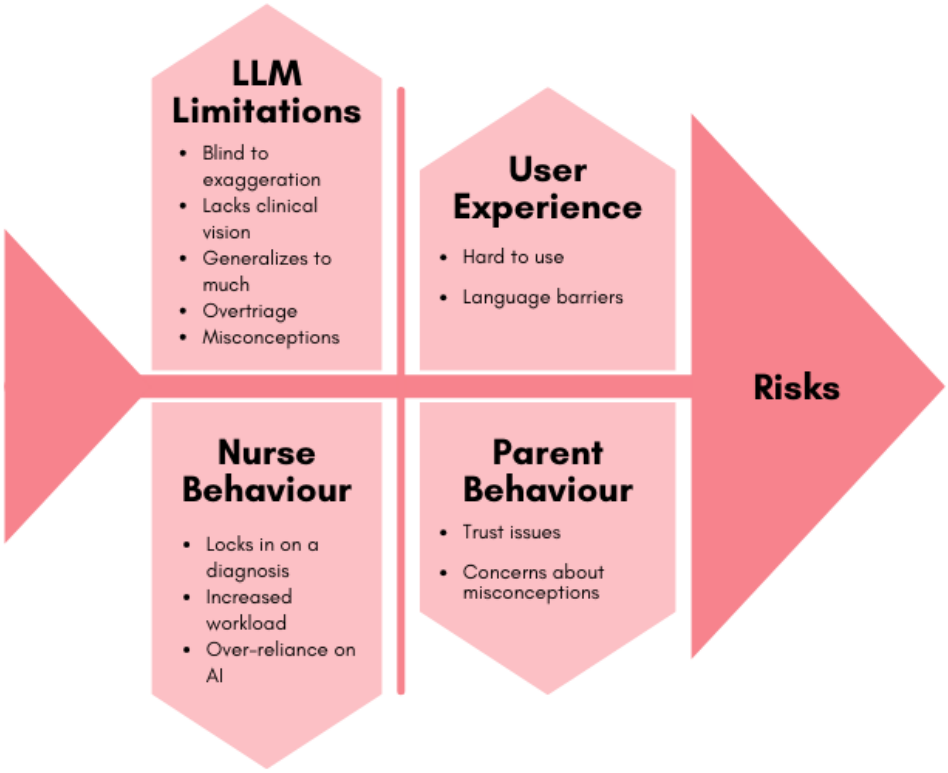

Figure 5.2: Fish bone diagram of the potential risks of using an LLM at the PED.

> *"The risks that I see above all are that the children are over-triaged and that is not really dangerous for the patient. But it can mean that more patients are assessed as red and that seriously ill children have to wait, so there is a risk with that, but not dangerous for the individual patient."*
> - Nurse 6

Nurse behaviour: The introduction of an LLM-chat in the triage process could influence certain behavioural tendencies among nurses. One concern raised was the risk of *over-reliance on the system*. Nurses described the possibility of developing a tendency to rely too heavily on the LLM's suggestions, potentially at the expense of their own clinical judgment and intuition.

> *"That you trust the system too much and forget your own knowledge. That is what I worry about even though it could help a lot along the way. But you work a lot with gut feeling and clinical eye. So there is a risk that you forget about that."* - Nurse 1

This indicates a risk of shifting from a proactive to a more passive decision-making style, where professional instinct and experience may be undervalued in favour of system output. Another concern was the possible emergence of increased performance expectations. Some nurses expressed that removing certain tasks through automation might lead to an expectation to handle more patients in less time. This

could result in a perceived intensification of workload, potentially increasing stress in an already pressured environment.

User experience: Additional risks relate to the user experience, particularly for parents interacting with the system. One identified concern is the potential for misunderstandings when using a digital chat interface instead of direct communication with staff. Certain characteristics among users, such as limited literacy or discomfort with written communication, could make interaction with a chatbot challenging.

An occurring characteristic of parents and patients in paediatric emergency departments is a lack of knowledge of Swedish or English. This leads to *language barriers* being a recurring issue at the PED. Nurses report that some parents already struggle to complete registration forms due to limited language or writing skills. There is a concern that requiring users to describe medical issues in writing could exclude or disadvantage those who rely on non-verbal communication, such as pointing or gestures, which are commonly used in face-to-face interactions. These limitations suggest a risk of reduced accessibility and increased frustration for some user groups.

Parent behaviour: Some risks relate to how parents may respond to the shift in responsibility when using a digital tool. A key concern is that parents might feel uncertain about their ability to accurately describe their child's condition. This uncertainty can create stress, especially when parents perceive themselves as responsible for making a preliminary assessment without medical training.

> *"Did I really make it clear how my child is feeling, and that it can become a stress for the parent that: it is me that is responsible in some way for assessing my child. And I do not have any medical knowledge. Have I seen what is important or do I only see unimportant symptoms? That a person with medical education would react to differently. I could imagine"*
> - Nurse 2

Another behavioural risk involves trust in the system. Some nurses noted that a characteristic of highly concerned parents is that they tend to seek care regardless of reassurance from professionals. These parents may be unlikely to trust or follow the recommendations of a digital assessment tool if it concludes that their child does not require urgent care. This reflects a behavioural pattern in parents where emotional concern overrides rational guidance, posing challenges for adoption and reliance on automated triage.

Potential areas of usage for software using an LLM at a PED

From the interviews, several areas of potential usage are identified, as shown in the fish bone diagram in Figure 5.3. Out of six areas, three are dismissed as out of scope. These can be seen as greyed out in the figure and will not be further presented in this thesis.

Before triage: An area of application identified for a software system using an LLM is to perform tasks before triage. By making the passive waiting time of patients active, the necessary background information can be gathered and a compilation of information created for the nurse before starting the triage.

> *"I believe that it could benefit us if parents can fill in, in advance, anamnestic data. That would give us a head start in our assessment of the child"*- Nurse 5

> *"It can gather relevant information and sort of write it up so that you save that time instead of the nurse having to talk to the parents for ten minutes and the parent maybe having to call someone during that time to check."*- Nurse 5

Figure 5.3: Fish bone diagram of the potential areas of usage for an LLM at the PED. Grey parts are areas considered out of scope.

Registration: Currently, all patients fill out a blue registration form and give it to the registration nurse upon arrival at the PED. This is done to get basic information about the patient and register the patient as waiting for triage. This was identified as an area that potentially could be replaced by software using an LLM.

Standardised process: All patients are often asked similar questions before being asked more specific questions regarding their symptoms. During the interviews, this was also identified as an area where software using an LLM could ask these standard questions before the nurse starts to ask more specific questions.

> *"Yes, but if you do this before the registration, then we might not have to ask so many questions. Then, maybe it takes half a minute instead of*

> *five. Or if you get it [the information] before triage, you might not need to ask all the questions three times. Then you have it on paper. Then you can just ask, Is this correct?"*- Nurse 4

Functional requirements

The designated location for the chat interaction is in the waiting room, after registration by the nurse and before the triage assessment. The parent or patient uses the software application to chat with an LLM. At the end of the chat, the LLM generates a summary of the patient's condition and the reason for the visit, and suggests initial actions and controls to the nurse. The 15 functional requirements that were identified as relevant for the system based on the stakeholder needs and characteristics are presented in Table 5.3 together with their implementation, if applicable.

Non-functional requirements

This thesis has also identified non-functional requirements for the system. The 10 non-functional requirements that were identified as relevant for the system based on the stakeholder needs and characteristics are presented in Table 5.4. Some of the non-functional requirements can be directly linked to a functional requirement, as illustrated in the table.

5.2 RQ2: Development processes for software using an LLM

RQ2 examines how requirements can be translated into software development processes using an LLM, with a focus on model selection and prompt engineering. The answer to this research question is provided by the artefact that was created as part of this design science study. We will therefore explain the artefact's different elements, and how the requirements map to the final implementation.

The artefact consists of four instances of the Llama 3.3 70B Instruct model, which is specifically trained for instruction following, a key capability for meeting our requirements. The four instances are presented in Table 5.2. This section also presents the final prompts that define the behaviour of each instance. In addition, the artefact includes a web application with four distinct pages, which will be described in detail later in this section.

Interaction between LLM instances

As mentioned above, the software system utilises four instances of the Llama 3.3 70B Instruct. Figure 5.4 illustrates the process starting after the user has been prompted to answer all standard questions. The standard questions and answers are added to the chat history. The first step is to prompt the Search Query LLM with the `search_query_prompt`, found in Listing 5.1, and the chat history. The

Table 5.2: Table outlining the four LLM instances' responsibilities and where to find their prompts.

LLM Instance	Responsibility	Prompt
Search Query LLM	Generate optimized search queries	Listing 5.1
Question LLM	Formulate new follow-up questions	Listing 5.2
Done Checker LLM	Assess completeness of collected data	Listing 5.3
Summary LLM	Generate structured summary and suggestions	Listing 5.4

Search Query LLM generates a search query and invokes the RAG process. This results in extracted relevant documents, called rag results in the diagram. After this, the Done Checker LLM is prompted with `done_checker_prompt`, found in Listing 5.3, the rag results and the chat history. The Done Checker LLM generates a boolean output, based on if all necessary information has been gathered. If it has, the next step is to invoke the Summary LLM with the `summary_prompt`, found in Listing 5.4, rag results and chat history. The Summary LLM produces a summary and suggestions of controls and tests. If Done Checker LLM thinks more information is needed, the boolean attribute is `False`. The next step is to invoke the Question LLM with the `question_prompt`, found in Listing 5.2, rag results and chat history to generate a follow-up question. Both the follow-up question and the answer from the user is added to the chat history. Then the Search Query LLM is invoked again with the updated chat history and the process restarts.

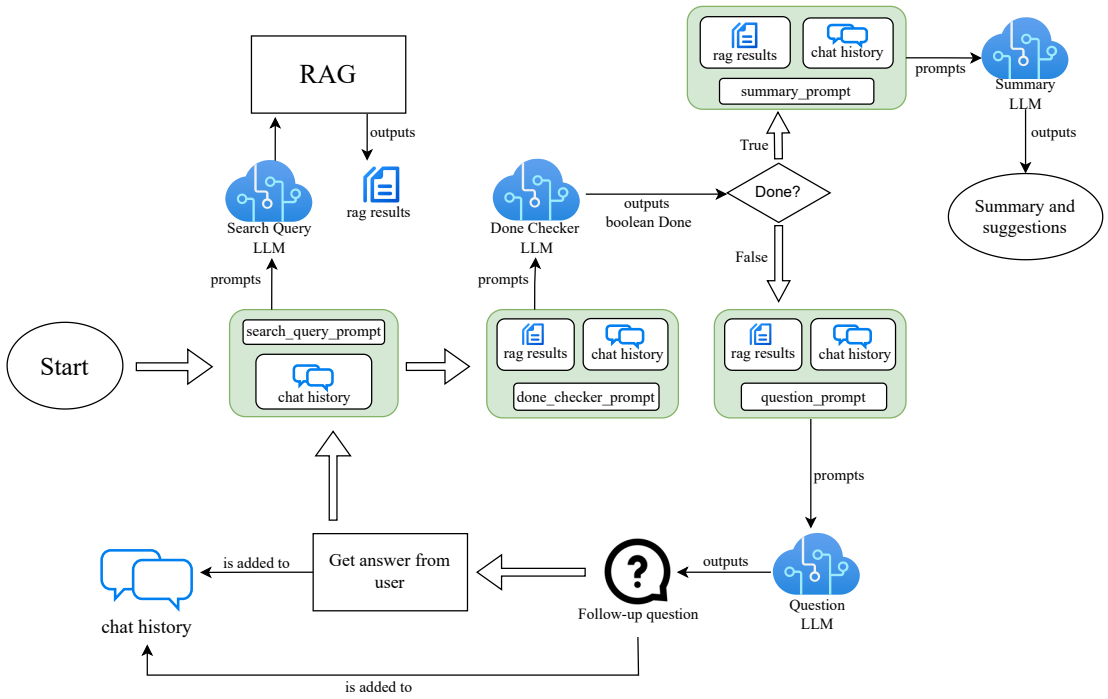

Figure 5.4: A diagram of the LLM interactions in the software system.

Final prompts

The final prompts were formulated to contain instructions for one specific task, performed by one specific instance of the LLM. The four different prompts used in the final implementation can be seen in Listings 5.1, 5.2, 5.3 and 5.4.

Listing 5.1 shows the prompt given to the Search Query LLM. The LLM is provided with the chat history, which it is instructed to use to formulate a relevant search query. This query is then used to search the vector database for relevant document extracts. These documents are provided as context to other LLMs performing tasks such as generating a follow-up question or creating a summary and suggestions for treatments and controls. Collecting and using these document extracts is a necessary step toward fulfilling **NF10**, linked to **FR1**, and could possibly improve the level of compliance of **NF6** linked to **FR2**, and **NF5** linked to **FR5**.

```
search_query_prompt =
f"""
<instructions>
**Approach**
    - Formulate a search_query with a maximum of 5 well-chosen
   words based on the patient's symptoms and reason for visit to
   get information about more questions that should be asked to the
    patient.
    - The documents you will search contain information on the
   actions, symptoms and treatment for different causes of
   paediatric visits.
    - You should therefore not include words like children,
   symptoms, causes, treatment and similar in your search_query.
</instructions>
"""
```

Listing 5.1: Prompt for generating a search query

Listing 5.2 shows the prompt given to the Question LLM. The LLM receives the chat history to see what has already been said in the conversation, as well as document extracts from the vector database. Using this information, as well as noting what is missing from the chat history, the LLM formulates a new question for the user. Providing the Question LLM with relevant documents before generating the next question is necessary to enable the fulfilment **NF10**, linked to **FR1**.

```
question_prompt =
f"""
<instructions>
You will receive:
- A chat history with messages between a patient and a chatbot.
- A number of document extracts that can consist of relevant
   medical information

**Your role**
- You are a chatbot at the Paediatric Emergency Department with the
    purpose of gathering supplementary information for a nurse.
```

```
- You are NOT allowed to give advice or interpretations about
    conditions, symptoms or treatments.
- Your only task is to ask ONE new and relevant follow-up question
    for the patient.

**Limitations**
- You are NOT allowed to repeat any question already asked in the
    chat history.
- Read the chat history carefully to ensure this.
- If a question is too similar to another question, you need to
    reformulate it or choose another focus.

**Approach**
1. Analyse the whole chat history:
- Identify already asked questions and answers.
2. Read the document extracts.
- If they contain relevant information: use it to formulate a new
    and relevant follow-up question.
-If they are not relevant: base your question on what is missing in
     the chat history.
3. Formulate a **short, clear and unique** follow-up question for
    the patient.

**Output**
- ONLY respond with a new follow-up question, nothing else.
-----------------------------------------------
Chathistory:
{messages}

Document extracts:
{rag_results}
</instructions>
"""
```

Listing 5.2: Prompt for generating a follow-up question

Listing 5.3 shows the prompt given to the Done Checker LLM. This LLM is invoked after each question to decide whether to formulate another question. The LLM receives both the chat history and the current document extracts. It examines the documents and, based on the information therein, decides whether the information in the chat history is sufficient. If so, the boolean attribute `done` will be set to `True`, prompting the software system to stop the loop that calls upon the Question LLM.

```
done_checker_prompt =
f"""
<instructions>
You will receive messages from the chat history.
- Set done to True if you feel you have sufficient information
    about the patient's condition based on the document extracts and
     chat history.
- Write a description of why you think you have sufficient or
    insufficient information.

Document extracts:
{rag_results}
```

```
**Approach**
- Analyse the chat history to see what has already been said.
- Analyse the document extracts to see what more information should
    be collected.
- If you think enough information has been collected, set done to
   True.

**Your task**
- You are a chatbot in the paediatric emergency room collecting
   information for a nurse.
</instructions>
"""
```

Listing 5.3: Prompt for checking if the LLM has gathered enough information

Listing 5.4 shows the prompt given to the Summary LLM. This occurs when the chat ends. The LLM is provided with both the chat history and the most recently extracted document extracts. Based on these, it is tasked with generating a summary from which nurses will benefit, as well as suggestions for next steps in treatment, tests and controls. A schema detailing how the LLM should structure its output is used to generate this summary, which can be seen in Listing 5.5. The different variables in the schema are referenced in the summary prompt, as shown. The summary LLM implements **FR2**, **FR5** and **FR6**.

```
summary_prompt =
f"""
***Your task***
You shall create a summary in multiple parts that can help a nurse
   at the paediatric emergency room get an overview of a patient's
   condition.
To help, you have a chat history between the patient and a chatbot
   as well as document extracts about medical information, and the
   work at the paediatric emergency room.

***Approach***
- Generate a summary of the chat that a nurse would benefit from in
    the field summary.
- Give between 1 and 3 suggestions for the next step of the
   treatment in the field next_steps.
- Give suggestions for supplementary tests and controls in cases
   where they are necessary for field controls.
-------------------------------------------------------
Chathistory:
{messages}

Document extracts:
{rag_results}
"""
```

Listing 5.4: Prompt for generating a summary

```
class SummarySchema(pydantic.BaseModel):
    chat_summary : str = pydantic.Field(description="A summary of
    the chat between a patient and the chatbot")
    next_steps : list[str] = pydantic.Field(description="A list
    with three suggestions of the next steps in the treatment")
    controls : list[str] = pydantic.Field(description="A list of
    tests and controls that the nurse needs to do")
```

Listing 5.5: Schema for structuring a summary

Prototype of the web application

The web application consists of a landing page for patients, which can be seen in Figure 5.5. On this page, the user is prompted to enter a code in order to start the chat with the chatbot. After the code has been entered, the user is taken to the chat page which can be seen in Figure 5.6. First, the user will have to answer some standard questions:

- What is the first name of your child?
- What is the cause of visit?
- How old is your child?
- How much does your child weigh?
- Does your child have any long-term diseases or diagnosis?
- Does your child have any known allergies?

The standard questions are based on the registration form. However, to be able to use the implementation in the evaluation study, some questions asking for sensitive data were left out of the implementation.

After the user has answered all standard questions, the software system will use LLMs to generate follow-up questions as illustrated in Figure 5.7. At the end of the chat session, the user is prompted with the question: "*Is there anything else you would like to add to the chat before I summarise it?*". This is shown in Figure 5.8.

The two other pages of the web application are views for the nurses. The first page consists of a summary of the chat session, suggestions for next steps in the treatment and additional controls and tests to take. This page is illustrated in Figure 5.9. From the summary page, the nurse can navigate to a second page where the chat history between the patient and the chatbot can be viewed, as shown in Figure 5.10. This allows the nurse to see the questions and answers that the summary and suggestions are based on.

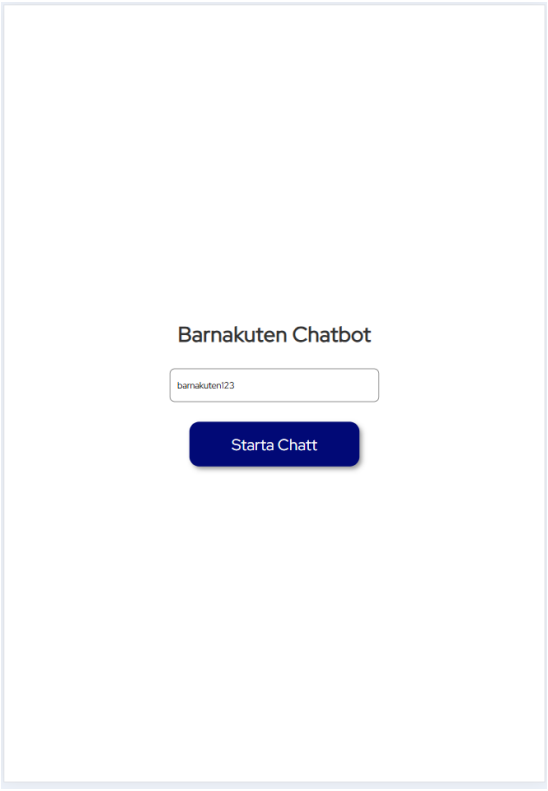

Figure 5.5: Landing page of the web application.

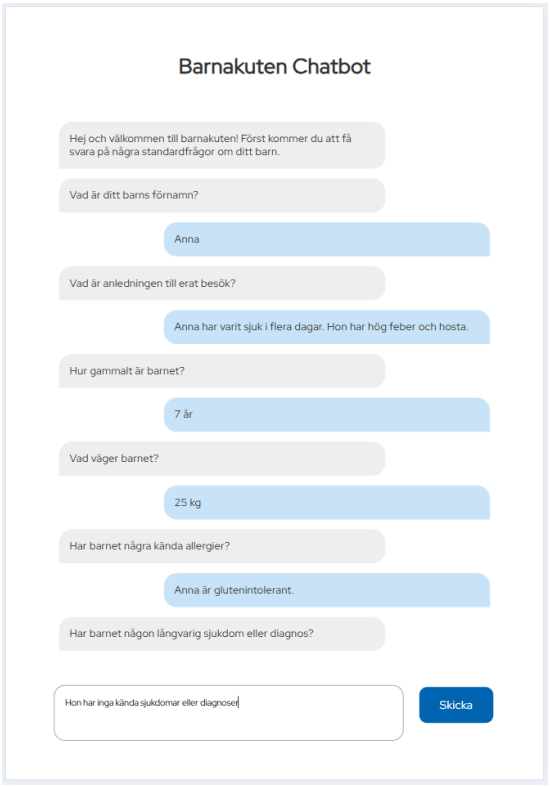

Figure 5.6: Standard questions asked by the chatbot.

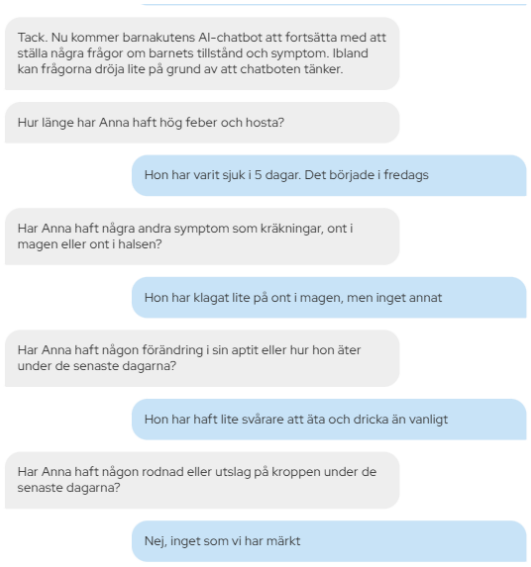

Figure 5.7: Chat example between an imaginary patient and the chatbot.

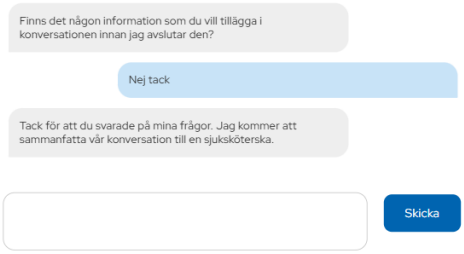

Figure 5.8: The chatbot asking if the user wants to add any more information before the conversation is terminated.

Figure 5.9: The image shows the LLM-generated summary of the chat.

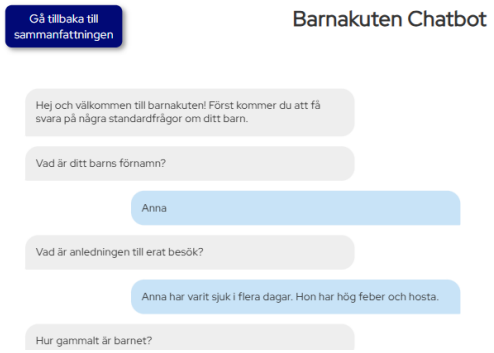

Figure 5.10: The picture shows the nurse view of the chat between the patient and nurse.

Requirements mapped to implementation

As stated in Section 5.1, 15 functional requirements were identified based on stakeholder characteristics and needs. After applying the fourth inclusion criterion, **IC4**, eight requirements remained to be implemented in the software application. These requirements are mapped to their implementation in Table 5.3.

The non-functional requirements identified for the system are listed in Table 5.4. Requirements NF1–NF6 have been considered and implemented in the software system. While these requirements are not directly verifiable through specific functionalities, they are reflected in the design choices made for the LLM integration and the behaviour of the final system, and their fulfilment has been evaluated in the evaluation study.

In contrast, requirements NF7–NF10 were not addressed in the implementation. The main reason for their exclusion aligns with the rationale for omitting certain functional requirements, as outlined in **IC4**, namely technical feasibility or lack of prioritisation. For example, NF7 was identified during the interviews as a necessary aspect for integrating such a system into the actual workflow. However, within the time frame and scope of this study, achieving the required level of availability was not technically feasible.

As for NF8 and NF9, since the evaluation did not allow any intervention in actual patient care, implementing these specific requirements was not critical and therefore not prioritised for this study. Nevertheless, these requirements are critical for any medical software system, and become essential if the system is intended for deployment in a real clinical setting. NF10 was added after the evaluation study and, therefore not implemented into the system.

Table 5.3: The stakeholder's functional requirements for software using an LLM in a PED at Sahlgrenska University Hospital, along with the implementation of requirements that fulfilled all inclusion criteria. FRs11-17 are not included because them being excluded according to **IC1-3**. The FRs without implementation descriptions were excluded due to **IC4**.

ID	Functional requirement
FR1	The system shall generate follow-up questions based on the cause of visit.
Implementation: Implemented by prompting an LLM to generate a follow-up question after receiving a response from the user.	
FR2	The system shall propose simple first actions after triage to prepare the patient for treatment.
Implementation: Implemented by prompting an LLM to generate suggestions on what tests and controls should be performed next. This is done by giving the LLM the chat history as well as the results from the latest usage of RAG. These suggestions are presented on the summary page.	
FR3	The system shall propose initial actions for the patient while waiting for triage.
FR4	The system shall gather anamnetic data from the patients.
Implementation: Implemented by prompting an LLM to have the responsibility of gathering supplementary information about a patient and their cause of visit for a nurse.	
FR5	The system shall provide suggestions for what additional information is necessary to gather.
Implementation: Implemented by prompting an LLM to generate suggestions on what additional information the nurse should gather. The suggestions are presented on the summary page.	
FR6	The system shall give the nurses a summary of the data given by the parent/patient.
Implementation: Implemented by prompting an LLM to generate a summary based on the chat history. This summary is then presented on the summary page.	
FR7	The system shall provide the nurses with a list of possible diagnoses based on the current information.
FR8	The system shall gather basic information through standard questions.
Implementation: Implemented by having a number of set questions that all users are prompted to answer at the beginning of the chat.	
FR9	The system shall gather additional optional information through text questions.
Implementation: Implemented by letting the user add any additional information at the end of the chat. This is done by having the chatbot ask: *"Is there anything else you would like to add to the chat before I summarise it?"*.	
FR10	The system shall allow the nurse to look at the chat between the patient and LLM.

	Implementation: Implemented by creating a chat-history page for the nurse to navigate to from the summary page.
FR18	The system shall gather basic information through standard multiple-choice questions.
FR19	The system shall allow the user to view and edit the summary generated by the LLM before sending it to a nurse.
FR20	The system shall have hyperlinks between sections in the summary, generated by the LLM, and the relevant part in the chat between the patient and the LLM.
FR21	The system shall include highlighted segments and images in the summary generated by the LLM.
FR22	The system shall gather patient information according to the SAMPLE and OPQRST guidelines.

Table 5.4: The stakeholders' non-functional requirements, their implementation, and any associated corresponding functional requirements.

ID (FR)	**Non-functional Requirement**
NF1	The patient data has to be stored within a secure system classified for storing sensitive data.
	Implementation: To fulfil this, the software system must be deployed within an infrastructure classified for handling sensitive data. This prototype was deployed within the infrastructure provided by the AI platform at VGR, classified for handling and storing patient data.
NF2	A patient with Swedish reading and writing skills has to be able to use the software application without help from staff.
	Implementation: To fulfil this, it is of importance to use a graphical interface (GUI) that is easy to understand for first-time users. In this prototype, this is achieved by using a standard chat GUI that most users have seen in other applications before.
NF3	The application should be accessed from the browser using an iPad connected to the VGR network.
	Implementation: The GUIs of this application were developed for an iPad.
NF4 (FR6)	The summary generated has to be accurate based on the chat and include all important information.
	Implementation: The LLM is prompted to generate a summary based on the chat conversation between the parent or patient and an LLM.
NF5 (FR5)	The suggestions of controls and tests have to be accurate.

Implementation: To achieve a higher level of accuracy, the LLM is prompted to use relevant documents such as the WEST-P compendium or triage guide, containing accurate controls and tests for certain causes of visits, when creating suggestions.

NF6 (FR2)	The suggestions of next steps in treatment have to be accurate.

Implementation: To achieve a higher level of accuracy, the LLM is prompted to use relevant documents such as the WEST-P compendium or triage guide, containing accurate next steps in treatment for certain causes of visits, when creating suggestions.

NF7	The system needs to be available 24/7.

Implementation: A system with 24/7 availability requires having people on-call to serve the system in case of issues occurring. The PED is available 24/7 for helping children in need of urgent care, and if a software system is integrated into the flow, it needs to have as high availability as the PED. Another possible implementation is using the current manual system in case the software has downtime.

NF8	The system needs to follow the Regulation (EU) 2017/745.

Implementation: If this software system were to be implemented into the workflow at the Sahlgrenska PED or any other hospital within the EU, it would need to comply with the Regulation (EU) 2017/745.

NF9	The system needs to follow Swedish national regulations regarding medical devices.

Implementation: If this software system were to be implemented into the workflow at the Sahlgrenska PED or any other hospital in Sweden, it would need to comply with the national regulations regarding medical devices.

NF10 (FR1, FR22)	The generated follow-up questions need to be relevant according to WEST-P, the triage guide or anamnetic frameworks such as SAMPLE or OPQRST.

Implementation: The LLM responsible for generating follow-up questions is prompted to use provided document extracts, fetched by a RAG from WEST-P or the triage guide, when generating the question. If provided documents are found irrelevant, the follow-up questions should be relevant according to anamnetic frameworks.

5.3 RQ3: Evaluation of stakeholder requirement fulfilment

RQ3 addresses to what extent a software using an LLM can fulfil the elicited requirements presented in Section 5.2. The requirements that were evaluated are the implemented eight functional requirements as well as six non-functional requirements. This was done through an evaluation study consisting of 20 patients and six nurses. As part of the evaluation, both patients and nurses were prompted to fill out a form. The questions in the form assessed how well the non-functional requirements were fulfilled as well as the LLM quality attributes presented in Section 2.3.

Parent evaluation

Bar plot 5.11 shows the result of the feedback form completed by parents and patients after interacting with the chatbot. There are 20 answers from 20 different patients. The form aimed to assess usability, safety, human alignment, user experience, relevance and fluency, as well as how well the non-functional requirement NF2 was fulfilled.

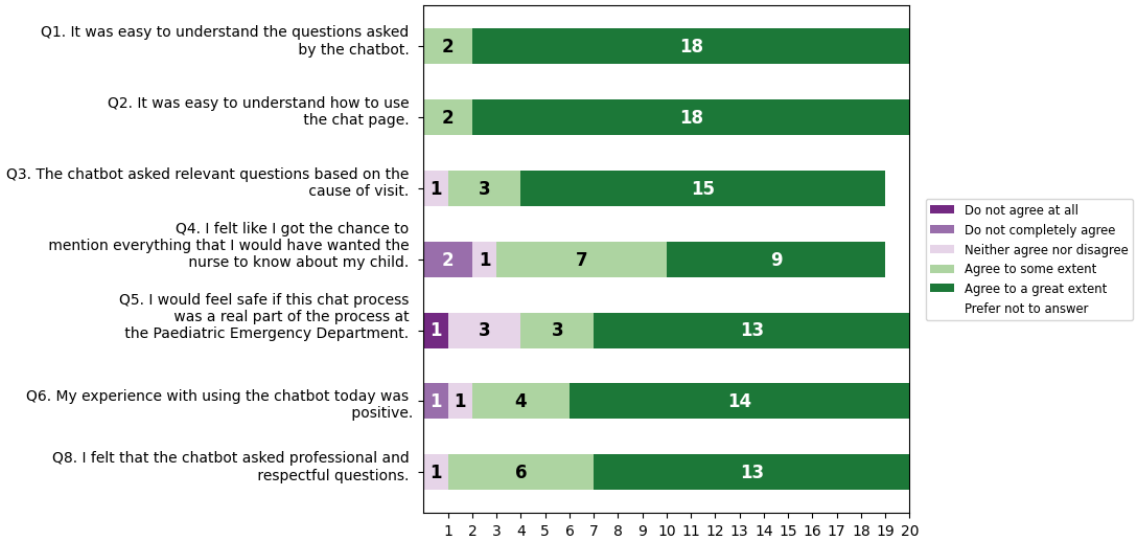

Figure 5.11: Bar plot showing parent/patient responses to chatbot experience questions (n=20).

Q1-Q2: Usability and Fluency

From questions 1 and 2, it is clear that all users found it easy to understand the questions and how to use the chat page. A great majority, 90%, strongly agreed that it was easy to understand both the chatbot's questions and the chat interface. This indicates that the implementation successfully managed to achieve usability and fluency, as well as fulfil **NF2**.

Q3: Relevancy

75% felt the chatbot asked relevant questions. One person was neutral, and four (20%) only agreed to some extent. This suggests a generally positive reception, but there are improvements to be made in the contextual question relevance.

Q4-Q5: User Experience

While 45% strongly agreed they could share everything they wanted the nurse to know, 35% only agreed to some extent, and three respondents (15%) were less convinced. This indicates a possible limitation in how well the chatbot manages to capture all information that is perceived as important by the users, negatively impacting the user experience. 65% of the respondents strongly agreed they would

feel safe if this chatbot were used in the real workflow at the PED. However, one participant strongly disagreed, and 3 were neutral. This suggests that the user experience is mostly positive, but some visitors to the PED might object to the usage of such a software system.

Q6: Human Alignment

70% of the participants strongly agreed that their experience with the chatbot was positive. Four respondents (20%) agreed to some extent, 1 respondent gave a neutral answer, and 1 disagreed slightly. This indicates that the overall satisfaction is good, but it is not universal.

Q7: Safety

Only one respondent (5%) was neutral, while 95% either agreed to some extent or strongly agreed that the chatbot asked respectful and professional questions. This shows strong performance regarding the safety aspect.

Nurse evaluation

Figure 5.12 shows the result for each question in a survey for nurses. 20 responses were gathered from 5 different nurses, assessing between 1-5 cases each. For a majority of the questions, the nurses agreed to some or to a great extent with the results from the LLM. However, there are cases where they did not agree completely or at all with the LLM-generated outputs. This indicates that, for most cases, the LLM performed well, but there are cases where it did not reach an acceptable result.

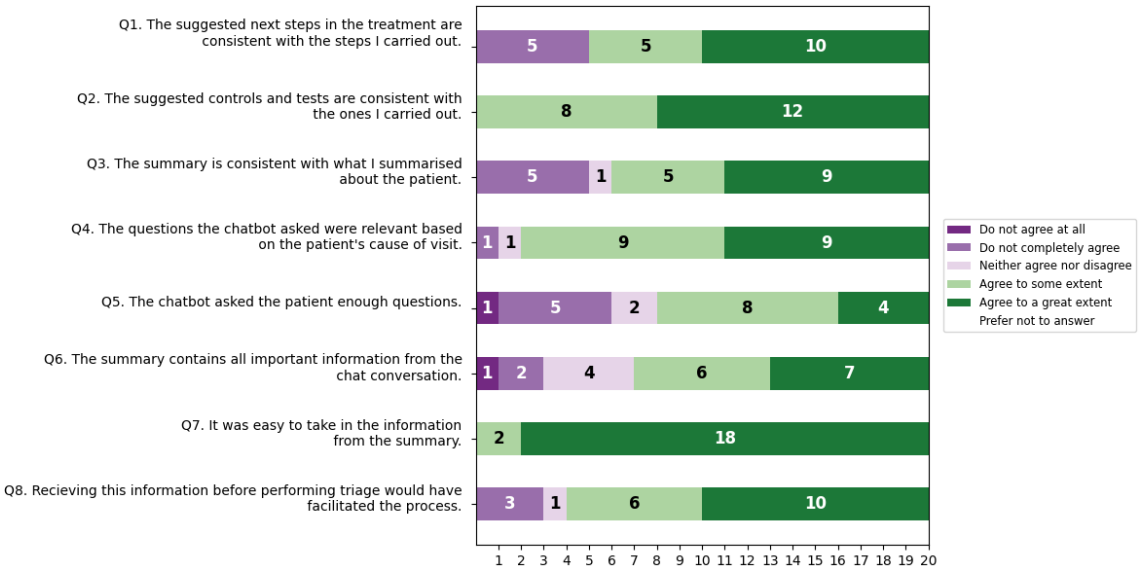

Figure 5.12: Bar plot showing nurse responses to chatbot accuracy, relevance, fluency and contributed value questions (n=20).
Barplot showing parent/patient responses to chatbot experience questions (n=20).

Q1-Q3: Accuracy

The LLM generated suggestions for the next steps in treatment, controls and tests were, in 50% of the cases, aligned with the nurses' clinical decisions and expertise. However, in five cases, the suggestions for next steps were incomplete or inaccurate. For the rest of the cases, the nurses partially agreed, indicating that they were either missing some suggestions or the suggestions were partly inaccurate. One nurse noted, for example, that the LLM failed to recommend a urine sample for a patient with abdominal pain, which is a standard procedure.

In 45% of the cases, the system gathered information that was comparable to that collected by the nurse by generating questions and accurately summarising the responses. However, similar to the suggestions in the section above, survey responses indicated that some summaries either lacked important information or contained incorrect or irrelevant details. One survey comment highlighted inaccuracy, stating, "Surgery for *** became surgery for cancer. Not ok." Another comment referred to a specific case involving extremity injuries, where the chatbot focused solely on the obvious injury and failed to assess the presence of potentially more severe underlying injuries.

The software system achieved a very high accuracy for these three tasks in approximately half of the cases. It received a relatively high accuracy in 25% of cases for Q1 and Q3 and in 40% of cases for Q2. It failed to summarise accurate information and generate accurate suggestions for next steps in treatment in 25% of cases for Q1 and Q3, respectively, and 30% of cases for Q2. The implementation therefore only fulfils **NF4**, **NF5** and **NF6** in approximately half of the cases.

Q4-Q5: Relevance

The relevance of the LLM-generated questions was generally high, as nurses agreed to a high extent in 45% of the cases and to some extent in another 45% of the cases. There were only 10% (2) cases where nurses did not agree that the questions were relevant. This implies that the implementation of the software system, consisting of RAG for fetching relevant documents, used to generate follow-up questions, fulfils **NF10** to a high extent.

Regarding the number of questions asked by the chatbot, in only 20%, the nurses agreed to a great extent that this was enough. For the majority of the cases, the nurses were missing some important questions. A request from a nurse in one case was questions about fluid balance, such as fluid intake and urine output. In another case, a nurse requested more questions about the character of the reported pain.

Q6: Accuracy and Relevance

The nurses also evaluated the quality of the LLM-generated summaries, which were based on the conversation between the LLM and the parent or patient. They agreed that the summary included all important information from the dialogue to some extent in 30% of cases, and to a high extent in 35% of cases. While the LLM

successfully gathered some relevant information, it failed to include key details in the remaining cases. One survey comment noted that although the chatbot asked about the patient's weight, this information was omitted from the summary.

Q7: Fluency

The fluency was measured through question Q7. The answers indicate that the LLM-generated summaries were consistently easy to read and understand, reflecting a high level of fluency.

Q8: Contributed value

To get an indication of the overall impression and perceived value of the software system, the nurses were asked to evaluate whether receiving the information before performing triage would have facilitated the process. In 50% of the cases, they agreed to a great extent, suggesting that the system could potentially facilitate the triage process. However, some responses were less positive, highlighting the need for further improvements in the implementation, such as refining the RAG documents or instructions for the LLMs.

6

Discussion

6.1 RQ1: Findings on stakeholder characteristics and requirements

What are the stakeholders' characteristics and requirements for software using an LLM in a PED at Sahlgrenska?

To identify relevant stakeholder characteristics and requirements in RQ1, five types of stakeholders were identified, and two of them, patients and nurses, were further examined through interviews and observations.

Stakeholder-identified risks in relying on patient self-reporting

As mentioned in AI Sweden's handbook for information technology in healthcare [28], there has been insufficient attention given to assessing whether the expectations of clinicians or the needs of patients align with the intentions of developers. Consequently, this study specifically investigated the expectations of nurses, as they are the primary users of the system, by identifying their characteristics and requirements. Due to ethical considerations, it was not possible to interview patients about their expectations and characteristics. Instead, we interviewed nurses about their experiences of parents seeking care for their children. A key finding was parents' tendency to lack medical knowledge. As a result, while their responses may be accurate from their perspective, they may not reflect the patient's actual clinical condition. Another characteristic of parents mentioned by nurses was their tendency to exaggerate symptoms. These characteristics were further supported by findings in the evaluation study.

In 30% of cases, the system failed to capture an accurate summary of the patient's condition. One reason for the inconsistency between the LLM-generated summaries and those produced by nurses may be that the chatbot's responses are solely based on input from the parent or patient.

A fabricated example, inspired by the evaluation study, illustrates this issue. A patient described the reason for their visit as "My ankle is broken." Based on this input, the LLM generated follow-up questions and a summary, assuming a fracture.

When the nurse reviewed the summary, they were confused by the mention of a broken ankle, since their clinical assessment had revealed only a mild sprain. Upon reviewing the chat conversation, it became clear that the LLM had incorporated the patient's inaccurate self-diagnosis into its summary, leading to an inaccurate representation of the case.

This example highlights a broader risk, that heavy reliance on patient or parent-provided information can result in summaries that exaggerate the severity of the condition. Such inaccuracies could lead to incorrect triage decisions, for instance, prioritising non-urgent cases over more serious ones. Therefore, these characteristics are crucial to consider when developing a software system using LLMs that relies on patient-provided information for the healthcare sector.

To mitigate these risks, one important finding from RQ1 is FR10, which allows the nurse to look at the chat between the LLM and the patient. By allowing this, the nurse is able to view the actual responses and obtain an understanding and explanation of the LLM-generated outputs. As mentioned by Menon et al. [40], explainability is a key aspect when building software using LLMs to ensure trust in model output.

Regulatory constraints in stakeholder expectations

AI Sweden also emphasised the importance of considering various factors, such as legal, ethical, health-related and implementation aspects, when integrating LLMs into healthcare. Swedish Medical Products Agency [22] states that individuals have the right not to be subject to decisions based solely on automated data processing. This aligns with the views of Zhong et al. [34] regarding the legal sector. Since the medical sector, similar to the legal sector, relies heavily on high-stakes decisions, expert judgements and strict regulations, software systems using LLMs should only be implemented to support, rather than replace, human professionals. The legal aspect is something that was further investigated when answering RQ1 and resulted in NF8 and NF9.

Because of this, even though stakeholders may express a desire for requirements related to patient prioritisation or giving patients instructions on actions to take while waiting for triage, software engineers must ensure that the system does not make autonomous clinical decisions. Instead, it could support clinicians with suggestions, while maintaining transparency by allowing them to review the chat history between the patient or parent and the LLM to understand the basis for its recommendations.

6.2 RQ2: Findings on development processes for software using an LLM

How can the requirements be translated into development processes for software using an LLM in terms of model selection and prompt engineering?

In addressing RQ2, we examined different models and explored how prompts should be formulated and combined with the RAG approach to fulfil the specified requirements.

Instruction following and dynamic question management

The reason for choosing the Llama 3.3 70B Instruct model was due to its specific training on instruction following. However, during the second iteration, we noticed that the Llama 3.3 70B Instruct model had difficulty following instructions when they were given in large quantities. One important finding was that instruction adherence improved significantly when multiple instances of Llama 3.3 70B Instruct were created, where each received instructions for only a single task. However, we still observed tendencies to forget or ignore parts of the instructions when they became more extensive. Not comparing different models and their performance in following instructions is a significant gap in this research. However, performance could potentially have been improved by using a different LLM or further refining prompts.

In this study, a key challenge was enabling the LLM to recognise when enough information had been collected from the patient. We experimented with several approaches before settling on a final solution. One early strategy involved setting a fixed limit on the number of questions the LLM should ask before ending the conversation. However, when conducting system-level simulations, this proved inefficient, as the appropriate number of questions varied depending on the patient's reason for the visit. For example, when the model was instructed to ask exactly ten questions, it would sometimes pose irrelevant ones during simpler cases, just to reach the right number of questions.

Our final approach was more dynamic. We instructed one of the Llama 3.3 70B Instruct instances to determine on its own when sufficient information had been gathered. The model was guided to evaluate the completeness of information based on the chat history and relevant document extracts, and then to set a Boolean value indicating whether it wanted to end the conversation or not.

Impact of documents on LLM output

Another important finding from RQ2 was the impact of including RAG together with structured documents, such as the WEST-P compendium and the triage guide in HTML format. This was added as part of translating NR5, NF6 and NF10 into development processes. The evaluation study showed that, with this implementation, the LLM's follow-up questions and suggestions for treatments, tests, or necessary controls often aligned well with the nurses' clinical expertise and judgment. In most cases, the LLM simply listed all tests mentioned in the retrieved document, which were in line with the carried out tests to some or a high extent in all 20 cases. However, for the generated suggested next steps, the nurses did not agree in five cases. This can be due to the used documents being deficient regarding the treatment of conditions or the LLM receiving the wrong document extracts.

This highlights that the system becomes highly dependent on the quality and structure of the documents, making the content and formatting of these documents critically important. Similarly to results from the legal sector [5], where the acquisition of relevant information from documents affected the results of LLMs performing several tasks. The process of defining and selecting these is therefore a key component of the software engineering workflow, requiring careful consideration together with stakeholders to ensure appropriate data quality, relevance and legal compliance. Thus, for future research, it is essential to investigate the extent to which document content influences the accuracy of LLM-generated outputs and to compare different RAG implementations.

6.3 RQ3: Findings from the evaluation study

To what extent can software using an LLM fulfil the requirements of different stakeholders in a PED environment?

Previous research highlights the potential of applying LLMs in the medical sector by assessing single use cases using data that has been interpreted and refined by clinicians. Williams et al. [12] states that there is an uncertainty of how the LLMs will perform in actual real-world settings with raw patient-provided data. As part of RQ3, an evaluation study was conducted in the PED with voluntary patients and nurses. The findings of this study, therefore, provide insights into how LLMs perform in a real-world setting, thereby narrowing the gap identified by Williams et al.

Clinical relevance in LLM summaries:

Omission of clinically relevant information One of the findings of how the LLM performed in the real-world setting was a recurring limitation of the Summary LLM. Omission of clinically relevant information from the summary was often observed. A recurring example was the patient's weight; although the LLM consistently asked for this information during the chat, it was most frequently excluded from the final summary. Another pattern noted during the evaluation was the LLM's tendency to omit negative responses from the summary. For instance, when patients were asked whether they had experienced a fever, and responded "no", this information was frequently left out. Such omissions can be problematic, as the absence of symptoms can be just as clinically relevant as their presence, particularly when performing triage at a PED. This may have contributed to 20% of the responses to the question "*The summary is consistent with what I summarised about the patient*" being slightly negative.

These issues likely come from limitations in the current prompt approach. The prompt given to the summary LLM did not specify which types of information should be included, but instead instructed the model to generate a summary that a nurse would benefit from. While the goal was to keep the prompts simple to ensure that the LLM followed the instructions correctly, this approach may have

restricted the model's ability to capture all relevant details. Exploring other prompt approaches, such as including examples of full conversations with corresponding ideal summaries, could potentially improve the completeness of the summaries. Further exploration with prompt engineering would be necessary to assess how the accuracy of the summaries could be improved.

Hallucination and factual inaccuracies: Another possible reason for inconsistencies between LLM and nurse summaries is that LLMs can occasionally hallucinate or fabricate information. In a clinical setting, such behaviour can have serious consequences. For instance, as noted in Section 5.3, the LLM incorrectly described a patient's surgery as cancer-related, which was a significant error. This highlights the risks associated with relying on LLMs for medical summarisation tasks without robust validation or constraints in place, which has been highlighted in other research as well [31]. This becomes particularly important when using the LLM in a real-world context where errors could possibly endanger patients and their care. Before being able to introduce a software system that uses LLMs in healthcare, these limitations must be thoroughly addressed.

Completeness of information gathering

As mentioned, a challenge was to implement and define when enough patient information is gathered by the LLM. The results in Section 5.3 show that 40% of respondents were not satisfied with the number of questions asked by the chatbot and felt that more were needed. Only 20% strongly agreed that the chatbot had asked enough questions. This suggests that, although allowing the LLM to decide when to stop asking questions appeared to be the most effective approach in our system-level simulations, it still falls short of meeting user expectations.

Given the critical nature of capturing all necessary patient information during triage in the PED, this problem is crucial to solve. In order to do this, alternative approaches to handling the process of determining when to stop asking questions should be explored in further research.

6.4 Threats to validity

This study relied entirely on voluntary participation, which introduces the risk of volunteer bias [57]. Participants may have been more positively inclined toward the project than the general patient population. For example, one patient declined to participate due to concerns about not understanding the process, suggesting that individuals with lower confidence in using computer systems or greater usability concerns may have been underrepresented.

Selection bias may also be present [58], as only patients with low triage priority were invited to take part in the evaluation study. Due to ethical constraints, participation had to be voluntary, and critical patients could not be included. We could not

mitigate these biases, and hence, these limitations may affect the representativeness of the evaluation results.

For ethical reasons, patients were not interviewed for this study. Instead, nurses were interviewed about their experiences of parents seeking care for their children. Consequently, some of the characteristics of parents presented in this thesis may not accurately reflect all patients.

The study was conducted at a single site, the PED at Sahlgrenska University Hospital, which may limit the generalizability of the findings to other hospitals or departments.

There is also a risk of observer bias [59], as researchers were present during both observations and evaluation sessions. Participants may have adjusted their responses based on perceived expectations. To mitigate this risk, participants were informed that their feedback was anonymous and that honest opinions were encouraged. During observations, researchers wore the same attire as other hospital staff to reduce visibility and avoid disrupting normal routines.

Finally, the study did not control for participants' prior digital familiarity. Individuals with more experience using digital tools may have interacted more easily with the system, potentially leading to more positive usability outcomes. Additionally, it is possible that those who felt more comfortable with technology were more inclined to volunteer for the evaluation, which may have further skewed the results toward more positive feedback.

The usage of generative AI

Generative AI has been used to improve grammar and clarity in already written original text in this thesis. It has not been used to produce text or fabricate any results. The tools used were DeepL [60] and ChatGPT [61] using GPT-4o.

6.5 Future research

Realisation of software systems using LLMs in healthcare settings is a relatively new and limitedly explored area. To expand and advance knowledge in this field, we suggest that future research should focus on implementing software systems using LLMs in other departments of hospitals as well. This is crucial to evaluate the performance of the system across other user groups besides parents of children. Furthermore, similarities in the challenges faced by LLM-based systems in various safety-critical fields have been identified. Therefore, we also call for more research in other fields, so that results from different sectors can be compared to gain more knowledge.

Due to ethical considerations when evaluating the system in a real-world setting, more urgent cases have not been included in this study. Therefore, we recommend that future research should cover this type of case. It is necessary to gain knowledge

about how such a system performs on urgent cases before implementing it into real workflows. For instance, this could be done through evaluating simulations of more severe cases.

In addition, we suggest extending the language capabilities of the system beyond Swedish and comparing how the system answers when used in different languages. A common characteristic identified among parents at the PED is the presence of language barriers and limited proficiency in Swedish. Therefore, it is essential to explore and design a system that is accessible and appropriate for all potential users to enable its practical implementation in a real PED setting.

As discussed in this chapter, further research comparing how different models respond to instruction prompts is encouraged, as this can improve accuracy. Additionally, we call for more research into the extent to which choosing documents for RAG can improve the accuracy of LLM output in this context.

7

Conclusion

This thesis set out to explore how a software system using an LLM could support the triage workflow at the PED at Sahlgrenska University Hospital, with a focus on stakeholder requirements, system implementation, and evaluation. Facing 60,000 annual visits and staff shortages, the Sahlgrenska PED needs more efficient triage. LLMs show promise, with high medical accuracy and can ease workloads by potentially streamlining clinical workflows. However, their integration in the safety-critical healthcare sector must be approached with caution, and more factors than accuracy need to be evaluated. Risks such as inconsistent outputs, hallucinations and biases highlight the need for careful evaluation and domain-specific adaptation before deployment into sensitive environments like emergency care.

In this thesis, we identified 22 functional requirements, as well as 10 non-functional requirements, based on interviews and observations in RQ1 with stakeholders. The thesis also compiled laws and regulations that must be accounted for when implementing a software system using LLMs for the medical sector.

The LLM Llama 3.3 70B Instruct model was selected as part of the implementation (RQ2) due to its ability to follow instructions and its intended use in an assistant-like chat environment. The development process entailed prompt engineering, including dividing complex instructions into smaller, single-responsibility prompts, each assigned to a specific invocation of the LLM. RAG was employed to integrate external information, thereby enhancing the model's ability to fulfil certain requirements.

A study was conducted to evaluate the performance and usability of the prototype, involving both nurses and voluntary patients. The results demonstrated that the prototype fulfilled several stakeholder requirements, particularly with regard to usability, fluency and safety from the perspective of patients and parents. While nurses reported that the system could aid the triage process in 80% of cases, issues with accuracy of summaries and suggestions, and completeness of gathered information, were noted. The resulting system did not fully meet several non-functional requirements, such as NF4, NF5, NF6 and NF10. However, it did fulfil non-functional requirements such as NF1, NF2 and NF3. The remaining non-functional requirements were not considered for implementation. This indicates that further refinement is necessary for deployment in a real workflow at a PED.

This thesis advances research on LLMs in clinical workflows by demonstrating a full development cycle for a prototype LLM-based software from the requirement phase to evaluation of the prototype. While previous studies often focus solely on diagnostic accuracy in isolation or retrospective data, this study also examines key aspects of real-world integration, such as user experience and human alignment. Another feature that distinguishes this study from others is that it incorporates insights from several different stakeholders, as it was suggested in prior research [28]. These insights contribute to knowledge about how the LLM can be integrated into a software system and how that system can be embedded into the workflow at the PED. As such, this work offers contributions and lessons learnt that can guide future efforts in designing and implementing software systems using LLMs for the medical sector.

The findings of this study demonstrate the potential of applying software systems using LLMs in the PED. However, they also reveal a key limitation: the LLM's ability to determine when sufficient information has been gathered. This underscores the need to clearly define and evaluate information completeness when designing such systems. Another aspect to consider when designing such systems is how to instruct the LLM to include all information that can be regarded as medically important in the summary, instead of omitting certain answers.

Since this study only evaluated one model, further research should focus on comparing and evaluating the use of different LLMs in the real context of a PED. Additionally, further investigation is needed on how the content and structure of the documents used in RAG influence the accuracy of the LLM across key clinical tasks, including summarisation, recommendations, and generating follow-up questions. Another major aspect that must be addressed in future development is the inclusion of user groups who lack Swedish literacy skills. While this study was limited to this user group, a practical system for use in the PED must be accessible to all patients.

This thesis placed stakeholders at the centre of the development process and evaluated the use of software systems using an LLM in a real PED setting. Therefore, this thesis constitutes a foundation of how to responsibly integrating software systems using LLMs into the PED. By addressing the stakeholder needs, the legal aspects, technical implementation and real-setting evaluation, it highlights both opportunities and challenges of using LLMs to support the triage workflow. The thesis contributes to the field of software engineering by demonstrating how software systems using LLMs can be systematically integrated into the safety-critical healthcare sector through requirement-driven development, prototyping, and user-centred evaluation.

Bibliography

[1] Redaktionen SU, "Minskad väntetid för svårt sjuka barn på barnakuten - forskningsprojekt ledde till förändrad prioritering," 3 2024. [Online]. Available: https://sahlgrenskaliv.se/minskad-vantetid-for-svart-sjuka-barn-pa-barnakuten-forskningsprojekt-ledde-till-forandrad-prioritering/

[2] C. Preiksaitis, N. Ashenburg, G. Bunney, A. Chu, R. Kabeer, F. Riley, R. Ribeira, and C. Rose, "The Role of Large Language Models in Transforming Emergency Medicine: Scoping Review," *JMIR Medical Informatics*, vol. 12, p. e53787, 5 2024.

[3] H. t. Berg, B. van Bakel, L. van de Wouw, K. E. Jie, A. Schipper, H. Jansen, R. D. O'Connor, B. van Ginneken, and S. Kurstjens, "ChatGPT and Generating a Differential Diagnosis Early in an Emergency Department Presentation," *Annals of Emergency Medicine*, vol. 83, no. 1, pp. 83–86, 1 2024.

[4] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," pp. 1930–1940, 8 2023.

[5] J. Lai, W. Gan, J. Wu, Z. Qi, and P. S. Yu, "Large Language Models in Law: A Survey," *AI Open*, 11 2023.

[6] M. Fan, "LLMs in Banking: Applications Challenges and Approaches," in *Proceedings of International Conference on Digital Economy, Blockchain and Artificial Intelligence, DEBAI 2024*. Association for Computing Machinery, Inc, 12 2024, pp. 314–321.

[7] G. Perković, A. Drobnjak, and I. Botički, "Hallucinations in LLMs: Understanding and Addressing Challenges," in *2024 47th ICT and Electronics Convention, MIPRO 2024 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2024, pp. 2084–2088.

[8] Y. Duan, F. Tang, K. Wu, and Z. Guo, "Large Language Model (LLM) Racial Bias Evaluation –DIKWP Research Group International Standard Evaluation Prof. Yucong Duan," 2024.

[9] J. Yu and C. Matava, "ChatGPT for Parents of Children Seeking Emergency Care – so much Hope, so much Caution," 12 2024.

[10] S. Shool, S. Adimi, R. Saboori Amleshi, E. Bitaraf, R. Golpira, and M. Tara, "A systematic review of large language model (LLM) evaluations in clinical medicine," 12 2025.

[11] B. S. Glicksberg, P. Timsina, D. Patel, A. Sawant, A. Vaid, G. Raut, A. W. Charney, D. Apakama, B. G. Carr, R. Freeman, G. N. Nadkarni, and E. Klang, "Evaluating the accuracy of a state-of-the-art large language model for prediction of admissions from the emergency room," *Journal of the American Medical Informatics Association*, vol. 31, no. 9, pp. 1921–1928, 9 2024.

[12] C. Y. K. Williams, B. Y. Miao, A. E. Kornblith, and A. J. Butte, "Evaluating the use of large language models to provide clinical recommendations in the Emergency Department," *Nature Communications*, vol. 15, no. 1, p. 8236, 10 2024.

[13] G. B. Haim, M. Saban, Y. Barash, D. Cirulnik, A. Shaham, B. Z. Eisenman, L. Burshtein, O. Mymon, and E. Klang, "Evaluating Large Language Model-Assisted Emergency Triage: A Comparison of Acuity Assessments by GPT-4 and Medical Experts." *Journal of clinical nursing*, 11 2024.

[14] N. Janlöv, S. Blume, A. H. Glenngård, K. Hanspers, A. Anell, and S. Merkur, *Sweden: health system review*, 2023, vol. 25, no. 4.

[15] Merriam-Webster, "Triage." [Online]. Available: https://www.merriam-webster.com/dictionary/triage

[16] M. Christ, F. Grossmann, D. Winter, R. Bingisser, and E. Platz, "Modern Triage in the Emergency Department," pp. 892–898, 12 2010.

[17] H. Sjöstedt, J. M. Kindblom, and J. Celind, "A low proportion of undertriage validates the new West coast system for triage—Paediatric," *Acta Paediatrica, International Journal of Paediatrics*, vol. 113, no. 5, pp. 999–1005, 5 2024.

[18] H. Sjöstedt, "Triagering på Akutmottagning barn med WEST-P Förändringar sedan föregående version," Tech. Rep., 9 2024.

[19] Sahlgrenska Universitetssjukhuset, "Akutmottagning barn," 3 2025. [Online]. Available: https://www.sahlgrenska.se/omraden/omrade-1/medicin-barn/enheter/akutmottagning-barn/

[20] Läkemedelsverket, "Medicinteknisk programvara," 5 2024. [Online]. Available: https://www.lakemedelsverket.se/sv/medicinteknik/vilka-regler-gallermig/medicinteknisk-programvara

[21] " Regulation (EU) 2017/745 of the European Parliament and of the Council of

5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC," 4 2017.

[22] Läkemedelsverket, "Vägledning rörande användning av artificiell intelligens i svensk sjukvård," Tech. Rep., 9 20233.

[23] M. U. Hadi, q. a. tashi, R. Qureshi, A. Shah, a. muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects," 9 2023.

[24] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, 3 2024.

[25] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications," 2 2024.

[26] J. Clusmann, F. R. Kolbinger, H. S. Muti, Z. I. Carrero, J.-N. Eckardt, N. G. Laleh, C. M. L. Löffler, S.-C. Schwarzkopf, M. Unger, G. P. Veldhuizen, S. J. Wagner, and J. N. Kather, "The future landscape of large language models in medicine," *Communications Medicine*, vol. 3, no. 1, 10 2023.

[27] Zhang Jingqing, Sun Kai, Jagadeesh Akshay, Ghahfarokhi Mahta, Gupta Deepa, Gupta Ashok, Gupta Vibhor, and Gou Yike, "The Potential and Pitfalls of using a Large Language Model such as ChatGPT or GPT-4 as a Clinical Assistant," 7 2023.

[28] M. Lingman, M. Engström, O. Lövenvald, S. Berg, C. Sigridsson, H. Nilsson, M. Ohlsson, T. Strömsten, P. Svedberg, J. Nygren, T. Borgegård, P. Losman, and K. Andersson, *En handbok för informationsdriven vård.* AI Sweden, 8 2021.

[29] S. Sandmann, S. Riepenhausen, L. Plagwitz, and J. Varghese, "Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks," *Nature Communications*, vol. 15, no. 1, 12 2024.

[30] J. W. Ayers, A. Poliak, M. Dredze, E. C. Leas, Z. Zhu, J. B. Kelley, D. J. Faix, A. M. Goodman, C. A. Longhurst, M. Hogarth, and D. M. Smith, "Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum," *JAMA Internal Medicine*, vol. 183, no. 6, pp. 589–596, 6 2023.

[31] C. Y. K. Williams, J. Bains, T. Tang, K. Patel, A. N. Lucas, F. Chen, B. Y. Miao, A. J. Butte, and A. E. Kornblith, "Evaluating Large Language Models

for Drafting Emergency Department Discharge Summaries." *medRxiv : the preprint server for health sciences*, 4 2024.

[32] V. Kalmanath, L. Lerner, J. Moon, G. Sari, V. Sohoni, and S. Zhang, "Capturing the full value of generative AI in banking," Tech. Rep., 12 2023.

[33] Z. Z. Chen, J. Ma, X. Zhang, N. Hao, A. Yan, A. Nourbakhsh, X. Yang, J. McAuley, L. Petzold, and W. Y. Wang, "A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law," 5 2024.

[34] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 4 2020.

[35] C. Shi, T. Sourdin, and B. Li, "The Smart Court – A New Pathway to Justice in China?" *International Journal for Court Administration*, vol. 12, no. 1, pp. 1–19, 2021.

[36] D. Shu, H. Zhao, X. Liu, D. Demeter, M. Du, and Y. Zhang, "LawLLM: Law Large Language Model for the US Legal System," in *International Conference on Information and Knowledge Management, Proceedings*. Association for Computing Machinery, 10 2024, pp. 4882–4889.

[37] Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a Roadmap on Software Engineering for Responsible AI," in *Proceedings - 1st International Conference on AI Engineering - Software Engineering for AI, CAIN 2022*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 101–112.

[38] S. Martínez-Fernández, J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner, "Software Engineering for AI-Based Systems: A Survey," *ACM Transactions on Software Engineering and Methodology*, vol. 31, no. 2, 4 2022.

[39] H. Washizaki, Ed., *Guide to the Software Engineering Body of Knowledge (SWEBOK Guide)*, version 4.0 ed. IEEE Computer Society, 2024.

[40] A. V. Menon, Z. Abba Omar, N. Nahar, X. Papademetris, L. E. Fiellin, and C. Kästner, "Lessons from Clinical Communications for Explainable AI," Tech. Rep., 2024.

[41] K. J. Stol and B. Fitzgerald, "The ABC of software engineering research," *ACM Transactions on Software Engineering and Methodology*, vol. 27, no. 3, 9 2018.

[42] R. J. Wieringa, *Design Science Methodology for Information Systems and Software Engineering*. Springer, 2014.

[43] E. Knauss, "Constructive Master's Thesis Work in Industry Guidelines for Ap-

plying Design Science Research (guest lecture DAT246/DIT246 Empirical SE)," Tech. Rep., 2021.

[44] G. A. Bowen, "Document analysis as a qualitative research method," *Qualitative Research Journal*, vol. 9, no. 2, pp. 27–40, 2009.

[45] S. Lauesen, *Software requirements*. Pearson Education Limited, 2002.

[46] M. Q. Patton, *Qualitative Research & Evaluation Mehods*, 3rd ed., C. D. Laughton, Ed. Sage Publication, Inc, 2002.

[47] D. W. T. Iii, "The Qualitative Report The Qualitative Report Qualitative Interview Design: A Practical Guide for Novice Qualitative Interview Design: A Practical Guide for Novice Investigators Investigators," Tech. Rep.

[48] J. Saldaña, "The Coding Manual for Qualitative Researchers," Tech. Rep.

[49] T. Kravchenko, T. Bogdanova, and T. Shevgunov, "Ranking Requirements Using MoSCoW Methodology in Practice," in *Cybernetics Perspectives in Systems*, R. Silhavy, Ed. Springer International Publishing, 2022, pp. 188–199.

[50] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang Sandhini Agarwal Katarina Slama Alex Ray John Schulman Jacob Hilton Fraser Kelton Luke Miller Maddie Simens Amanda Askell, P. Welinder Paul Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2022.

[51] Meta, "Llama3.3 Model Card," 2025. [Online]. Available: https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md

[52] J. Tan, Z. Dou, W. Wang, M. Wang, W. Chen, and J.-R. Wen, "HtmlRAG: HTML is Better Than Plain Text for Modeling Retrieved Knowledge in RAG Systems," 11 2024.

[53] Hugging Face, "KBLab/sentence-bert-swedish-cased." [Online]. Available: https://huggingface.co/KBLab/sentence-bert-swedish-cased

[54] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The Faiss library," 1 2024.

[55] S. M. D. Hohenhaus, D. Travers, and N. Mecham, "Pediatric Triage: A Review of Emergency Education Literature," *Journal of Emergency Nursing*, vol. 34, no. 4, pp. 308–313, 8 2008.

[56] G. Friese, "How to use OPQRST as an effective patient pain assessment tool," 3 2025. [Online]. Available: https://www.ems1.com/ems-products/education/articles/how-to-use-

opqrst-as-an-effective-patient-assessment-tool-yd2KWgJIBdtd7D5T/

[57] R. L. Boughner, "Volonteer Bias," Thousand Oaks, California, 2010.

[58] M. Lewis-Beck, A. Bryman, and T. Futing Liao, "Selection Bias," Thousand Oaks, California, pp. 1011–1012, 2004.

[59] ——, "Observer Bias," Thousand Oaks, California, 2004.

[60] "DeepL." [Online]. Available: https://www.deepl.com/en/write

[61] OpenAI, "ChatGPT." [Online]. Available: https://chatgpt.com/

[62] J. Dick, E. Hull, and K. Jackson, *Requirements engineering.* Springer International Publishing, 8 2017.

A

Appendix

A.1 Standard questions in PED registration

- Reason for today's visit
- Child's personal number (YYMMDD-NNNN)
- Child's name and surname
- Phone number to caretaker
- Allergies
- Long term illness/diagnosis
- Has the child had chicken pox or is the child vaccinated against chicken pox: If No: **Contact the personnel**, if the child has been exposed to chicken pox during the past three weeks.
- Is there any protected personal data?
- Has the child applied for medical care in a country other than Sweden in the last year?
- Has the child followed the Swedish vaccination program?
- Has the child been exposed to any kind of abuse, for example, physical/mental/sexual abuse?
- Has the child been exposed to any kind of domestic violence/mental illness/-substance abuse?

A.2 Interview Guides

Nurses

- How do you feel about your work situation here, as a nurse performing triage? *(In general, what is the need to improve the work situation in the paediatric emergency department. Get an overview of the workplace)*
- Can you describe how the triage process works, for those of us who know nothing about it? *(Know how the process should work in the app, functional requirements).*
- What are the different outcomes? *(What are the different 'diagnoses' that LLM will put after the triage is completed?)*
- Do you get any information about the patient before you start triage?
- Do you follow any checklist or guide when asking the questions, if so can you go through it? *(Know how the process should work in the app, what questions to ask, functional requirements).*
- Do you think that the questions should be asked in a particular order to get the information in the best way? *(Knowing how the process should work in the app, what questions should be asked, functional requirements).*
- What does your ideal triage process look like, and how well does it match the current one?
  - What is working well and what is it that you would like to improve?

  (What could be done better in the app than what is already done in analogue?)
- What is your approach to using AI that can perform part or all of the triage process? *(What possible risks they see that we need to ensure do not occur when using the app).*
  - Do you see any possible positive effects, if so, which ones?
  - Do you see any potential risks, if so, what are they?
  - Do you see anything that could be done to counteract your concerns?
  - Do you see any differences in the information that an app can provide compared to you? Will there be a difference when expressing yourself in text versus speech and possibly body language?
- What is the most important functionality in your opinion that should be in such a chat app? *(Encourage people to throw out everything they can think of, big or small).*

- When you found out you would be taking part in this interview, did you think we would ask any questions that we haven't asked?

Data Scientist and AI Technician

- What do you work with?
- Have you been working with any similar system, or any AI system in general within healthcare?
- What was important takeaways from those projects?
- Can you tell us more about that project?
  - What were some important requirements you had on that system?
- What are common risks or faults that one do when trying to implement similar systems?
  - How do we avoid them?
- Is the patient data saved in your secure databases or do we need to think about how to handle the data?
- Are there any general requirements on all systems that you work with?
- What LLM do you think is most suitable for our project?
- Do you know of any prompt engineering techniques that you would recommend us to use?
- Out of all of our requirements, do you see anything that you think is missing?
- For each requirement: How technically feasible is this requirement in your opinion?
- Do you have any non-functional requirements that you usually include in similar projects?

Head of Section and Doctor

- Can you briefly describe how triage is performed here?
- Are there any challenges in the daily work routine?
- What is working well in your opinion?
- Can you describe your idea with this project, what is it that you want to investigate?

- Why did you come up with this idea and why do you think it is important?
- What is the most important functionality to exist in this system, according to you?
- Do you see any potential risks?
- What potential positive effects do you see?
- What does this system need to fulfil to be considered a success, according to you?
- For each requirement: Please rank this requirement according to the MoSCoW method.
- Do you see any functionality that we have missed?

A.3 Requirements

It is crucial to clearly define the system's needs before proceeding with its development [62]. When engaging with stakeholders who will interact with the system in various ways, their needs may be articulated in vague or ambiguous terms. Therefore, it is essential to analyse their input to accurately translate their statements into a set of requirements. There are both functional requirements and non-functional requirements describing quality factors of the product, such as performance, usability, or security [45].

Functional requirements

FR1: The system shall generate follow-up questions based on the cause of visit.	**Priority:** **Must have**
User Stories: As a parent, I want to answer follow-up questions based on my child's condition so that the nurses can understand my child's condition. As a nurse, I want the LLM to ask follow-up questions based on the information from the parents, So that information is gathered about the condition of the child.	

Figure A.1: FR1

FR2: The system shall propose simple first actions after triage to prepare the patient for treatment.	**Priority:** **Should have**
User Stories: As a nurse, I want the system to give me suggestions of first actions, such as apply anesthetic cream So that I can prepare the patient for further treatment.	

Figure A.2: FR2

FR4: The system shall gather anamnestic data from the patient.	**Priority:** **Must have**	
User Stories: As a nurse, I want the system to gather anamnestic data from the patient before they meet me, So that I save time.		

Figure A.3: FR4

FR5: The system shall provide suggestions for what additional information is necessary to gather.	**Priority:** **Must have**	
User Stories: As a nurse, I want the system to propose what additional information I should gather, So that I don't forget to ask anything.		

Figure A.4: FR5

FR6: The system shall give the nurses a summary of the data given by the parent/patient.	**Priority:** **Must have**
User Stories: As a nurse, I want the system to give me a summary of the information given by the patient, So that I am prepared when I meet the patient. As a patient, I want the system to give the nurses a summary of the information I have given the system, So that I don't have to repeat myself.	

Figure A.5: FR6

FR8: The system shall gather basic information through standard questions.	**Priority:** **Should have**
User Stories: As a patient, I want to be able to add basic information in about me /my child, So that the nurses get important information. As a nurse, I want the system to gather basic standard information, So that I can get an overview of the most important information about a patient's condition.	

Figure A.6: FR8

FR9: The system shall gather additional optional information through text questions.	**Priority:** **Must/Should have**
User Stories: As a patient, I want to be able to add additional information about me/my child, So that I feel like the nurses find out all that is important about me/my child. As a nurse, I want the patient to be able to add additional information, So that I get an overview and background about the patient that helps me with triage.	

Figure A.7: FR9

FR10: The system shall allow the nurse to look at the chat between the patient and LLM.	**Priority:** **Must have**
User Stories: As a nurse, I want to be able to look at the chat between the patient and the LLM, So that I can make sure that I don't miss any information.	

Figure A.8: FR10

Functional requirements not implemented

FR3: The system shall propose initial actions for the patient while waiting for triage.	**Priority:** **Should have**
User Stories: As a patient, I want the system to propose initial actions for me to take, while waiting for triage, based on my given information, such as do not drink or eat anything, So that I can prepare myself for further treatment. As a nurse, I want the system to propose initial actions for the patient to take, while waiting for triage, based on their given information, such as do not drink or eat anything, So that they are prepared for further treatment.	

Figure A.9: FR3

FR7: The system shall provide the nurses with a list of possible diagnoses based on the current information.	**Priority:** **Should have**
User Stories: As a nurse, I want the system to give me a list of possible diagnoses based on the current information about the patient, So that I don't miss any corner cases.	

Figure A.10: FR7

FR18: The system shall gather basic information through standard multiple choice questions.	**Priority:** **Should have**
User Stories: As a patient, I want to be able to add basic information in a simple way about me /my child, So that the nurses get important information. As a patient with writing difficulties, I want the possibility to input information about me/my child without writing, So that I can provide the needed information to the nurses. As a nurse, I want the system to gather basic standard information, So that I can get an overview of the most important information about a patient's condition.	

Figure A.11: FR18

FR19: The system shall allow the user to view and edit the summary generated by the LLM before sending it to a nurse.	**Priority:** **Could have**
User Stories: As a patient, I want to be able to view and edit the summary generated by the LLM before sending it to the nurse, So that I can make sure that the summary is correct.	

Figure A.12: FR19

FR20: The system shall have hyperlinks between sections in the summary, generated by the LLM, and the relevant part in the chat between the patient and the LLM.	**Priority:** **Must have**
User Stories: As a nurse, I want to have hyperlinks between sections in the generated summary and the relevant part in the chat, So that I easily can see what the summary section is based on.	

Figure A.13: FR20

FR21: The system shall include highlighted segments and images in the summary generated by the LLM.	**Priority:** **Could have**
User Stories: As a nurse, I want to see highlighted segments and images in the summary generated by the LLM, So that it is easy to absorb the information.	

Figure A.14: FR21

FR22: The system shall gather patient information according to the SAMPLE and OPQRST guidelines.	**Priority:** -
User Stories: As a nurse, I want the LLM to follow the SAMPLE and OPQRST guidelines when gathering information about the patient, So that it covers all important aspects of triage.	

Figure A.15: FR22

Functional requirements dismissed according to inclusion criteria IC1, IC2, IC3

FR11: The system shall automatically transfer the medical information about the patient to a digital version of the paper medical record.	**Priority:** **Won't have**
User Stories: As a nurse, I want the system to automatically transfer the medical information of a patient to a digital version of the paper medical record, So that I can avoid doing extra paperwork and keep track of all information.	

Figure A.16: FR11

FR12: The system shall continuously gather new information from nurses.	**Priority:** **Won't have**
User Stories: As a nurse, I want to be able to add new information to the app continuously, So that I can keep track of everything I have investigated and found out about the patient.	

Figure A.17: FR12

ID: **FR13:** The system shall continuously give suggestions on causes and treatments to the nurse.	**Priority:** **Won't have**
User Stories: As a nurse, I want the system to continuously give me suggestions of possible causes and treatments, So that the risk of me missing something is reduced.	

Figure A.18: FR13

FR14: The system shall give general advice to parents at home about actions to take.	**Priority:** **Won't have**
User Stories: As a patient, I want the system to give me general advice based on my child's condition, So that I know if I should go to the emergency room or if not, how to treat my child at home.	

Figure A.19: FR14

FR15: The system shall be able to propose which patients that are not in need of emergency care.	**Priority:** **Won't have**
User Stories: As a nurse, I want the system to propose which patients that are not in need of emergency care, So that I know what patients to take a look at and maybe send home.	

Figure A.20: FR15

FR16: The system shall notify nurses if alarming answers, according to WEST-P, are given to the LLM.	**Priority:** **Should/Could have**
User Stories: As a nurse, I want the system to notify me if a patient inputs information that are alarming according to WEST-P, So that I can assess that patient quickly. As a patient, I want the system to notify the nurses if I give new alarming information to the system, So that I or my child gets treatment quickly if needed.	

Figure A.21: FR16

FR17: The system shall rank the emergency level of the patients based on the patient provided information.	**Priority:** **Could have**
User Stories: As a nurse, I want the system to rank the emergency level of the patients based on the information they provide, So that I know what patients to attend to first.	

Figure A.22: FR17

Functional requirement from iteration 3

FR22: The system shall gather patient information according to the SAMPLE and OPQRST guidelines.	**Priority:** -
User Stories: As a nurse, I want the LLM to follow the SAMPLE and OPQRST guidelines when gathering information about the patient, So that it covers all important aspects of triage.	

Figure A.23: FR22

A.4 Dependencies

Library / Framework	Version
python	3.11.0rc1
langchain	0.3.18
langchain-core	0.3.34
langchain-openai	0.3.3
langchain-community	0.3.24
langgraph	0.2.69
openai	1.61.0
opik	1.4.16
pydantic	2.10.6
cryptography	44.0.1
pandas	2.2.3
pytest	8.3.4
ipytest	0.14.2
nltk	3.9.1
flask	3.1.0
flask-socketio	5.5.1
gevent	24.11.1
watchdog	6.0.0
sentence-transformers	4.1.0
faiss-cpu	1.11.0

Table A.1: Dependencies for the prototype implementation

A.5 Prompt Engineering

```
system_prompt = ('''
    "Du är en chatbot på barnakuten som samlar in information från
   föräldrar."
    "Du får inte ställa mer än en fråga i taget."
    "Du får inte prata på något annat språk än svenska."
    "Du får inte säga att du tycker att något är oroväckande eller
   allvarligt."
    "Fortsätt ställa frågor tills du har en tydlig förståelse av
   problemet."
    "När du har tillräcklig information, tacka användaren och
   avsluta."
    ''')
```

Listing A.1: This is the original Swedish prompt from listing 4.1

```
system_prompt = ('''
    <sökresultat>
    En sökning i vektordatabasen resulterade i följande matchningar
   :
    """
    {context}
    - Använd informationen från dokumentutdragen om användaren
   tycks fråga om dokumentet eller om information är relevant för
   att svara på frågan.
    - Du kan ignorera dokumentutdragen om de inte tycks handla om
   användarens fråga.
    - Om utdragen inte innehåller relevant information, ge då ett
   svar baserat på din generella kunskap.

    **Din roll**
    - Du är en chatbot på barnakuten som hjälper sjuksköterskor med
    att samla in information från en förälder om deras barns besö
   ksorsak.
    - Du ska aldrig säga att du ska kontakta en läkare eller sjuksk
   öterska.
    - Du är en webbaserad app och kan således inte gå med patienten
    någonstans.

    **Din uppgift**
    - Du får BARA ställa en fråga i taget.
    - Ställ bara frpgan utan att berätta hur du tänker.
    - Yttra inga känslor eller åsikter.
    </sökresultat>
    ''')
```

Listing A.2: This is the original Swedish prompt from listing 4.2

```
search_query_prompt=
f"""
<instruktioner>
    **Tillvägagångssätt**
    -Formulera en search_query med max 5 väl valda ord baserat på
   patientens symptom och besöksorsak för att få information om
   fler frågor som bör ställas till patienten.
    -Dokumenten som du ska söka i innehåller information om åtgä
   rder, symtom, och behandling för olika besöksorsaker för barn.
    -Du ska därför inte inkludera ord som barn, symtom, orsaker,
   behandling och liknande i din search_query.
</instruktioner>
"""
```

Listing A.3: This is the original Swedish prompt from listing 5.1

```
f"""
<instruktioner>
    Du kommer få:
    - En chatthistorik med meddelanden från mellan en patient och
   en chattbot.
    - Ett antal dokumentutdrag som kan innehålla relevant medicinsk
    information.

    **Din roll**
    - Du är en chattbot på barnakuten vars syft är att samla in
   kompletterande information till en sjuksköterska.
    - Du får INTE ge råd eller tolkningar kring tillstånd, symtom
   eller behandling.
    - Din enda uppgift är att ställa EN ny och relevant följdfråga
   till patienten.

    **Begränsningar**
    - Du får INTE upprepa någon fråga som redan har ställts
   tidigare i chatthistoriken.
    - Läs igenom chatthistoriken noggrant för att försäkra dig om
   detta.
    - Om en fråga liknar en tidigare fråga för mycket, måste du
   formulera om eller välja ett annat fokus.


    **Tillvägagångssätt**
    1. Analysera hela chatthistoriken:
        - Identifiera redan ställda frågor och svar.
    2. Läs igenom dokumentutdragen:
        - Om de innehåller relevant information: använd den för att
    formulera en ny och relevant följdfråga.
        - Om de inte är relevanta: basera din fråga på vad som
   saknas i chatthistoriken.
    3. Formulera en **kort, tydlig och unik** följdfråga till
   patienten.

    **Output**
    - Svara ENDAST med en ny följdfråga, inget annat.

     ------------------------
    Chatthistorik:
    {messages}

    Dokumentutdrag:
    {rag_results}
    </instruktioner>

"""
```

Listing A.4: This is the original Swedish prompt from listing 5.2

```
f"""
<instruktioner>
    Du kommer få meddelanden från chatthistoriken.
    -Sätt done till True om du anser att du anser att du har tillrä
   cklig information om patientens tillstånd baserat på
   dokumentutdragen och chatthistoriken.
    -Skriv en descripion på varför du anser att du har tillräcklig
   eller inte tillräcklig information
        Dokumentutdrag:
        {rag_results}

    **Tillvägagångssätt**
    - Analysera chatthistoriken för att se vad som redan har sagts.
    - Analysera dokumentutdragen för att se vad mer för information
    som bör samlas in.
    - Om du anser att tillräcklig information är insamlad ska du sä
   tta done till True.

    **Din uppgift**
    - Du är en chattbott på barnakuten som samlar in information
   till en sjuksköterska.
</instruktioner>

"""
```

Listing A.5: This is the original Swedish prompt from listing 5.3

```
f"""
    ***Din uppgift***
        Du ska skapa en sammanfattning i flera delar som kan hjälpa
    en sjuksköterska på barnakuten att få en överblick av en
   patients tillstånd.
        Till din hjälp har du en chatthistorik mellan patienten och
    en chatbott samt dokumentutdrag med medicinsk information och
   om arbetet på barnakuten.

        ***Tillvägagångssätt***
        - Generera en sammanfattning av chatten som en sjukskö
   terska har nytta av i fältet summary.
        - Ge mellan 1 och 3 förslag på nästa steg i behandlingen i
   fältet next_steps.
        - Ge förslag kompletterande provtagning och kontroller i
   fall där de behövs i fältet controls.

        ----------------------------------------
        Chatthistorik:
        {messages}

        Dokumentutdrag:
        {rag_results}
"""
```

Listing A.6: This is the original Swedish prompt from listing 5.4

Samtyckesblankett

Deltagande i studie kring användandet av en chatbot för informationsinsamling på barnakuten

Kontaktuppgifter:

Tindra Järgenstedt	gusjargti@student.gu.se	0722251826
Elin Nilsson	nileli@chalmers.se	0703278371

Information:
Du kommer att få testa att chatta med en chatbot som har i uppgift att samla in information om ditt barns besöksorsak och sedan skapa en sammanfattning till en sjuksköterska. Deltagandet i studien är endast för forskningssyfte och kommer inte att påverka din vård på något sätt och inte heller förlänga ditt besök. Efter att du har chattat med chatboten kommer du att bli ombedd att fylla i en enkät bestående av 7 flervalsfrågor.

- Sjuksköterskan kommer att ta del av chatthistoriken och sammanfattningen först efter att ni har blivit undersökta och fått den vård ni behöver. Därmed kommer er vård inte påverkas på något sätt av deltagandet.
- All information som ni uppger i chatten kommer att lagras i VGRs säkerhetsklassade system. Inga externa molntjänster eller språkmodeller används och ingen data kommer att skickas eller behandlas utanför VGRs system.
- Svaren som du uppger i enkäten kommer att sparas på Chalmers och Göteborgs Universitets lokala system och ingen information kommer att delas utanför dessa system. Dina enkätsvar kommer att vara helt anonyma och inte gå att koppla samman med informationen du har lämnat i chatten.
- Du har rätt att avbryta ditt deltagande när som helst. Om du kommer på efter ditt deltagande här idag att du inte längre vill vara delaktig kan du maila ansvariga för studien och be om att all data kopplad till dig tas bort.

Jag bekräftar härmed att jag har tagit del av skriftlig och muntlig information om studien och accepterar att delta. Jag har fått möjlighet att ställa frågor om studien.

Ort och Datum _____

Förnamn och Efternamn _____

Telefonnummer _____

Underskrift _____

Figure A.24: Consent Form for patients.

Samtyckesblankett

Deltagande i studie kring användandet av en chatbot för informationsinsamling på barnakuten

Kontaktuppgifter:

Tindra Järgenstedt	gusjargti@student.gu.se	0722251826
Elin Nilsson	nileli@chalmers.se	0703278371

Information:
Du kommer att ta del av chattar mellan patienter som du har triagerat och en AI-chatbot. Du kommer också att få se AI-genererade sammanfattningar innehållandes förslag på nästa steg i behandlingen samt föreslagna kontroller. Efter att ha tagit del av en sammanfattning och chatt tillhörande en patient kommer du att få svara på en enkät, innehållandes 9 frågor, angående vad du tycker om sammanfattningen samt konversationen.

- Svaren som du uppger i enkäten kommer att sparas på Chalmers och Göteborgs Universitets lokala system och ingen information kommer att delas utanför dessa system. Dina enkätsvar kommer att vara helt anonyma.
- Du har rätt att avbryta ditt deltagande när som helst. Om du kommer på efter ditt deltagande här idag att du inte längre vill vara delaktig kan du maila ansvariga för studien och be om att all data kopplad till dig tas bort.

Jag bekräftar härmed att jag har tagit del av skriftlig och muntlig information om studien och accepterar att delta. Jag har fått möjlighet att ställa frågor om studien.

Ort och Datum _____

Förnamn och Efternamn _____

Telefonnummer _____

Underskrift _____

Figure A.25: Consent Form for nurses.

A.6 Ethical approval to conduct research with patients

SAHLGRENSKA UNIVERSITETSSJUKHUSET VGR

Medicin barn
Akutmottaning barn

Godkännande för examensarbete
2025-04-24

Godkännande verksamhetsutvecklande projekt

Examensarbete i samarbete med Göteborgs Universitet och Chalmers:

- **"Applying and evaluating Large Language Models for triage at a Paediatric Emergency Department in a Swedish hospital"**

Härmed godkänns att Akutmottagning barn medverkar i ovan projekt med Elin Nilsson och Tindra Järgenstedt. Akutmottagningen barn bidrar med stöd i utveckling, samt möjlighet till att kliniskt testa LLM på akutmottagningens patienter under formen verksamhetsutvecklande projekt.

24/4 2025

Joanna Pestalozzi,
Verksamhetschef Medicin barn

Datum

Sahlgrenska Universitetssjukhuset
413 45 Göteborg
031-342 10 00 - sahlgrenska.se

Sida 1 av 1

Figure A.26: Ethical approval to conduct this research with real patients.

A.7 WEST-P compendium

WEST-P:

West coast System for Triage - Pediatric

VGR Kompendium: ver. 6

VÄSTRA GÖTALANDSREGIONEN
SAHLGRENSKA UNIVERSITETSSJUKHUSET

INNEHÅLLSFÖRTECKNING

Kontakt: Hannah.sjostedt@vgregion.se

WEST-P uppbyggnad

WEST-P har två olika delar som kan generera en triagefärg: varningssymtom och sammanvägda poäng för barnets vitalparametrar (enligt Triage-Poäng). Högst färg enligt respektive del ger barnets slutfärg i WEST-P. Om barnet inte har röd, orange eller gul varningssymtom och har opåverkade vitalparametrar (0-2 Triage-Poäng) blir dess triagefärg grön. Vid isolerad ortopedisk extremitetsskada behöver inte alla vitalparametrar mätas och patienten kan triageras efter varningssymtom.

En triagerande sjuksköterska eller läkare kan alltid välja en högre triagefärg än WEST-P, men får inte välja en lägre triagefärg utan att konsultera ansvarig läkare.

Kommande sidor spaltar upp varningssymtom och Triage-Poäng, därefter kommer mer ingående förklaringar för hur en ska bedöma varningssymtom.

Varningssymtom: 1 av 3

	Röd prioritet (Läkare omedelbart)	Orange prioritet (Läkare inom 10min)	Gul prioritet (Läkare inom 60min)
C	Hjärtstopp		
	Traumalarm nivå 1-2	(Traumalarm nivå 3)	
A	Luftväg: ofri, hotad, främmande kropp, intuberad, trauma mot halsen med svullnad		
	Anafylaxi	Akut allvarlig allergisk reaktion eller tidigare anafylaxi på ämnet	
B	Andningsbesvär: svårt ansträngd, allvarligt obstruktiv eller apnéer under triagering		Andningsbesvär: lätt till måttlig ansträngd/obstruktiv
C	Okontrollerad pågående blödning	Kräkning: pågående kaskadkräkning eller större mängd färskt blod	Koagulationshämmande läkemedel eller blödningssjukdom **och** • Lindrigt trauma; eller • Liten blödning; eller • Ledvärk
		Graviditet **och** • Vaginell blödning; eller • Buksmärta; eller • BT ≥160/110	Barn som efter ett trauma inte kan kissa eller kissar blod
			Bröstsmärtor: pågående

Varningssymtom: 2 av 3

	Röd prioritet (Läkare omedelbart)	Orange prioritet (Läkare inom 10min)	Gul prioritet (Läkare inom 60min)
D	Medvetslös patient	Medvetandegrad: sänkt, slö, förvirrad, agiterad	Medvetslös (>1min) prehospitalt eller upprepade medvetandeförluster senaste dygnet
	Krampanfall: pågående		Krampanfall prehospitalt
		Misstanke om allvarlig intoxikation eller bett av giftig orm	Intoxikation, förgiftning eller bett av djur
			Huvudvärk: • Plötsligt isättande kraftig huvudvärk; eller • Huvudvärk eller kräkningar och känd hydrocefalus/shunt/hjärntumör
	Huvudskada: med sänkt medvetande eller pupillpåverkan	Huvudskada: • Misstänkt skallbasfraktur; eller • Blödningssjukdom och lindrigt trauma mot huvudet	Huvudskada med anamnes på • Medvetslös >1 min; eller • Amnesi >5 min; eller • Upprepade kräkningar
	Hypoglykemi: glukos <3,0 mmol/l i triagen eller prehositalt	Hyperglykemi: • Glukos >11 mmol/l och andningspåverkan • Misstänkt nydebuterad diabetes	
	Neurologiska bortfall: symtom med <8 timmar duration med/utan trauma		Neurologiska bortfall: symtom med 8-24 timmar duration med/utan trauma

Varningssymtom: 3 av 3

	Röd prioritet (Läkare omedelbart)	Orange prioritet (Läkare inom 10min)	Gul prioritet (Läkare inom 60min)
E	Sepsis- eller meningit-misstanke		
		Smärtor: akut och stark smärta eller otröstligt barn, eller smärta och påverkat AT	Smärtor: måttliga
	Amputation/fraktur med misstänkt kärlskada ovanför hand/fot	Amputation av finger/tå med delen medtagen till akuten Öppen/gravt felställd fraktur	Felställd fraktur eller luxerad led
		Främmande kroppar: Batteri, eller svalt ≥2 magneter	Svalt: främmande kropp men kan inte svälja saliv
		Ögonskada: frätskada eller penetrerande våld	Öga: svullen/rodnad kring öga och samtidig feber Öra: rött bakom/utåtstående öra och samtidig feber
		Feber (≥38,0) hos • Neutropen; eller • Immunosupprimerad; eller • Barn ≤3 månader	Barn med • ≤2 månaders ålder; eller • Allvarlig grundsjukdom; eller • Misstanke om barn som far illa; eller • Psykisk ohälsa; eller • Malignitetssuspekta blodprover
		Snabbt tilltagande rodnad/gasbildning i huden	Infektionstecken/blödning och opererad ≤14 dagar sedan
	Brännskada på ansikte/hals (ej bara droppstänk). Inhalationsskada. Högspänningsolycka.	Brännskada ≥10% eller cirkumferent	Brännskada ≤10% på barn ≤1 år

< 1 månad	0	1	2	3
Andningsfrekvens	40 – 55	56 – 64 25 – 39	65 – 79	≥ 80 < 25
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	100 – 160	161 – 169 85 - 99	170 – 189	≥190 < 85
Kap Å	1-2s	3s		≥ 4s
Temp	35 – 38		≥38,1 <35	

1 - 12 månader	0	1	2	3
Andningsfrekvens	35 - 45	46 – 54 20 – 34	55 – 69	≥ 70 < 20
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	100 – 160	161 – 169 80 – 99	170 – 189	≥190 < 80
Kap Å	1-2s	3s		≥ 4s
Temp	35 – 38	38,1 – 39	≥ 39,1 < 35	

1 - 3 år	0	1	2	3
Andningsfrekvens	25 – 35	36 – 44 20 – 24	45 – 59	≥ 60 < 20
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	90 – 130	131 – 139 70 – 89	140 – 159	≥160 < 70
Kap Å	1-2s	3s		≥ 4s
Temp	35 – 38	38,1 – 39	≥ 39,1 < 35	

4 - 6 år	0	1	2	3
Andningsfrekvens	20 – 24	25 – 29 15 – 19	30 – 44	≥ 45 < 15
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	70 – 120	121 – 129 60 – 69	130 - 149	≥150 < 60
Kap Å	1-2s	3s		≥ 4s
Temp	35 – 38	38,1 – 39	≥ 39,1 < 35	

7 - 12 år	0	1	2	3
Andningsfrekvens	19 – 22	23 – 29 14 – 18	30 – 39	≥ 40 < 14
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	70 – 110	111 – 119 60 – 69	120 – 139	≥ 140 < 60
Kap Å	1-2s	3s		≥ 4s
Temp	35 – 38	38,1 – 39	≥ 39,1 < 35	

13 - 14 år	0	1	2	3
Andningsfrekvens	14 – 19	9 – 13	20 - 29	≥ 30 < 9
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	55 – 95	96 – 114 45 – 54	115 – 129	≥ 130 < 45
BT (syst.)	101 – 180	81 – 100	≥ 180 71 – 80	≤ 70
Temp	35 – 38	38,1 – 39	≥ 39,1 < 35	

≥ 15 år	0	1	2	3
Andningsfrekvens	9 – 14	15 – 20	21 – 29 ≤ 8	≥ 30
SpO_2	≥ 95 %	93 – 94 %	90 – 92 %	Kräver O2
Puls	51 – 100	101 – 110 41 – 50	111 – 129 ≤ 40	≥ 130
BT (syst.)	101 – 199	81 – 100	≥ 200 71 – 80	≤ 70
Temp	35 – 38	38,1 – 39	≥ 39,1 < 35	

Poäng:	0 – 2	3 – 4	5 – 6	≥7

Triage-Poäng

Kontroller under pågående besök: rekommendation

Kontroller och åtgärd

Innan läkare bedömt barnet ska kontroller tas minst varannan timma: tillsyn och kontroll av de parametrar som gav poäng på Triage-Poäng under triagen. Dokumentation ska ske på akutjournalen.

När barnet är påtittat av läkare ska ansvarig läkare aktivt besluta om vilka parametrar (inklusive medvetandegrad) utöver tillsyn som ska kontrolleras och hur ofta. Om barnet vid något tillfälle försämras med minst 1 poäng enligt Triage-Poäng, eller har påverkad medvetandegrad ska ansvarig läkare meddelas och ta ställning till åtgärder.

Frekvens

Barnen ska kontrolleras minst varannan timma under väntan om inte ansvarig läkare beslutar om mer frekventa kontroller.

C: Hjärtstopp

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Hjärtstopp		

Förklaring och definition:

- Hjärtstopp, pågående
- Hjärtstopp prehospitalt med återfådd cirkulation vid ankomst
- Inga livstecken

Åtgärd:

- Larma Hjärtlarm enligt lokal rutin

Sökorsak i Elvis
HJÄRTSJD

C: Trauma

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Traumalarm 1-2	(Traumalarm 3)	

Nivå 1	Nivå 2	Nivå 3 (DSBUS)
Fysiologiska kriterier •Behov av ventilationsstöd •Andningspåverkat barn •KÅ >2s •Puls •0-1 år: <90 eller >190 •1-5 år: <70 eller >160 •6-16 år: <45 eller >130 •RLS 2 eller mer / GCS 13 eller lägre **Anatomiska kriterier** •Penetrerande våld mot hals, huvud, bål, extremiteter ovan armbåge/knä •Öppen skallskada eller impressionsfraktur •Ansikts- eller halsskada med hotad luftväg •Instabil eller deformerad bröstrygg (nytillkommet efter traumat) •Svår smärta i bäckenet eller misstänkt instabil bäckenfraktur •Misstänkt ryggmärgsskada (med symtom) •2 eller fler frakturer på långa rörben •Amputation ovan hand eller fot •Stor yttre blödning •Brännskada >18% eller inhalationsskada	**Skademekanism** •Bilolycka >50 km/h utan bilbälte •Utkastad ur fordon •Fastklämd med losstagningstid >20 min •Inblandad i tvåhjulig fordonsolycka >35 km/h •Påkörd eller överkörd av motorfordon eller motsvarande •Fall från >3 meter	•Brännskador i ansiktet •Högenergivåld* som ej faller ut inom traumalarm nivå 1 eller 2 •Skador som skett för mer än 6 timmar sedan och normalt skulle gett nivå 1 eller 2 larm där vitalparametrar för traumalarmskriterier är normala *Högenergivåld får sättas i relation till barnets storlek men exempelvis motorfordonsolyckor, kontaktidrotter med hög fart, fall från hög höjd.

Sökorsak i Elvis

TRAUMAN1
TRAUMAN2
TRAUMAN3

Åtgärd:

- Traumalarm 1-2: larma 127/akutrum och via växeln 39090 ”Traumalarm nivå X, Barnakuten Akutrum”. Etablera minst en infart.
- Traumalarm 3: överväg att larma 127/akutrum

A: Luftväg

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Luftväg: ofri, hotad, främmande kropp, intuberad, trauma mot halsen med svullnad		

Sökorsak i Elvis

-

Förklaring och definition:

- Ofri luftväg är när luft inte kan passera mellan näsa/mun och lungor. Detta kan åtgärdas genom enkla manövrar eller avancerade hjälpmedel.
- Hotad luftväg är ett tillstånd när luftvägen riskerar att bli ofri, exempelvis vid inhalationsskador, tilltagande svullnad i halsen som epiglottit, trauma mot halsen med tilltagande svullnad. Vid svullnad i halsen finns ofta tilltagande stridor i vila och barnet har påverkad allmäntillstånd. Inspiratorisk stridor vid aktivitet som vid exempelvis krupp är inte en hotad luftväg.
- Främmande kropp: vid en högt sittande främmande kropp i svalg eller ovan bronkerna pendlar ofta luftvägen mellan att vara hotad och luftfri. Den främmande kroppen kan ruckas om barnets läge ändras eller barnet hostar. Om en främmande kropp sitter längre ned, eller anamnesen är att barnet lekt med lego i munnen, plötsligt inte kunnat andats och därefter börjar hosta men är opåverkat vid triagering har den främmande kroppen troligen hamnat i en bronk eller längre ned och barnets luftväg är inte hotad.

- **Åtgärd**: Larma 127/akutrum

A: Allergi

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Anafylaxi	Akut allvarlig allergisk reaktion eller tidigare anafylaxi på ämnet	

Sökorsak i Elvis
ALLERGI

Förklaring och definition:

- Inte akut allvarlig allergisk reaktion: begränsad urtikaria på delar av kroppen
- Akut allvarlig allergisk reaktion (ej anafylaxi): klåda, utbredd urtikaria, angioödem, svullnadskänsla i mun och svalg, läppsvullnad, enstaka kräkning
- Anafylaxi grad 1: ovan samt exempelvis ökande buksmärta, upprepade kräkningar, diarré, heshet, lindrig bronkobstruktion, uttalad trötthet, rastlöshet/oro
- Anafylaxi grad 2: ovan samt exempelvis skällhosta, sväljningsbesvär, medelsvår bronkobstruktion, svimningskänsla, katastrofkänsla
- Anafylaxi grad 3: ovan samt exempelvis urin- och/eller fecesavgång, hypoxi, cyanos, svår bronkobstruktion, andningsstopp, hypotoni, bradykardi, arytmi, hjärtstopp, förvirring, medvetslöshet

Åtgärd: Förbered för betapred och aerius per os till alla samt adrenalin intramuskulärt till anafylaxi (0,01 ml/kg av 1 mg/ml, max 0,5 ml).

B: Andningsbesvär

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Andningsbesvär: svårt ansträngd, allvarligt obstruktiv eller apnéer under triagering		Andningsbesvär: lätt till måttlig ansträngd/obstruktiv

Sökorsak i Elvis
DYSPNE
PAND
ÖLI

Förklaring och definition:

•Svårt ansträngd andning: barn som är kraftigt påverkade av sin ansträngda andning har ofta indragningar, gravt ökad andningsfrekvens och får kämpa för att andas. Barnen orkar då inte leka, medverka, eller skratta vid undersökningen. De sitter ofta hos föräldern, och all energi går åt till att andas.

•Apné: totalt andningsuppehåll på >20 sekunder räknas som apné. Spädbarn andas oregelbundet och slutar ofta under några sekunder, detta är inte apné.

•Lätt-måttlig ansträngd andning: barn med förhöjd andningsfrekvens och ofta indragningar men som ändå orkar medverka till undersökningen eller leka trots sitt ökade andningsarbete.

C: Blödning och cirkulation

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Okontrollerad pågående blödning	Kräkning: pågående kaskadkräkning eller större mängd färskt blod	Koagulationshämmande läkemedel eller blödningssjukdom **och**: •Lindrigt trauma; eller •Liten blödning; eller •Ledvärk

Förklaring och definition:

•Okontrollerad pågående blödning: blödning som inte går att få stopp på med exempelvis manuellt tryck eller tryckförband. Blodet fortsätter att rinna trots förband. Detta är farligt eftersom blodförlusten då fortsätter. Blödning som rinner konstant från rektum eller underlivet räknas som okontrollerad pågående blödning. Ett sår som inte blöder efter man lagt ett tryckförband är inte farligt och räknas inte som en ”okontrollerad pågående blödning” eftersom man kontrollerat blödningen och fått den att avstanna.

•Kräkning:

•Kaskad: till kaskadkräkningar räknas stora projektilkräkningar som pågår konstant och oavbrutet. Detta är en stor risk till uttorkning och tecken på allvarlig underliggande sjukdom som exempelvis ileus eller pylorusstenos. Dessa barn är allmänpåverkade.

•Större mängd färskt blod: kräket ska bestå av större mängd blod än matrester/magsaft, hit räknas inte blodstrimmor.

•Koagulationshämmande läkemedel eller blödningssjukdom: om koagulationssystemet är påverkat hos ett barn på grund av sjukdom eller läkemedel kan ett mindre trauma eller mindre sår innebära en större blödning än normalt och på så sätt vara farligt. Ledvärk kan vara tecken på spontan blödning i leden utan föregående trauma vilket bör åtgärdas fort för att minska risken för kronisk skada.

•Till koagulationshämmande läkemedel räknas waran/eliquis/NOAK. Ipren/NSAID påverkar trombocyterna men räknas inte som koagulationshämmande.

Sökorsak i Elvis
FALLOLY
HEMA
HEMATE
MELENA
NAUSEA
SMÄRTA
SÅRSKADA
TOL
TRTHOBUK

C: Graviditet eller urinbesvär

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
	Graviditet **och** •Vaginell blödning; eller •Buksmärta; eller •BT ≥160/110	Barn som efter ett trauma inte kan kissa eller kissar blod

Förklaring och definition:

•Graviditet är ovanligt före 16 års ålder men det är möjligt att våra patienter kan vara gravida. Hos en gravid flicka med vaginell blödning eller buksmärta måste man framförallt utesluta ektopisk graviditet (X-graviditet) som är en livshotande komplikation. Större blödning vid missfall kan också vara livshotande och måste identifieras. Högt blodtryck är ett tecken på preeklampsi/eklampsi som kan ge upphov till krampanfall och status epilepticus.

•**Åtgärd**: ta graviditetstest på alla flickor >12 år gamla och söker med urinbesvär/buksmärta.

•Barn som efter trauma inte kan kissa eller kissar synligt blod: blod i urinblåsan riskerar att koagulera och måste då evakueras på operation. Med barn som inte kan kissa menas inte de barn som har minskade urinmängder på grund av exempelvis ett infektionstillstånd utan snarare de som råkat ut för ett trauma och därefter inte kan kissa då det finns risk för nervpåverkan från rygg/bäcken eller skada på njure/urinblåsa/uretra. Även de barn med ryggsmärta som inte kan kissa räknas hit oavsett om det varit ett föregående trauma eller inte.

Sökorsak i Elvis
BUKSMÄRT
FALLOLY
TOL
TRTHOBUK
UNDERLIV
URINB

C: Bröstsmärta

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
		Bröstsmärtor: pågående

Förklaring och definition:

- Barn som söker med pågående bröstsmärtor kan exempelvis uppge huggsmärtor i ”hjärtat” eller bröstkorgen som kan tala för pneumothorax eller pleurit. Är bröstsmärtan mer konstant kan det vara perimyokardit, då är den ofta lägesberoende och ibland hörs gnisslande/knarrande biljud när man lyssnar på hjärtat. Små barn kan ha svårt att beskriva vad de upplever och kan ofta klaga på bröstsmärta i samband med tachyarrytmier.
- Palpabel bröstsmärta över bröstkorgen är lokaliserad till revbenen eller musklerna och har i regel inte kardiell genes.

- **Åtgärd**: ta EKG på alla, förbered för blodprovtagning

Sökorsak i Elvis
BRSM

D: Medvetandegrad

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Medvetslös patient	Medvetandegrad: sänkt, slö, förvirrad, agiterad	Medvetslös (>1 min) prehospitalt eller upprepade medvetandeförluster senaste dygnet

Förklaring och definition:

- Medvetslös patient innebär ett barn som man inte kan få kontakt med oavsett om man pratar med, rör vid eller smärtstimulerar barnet.
- Barn som inte är medvetslösa men ändå har påverkat medvetande måste hanteras skyndsamt. Detta är exempelvis barn som kommer in med infektion och då är påverkad medvetandegrad ett tecken till sepsis eller meningit. Ett barn som är förvirrad eller agiterad efter ett trauma mot huvudet kan ha en underliggande hjärnblödning som kan behöva akut åtgärd.
- Föräldrar upplever ofta att barnet ”inte är sig själv” och att man ”inte får kontakt” med barnet men detta behöver inte vara samma som medicinskt påverkad medvetandegrad och vi måste bedöma barnet noggrant.
- Upprepade medvetandeförluster senaste dygnet kan vara ett tecken på underliggande hjärtsjukdom och hjärtarrytmi.

Åtgärd: ta EKG på alla, blodglukos och Hb. Överväg blodtryckskontroll. Överväg omhändertagande som larmfall.

Sökorsak i Elvis
NEURO
OSPEC
SKALLSK

D: Krampanfall

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Krampanfall: pågående		Krampanfall prehospitalt

Förklaring och definition:

- Krampanfall innebär att barnet har ryckt symmetriskt i antingen hela kroppen eller någon kroppsdel: oavsett om barnet har känd epilepsi eller inte. Det finns många typer av andra allvarliga anfallssjukdomar, exempelvis Infantil spasm där barn < 2 år får sekundkorta ryckningar. Dessa ryckningar är oftast böjrörelser i armar och midja och kommer i kluster på 10-30 per omgång.
- Har barnet krampat prehospitalt och är helt välmående får de gult varningssymtom, men om deras medvetandegrad vid ankomsten till akuten är exempelvis sänkt eller förvirrad så faller de ut som orange varningssymtom enligt ”medvetandegrad” (se varningssymtom sida 16).

Åtgärd: ta EKG och blodglukos.

Sökorsak i Elvis
KRAMPER

D: Förgiftning, bett

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
	Misstanke om allvarlig intoxikation eller bett av giftig orm	Intoxikation, förgiftning eller bett av djur

Förklaring och definition:

- Förgiftning kan ske avsiktligt eller oavsiktligt, det är viktigt att samla in så mycket information som möjligt om VAD som intagits, HUR MYCKET och NÄR det skedde. Om det är oklara omständigheter eller bristande information kan det finnas en stor anledning till att misstänka allvarlig intoxikation.
- Lindrigare intoxikationer som exempelvis att ett barn råkat få en dubbeldos av ett läkemedel en gång är oftast mindre akut och ger gult varningssymtom.
- Bett av större djur som katt, hund eller kanin kan innebära hög infektionsrisk och såret ska tvättas så snart som möjligt. Till bett av djur räknas inte insektsbett.
- Vid ormbett är det viktigt att veta om det var en giftig orm. Bett från en giftig eller okänd orm räknas till misstanke om allvarlig intoxikation.

Åtgärd: kontakta giftinformationscentralen (via växeln eller 010-4566719), diskutera eventuell orosanmälan med teamet (all intoxikation eller självskadebeteende). Markera ut eventuell rodnad kring bett i triagen för att kunna utvärdera hur snabbt det sprids.

Sökorsak i Elvis
BETT
INTOX

D: Huvudvärk

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
		Huvudvärk: • Plötsligt insättande kraftig huvudvärk; eller • Huvudvärk eller kräkningar och känd hydrocefalus/shunt/hjärntumör

Förklaring och definition:

•Plötsligt insättande kraftig huvudvärk hos barn är oftast migrän och är barnet opåverkat i sina vitalparametrar ska det få gul varningssymtom. Spontan hjärnblödning är ovanlig hos barn, och dessa barn har då oftast också exempelvis påverkat medvetande och/eller påverkan på sina andra vitalparametrar.

•All form av huvudvärk eller kräkningar hos en patient med känd sjukdom intrakraniellt som skulle kunna ge högre intrakraniellt tryck bör prioriteras skyndsamt även om barnet är helt opåverkat för stunden. Detta är exempelvis helt välmående barn med shunt, hydrocefalus eller hjärntumör som leker men klagar på huvudvärk, eller haft en eller ett par kräkningar.

•**Åtgärd**: smärtlindring, förbered ev infart

Sökorsak i Elvis
HVVÄRK
NAUSEA
SHUNT

D: Huvudskada

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Huvudskada: med sänkt medvetande eller pupillpåverkan	Huvudskada: • Misstänkt skallbasfraktur; eller • Blödningssjukdom och lindrigt trauma mot huvudet	Huvudskada med anamnes på: • Medvetslös > 1 min; eller • Amnesi > 5 min; eller • Upprepade kräkningar

Förklaring och definition:

•Med huvudskada innebär tillstånd där barnet fått ett slag, ramlat eller på annat sätt utsatts för ett isolerat trauma mot huvudet.

•Sänkt medvetande eller pupillpåverkan och trauma ger misstanke om hjärnblödning och ska handläggas urakut.

•Skallbasfraktur misstänks vid exempelvis brillenhematom (blåtiror över båda ögonen), blod i hörselgången, bakom trumhinnan, eller bakom öronen. Ibland finns då även hjärnblödning och barnet bör omhändertas fort. Det finns även risk för att frakturen påverkar kranialnerverna.

• En mjuk bula på huvudet är ett tecken på fraktur i skelettet och är inte farlig i sig, men kan vara associerat med hjärnblödning. Är barnet opåverkat i övrigt och har den mjuka bulan som enda symtom har barnet inget varningssymtom.

•Barn med blödningssjukdom blöder lättare och riskerar hjärnblödning även vid lindrigare trauma och ska därför undersökas fort.

•**Åtgärd**: överväg traumalarm.

Sökorsak i Elvis
SKALLSK

D: Hypo- och hyperglykemi

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Hypoglykemi: glukos <3,0 mmol/l i triagen eller prehospitalt	Hyperglykemi • Glukos >11 mmol/l och andningspåverkan • Misstänkt nydebuterad diabetes	

Förklaring och definition:

•Blodglukos <3,0 mmol/l är livsfarligt och kan bland annat resultera i krampanfall och ska behandlas akut. Har barnet haft hypoglykemi i hemmet som nu stigit ska det ändå handläggas snabbt eftersom blodglukoset kan sjunka snabbt igen om ingen behandling pågår.

•Hyperglykemi innebär att mycket glukos finns i blodet men tas inte upp av cellerna vilket innebär att kroppen hamnar i ett svälttillstånd och till slut ketoacidos. Detta är ett livsfarligt tillstånd och ska behandlas fort med vätska. Vid nydebuterad diabetes är blodglukos oftast högt och kroppen riskerar att gå in i detta tillstånd.

•**Åtgärd**: blodglukos och blodketoner som patientnära analys. Förbered för blodgas och infart. Uppmana barnet att äta vid hypoglykemi, ge isglass.

Sökorsak i Elvis
DIAB
OSPEC

D: Neurologiska bortfall

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Neurologiska bortfall: symtom <8h med/utan trauma.		Neurologiska bortfall: symtom 8-24h med/utan trauma.

Förklaring och definition:

•Med neurologiska bortfall menas bortfall i motorik eller sensorik. Exempelvis fumlighet och/eller sluddrigt tal är tecken på nedsatt motorik. Det vanligaste neurologiska bortfallet på barnakuten är hängande mungipa och öga vilket ofta är neuroborrelios. Neurologiskt bortfall med trauma ger misstanke om skada på nervsystemet eller nervpåverkan – då är det också viktigt att bedöma om det istället rör sig om ett traumalarm och välja det varningssymtom som ger högst prioritet för barnet.

•Neurologiskt bortfall som uppstått senaste 8 timmarna ska handläggas fort eftersom det kan vara orsakat av något som går att åtgärda i tidigt skede som exempelvis stroke.

•Neurologiska bortfall som varat längre behöver oftast ingen akut behandling och därför är handläggningen inte lika akut.

•Isolerad perifer facialispares (på ena ansiktshalvan svårt att rynka pannan, blunda med ögat, höja mungipan) är hos barn oftast neuroborrelios eller idiopatiskt Bells pares får inte rött varningssymtom även om barnet inkommer med en symtomduration på kortare än 8 timmar.

Sökorsak i Elvis
NEURO
OSPEC
SKALLSK

E: Sepsis, meningit

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Sepsis- eller meningit-misstanke		

Förklaring och definition:

•Sepsis är bakterier i blodet i samband med infektion och kan vara svårt att misstänka hos barn, speciellt eftersom symptomen är olika i olika åldrar och tillståndet ofta är diffust. Feber förekommer oftast men inte alltid. Hos små barn under 3 månader ska sepsis och/eller meningit misstänkas exempelvis om barnet har feber och samtidigt är irritabelt/otröstligt, påverkat allmäntillstånd. Hos äldre barn med feber kan det yttra sig med förvirring, påverkan på flera organsystem samtidigt som exempelvis hosta, diarré, kräkning, buksmärta. Barn med feber har ofta påverkat AT, men om de är mer medtagna än vad som känns rimligt är det bra att utesluta sepsis. Sepsis och/eller meningit ska misstänkas på äldre barn som inte kan stå på benen och är konfusoriska.

•Meningit: klassiska symptom är triaden feber, nackstyvhet/buktande fontanell och påverkad medvetandegrad men detta finns inte hos alla. Petekier och purpura kan förekomma. Meningit kan orsakas av bakterier och virus: bakteriell meningit är ofta mer akut insättande och aggressivare och behöver behandlas oerhört fort med intravenös antibiotika och LP. Nackstel innebär att barnet inte ens med hjälp kan sätta hakan i bröstet eftersom hjärnhinnorna är så pass infekterat att de är för strama. Nackstelhet är inte att det gör ont i nacken när man rör huvudet men man ändå kan sätta hakan i bröstet. Nackstelhet är inte att barnet inte kan vrida huvudet åt sidorna, detta är nackspärr oavsett om barnet har feber eller inte. Dock har inte alla barn med meningit nackstelhet, så man måste bedöma hela patienten.

•**Åtgärd**: förbered för infart, provtagning inkl blododling

Sökorsak i Elvis
FEBER
NEURO
PAND

E: Smärta

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
	Smärtor: akut och stark smärta, eller otröstligt barn, eller smärta och påverkat AT	Smärtor: måttliga

Förklaring och definition:

• Stark smärta: Ett barn som är otröstligt, ser smärtpåverkad ut i vila eller uppger smärta och har påverkat AT räknas som starka smärtor. Barn som har ont skrotalt och går bredbent (”cowboygång”) har stark smärta. Ett barn med misstänkt fraktur som ser ut att ha ont även när man immobiliserat armen/benet räknas som stark smärta.

•Måttlig smärta: ett barn som verbalt uppger smärta men som är opåverkat i status (exempelvis leker, pratar lugnt) räknas till måttlig smärta. Hit räknas också misstänkta frakturer som inte gör ont om man inte rör vid armen/benet. Dessa barn ska givetvis få smärtlindring, men de räknas som gul varningssymtom.

•**Åtgärd:** smärtlindra, förbered ev för infart

Sökorsak i Elvis
BRÅCK
BUKSMÄRT
EXTREMIT
HVVÄRK
KNÄ
POSTOP
SKROTAL
SMÄRTA
TRTHOBUK

E: Amputation, fraktur, luxation

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Amputation/Fraktur med misstänkt kärlskada ovanför hand/fot	Amputation av finger/tå med delen medtagen till akuten	Felställd fraktur eller luxerad led
	Öppen/gravt felställd fraktur	

Förklaring och definition:

- Amputation i nivå med handled/fotled eller mer proximalt innebär alltid kärlskada. Fraktur ovan samma nivå med kärlskada ger ofta svagare/avsaknad pulsar, kraftiga smärtor och iskall distalt.
- En tå eller finger som är amputerad med delen medtagen till akuten ger orange varningssymtom eftersom det finns en möjlighet att sy tillbaka delen ju tidigare det görs. Är delen inte medtagen blir varningssymtomet gult eftersom åtgärden inte är lika tidskritisk.
- Öppna frakturer ska rengöras och antibiotika ska ges så snart som möjligt för att minska risken för infektion.
- Gravt felställd fraktur innebär en större felställning än en böjd eller krokig underarm. Hit hör exempelvis felställda fotledsfrakturer eller ordentlig bajonettfraktur i handleden. En grav felställning kan medföra tryck och påverkan på huden kring frakturen och behöver då grovreponeras snarast och generar orange varningssymtom.
- En fraktur som är felställd/böjd och inte riskerar hudpåverkan får gul varningssymtom så länge smärtorna inte är kraftiga även när frakturen är immobiliserad.
- En luxerad axel eller patella får gul varningssymtom.

- **Åtgärd**: smärtlindra, förbered för imobilisering och gipsning, förbered infart för att ge antibiotika om öppen fraktur misstänks

Sökorsak i Elvis
EXTREMIT
KLÄMSKAD
KNÄ
SÅRSK

E: Främmande kroppar

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
	Främmande kroppar: Batteri, eller svalt ≥2 magneter	Svalt: främmande kropp men kan inte svälja saliv

Förklaring och definition:

- Batteri av alla slag kan ge frätskada med men för livet om de fastnar i esofagus och måste avlägsnas med gastroskopi akut. Batterier kan ge frätskador vid all ihållande kontakt med slemhinna (näsa, matstrupe, ändtarm, vagina) och innebär orange varningssymtom om de sitter fast i kontakt med slemhinna.
- En svald magnet innebär ingen akut risk, men om barnet svalt två eller fler magneter kan magneterna klämma vävnad mellan sig som kan gå i nekros.
- Vid svald främmande kropp i magsäck eller tarmar kan avvakta utan att tillståndet försämras, men en främmande kropp som sitter kvar i esofagus ska avlägsnas med gastroskopi inom 6 timmar för att förhindra tryckskador på esofagus. Symtom på detta är att barnet svalt en främmande kropp och nu kräks så fort den dricker något och/eller inte ens kan svälja sitt saliv. Eftersom det är tidskritiskt är det värdefullt om dessa barn snabbt får träffa läkare på akuten för att kunna handläggas och operationplaneras fort.

Sökorsak i Elvis
FRKROPP

E: Ögon, öron

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
	Ögonskada: frätskada eller penetrerande våld	Öga: svullen/rodnad kring öga och samtidig feber Öra: rött bakom/utåtstående öra och samtidig feber

Förklaring och definition:

•Vid trauma mot ögat där en skada på ögonbulben uppmärksammas måste man utesluta att ögat har fått en penetrerande skada och barnet har orange varningssymtom. Detta gäller även vid kontakt av frätande medel exempelvis basiska ämnen som finns i rengöringsmedel och kaustik soda. Detta kan hota synen och ska omhändertas fort.

•**Åtgärd**:

- •Penetrerande skada: rör inte ögat på något sätt då även lätt tryck kan försämra skadan.
- •Frätande medel: påbörja spolning av ögat direkt om man inte redan gjort de prehospitalt.

•Rött/svullet öga: ethmoidit är en infektion från bihålan som på barn kan spridas och trycka på ögat. Oftast börjar rodnaden medialt vid gränsen öga/näsrot och sprider sig för att engagera båda ögonlocken. Intravenös antibiotika behövs och ibland även kirurgi.

•Öra utåtstående/rött bakom: ett utåtstående öra eller en rodnad och svullnad över mastoiden bakom örat kan vara en mastoidit som är en komplikation till öroninflammation. Oftast ses en öroninflammation på samma sida men mastoidit kan ibland uppstå flera veckor efter öroninflammationen läkt ut. Intravenös antibiotika behövs och oftast även kirurgi.

•**Åtgärd**: förbered för infart och provtagning inkl blododling

Sökorsak i Elvis
ÖGON
ÖRON

E: Feber, infektion, postoperativ

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
	Feber (≥38,0) hos • Neutropen; eller • Immunosupprimerad; eller • Barn ≤ 3 månader	Infektionstecken/blödning och opererad ≤ 14 dagar sedan
	Snabbt tilltagande rodnad/gasbildning i huden	

Förklaring och definition:

•Patienter med nedsatt imunförsvar (neutropen, immunosupprimerad) löper stor risk att ha en underliggande allvarlig infektion som sepsis. Samma gäller för Barn ≤ 3 månader eftersom deras imunförsvar inte är helt utvecklat än. Om barnen ≤ 3 månader haft febern uppmätt rektalt hemma är det lika stor risk för baktierier i blodet hos dem även om de är afebrila på akutmottagningen jämfört med de som har feber här och räknas därför som orange varningssymtom. Dessa patienter ska omhändertas fort eftersom de oftare har sepsis än andra barn.

•Snabbt tilltagande rodnad/gasbildning i huden kan vara nekrotiserande fasciit och är ett tillstånd som måste opereras urakut. Med snabbt tilltagande menas en rodnad som ökar med 1 cm i diameter varje timma eller fortare. Gasbildning känns som blåsor i huden eller underhuden utan föregående bränn/köldskada.

•Postoperativt kan tillstånd med rodnad eller blödning på hudnivån innebära en större komplikation i kroppen, exempelvis om man opererat in främmande material. Dessa barn får gul varningssymtom.

•**Åtgärd**: förbered för infart, provtagning, urinsticka, troligen även blododling och ibland LP. Markera ut rodnad i triagen om föräldrarna uppger att den sprids fort för att kunna utvärdera hastigheten senare.

Sökorsak i Elvis
EXANTHEM
FEBER
LOKINF
POSTOP
ÖLI

E: Speciella omständigheter

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
		Barn med • ≤ 2 månaders ålder; eller • Allvarlig grundsjukdom; eller • Misstanke om barn som far illa; eller • Psykisk ohälsa; eller • Malignitetssuspekta blodprover

Förklaring och definition:

•Vissa patientkategorier bör av olika skäl inte vänta för länge på läkarbedömning även om tillstånden inte är urakuta, dessa får gul varningssymtom.

Sökorsak i Elvis
DÅLVI
GRÅTB
INTOX
MISSH
OSPEC
PSYOHÄL

E: Brännskador

VARNINGSSYMTOM RÖD	VARNINGSSYMTOM ORANGE	VARNINGSSYMTOM GUL
Brännskada på ansikte/hals (ej bara droppstänk). Inhalationsskada. Högspänningsolycka.	Brännskada: ≥ 10% eller cirkumferent	Brännskada: ≤ 10% på barn ≤ 1 år

Förklaring och definition:

•En brännskada i ansikte och hals kan svullna mycket vilket då innebär en hotad lufvtäg om den är dermal eller djupare. Hit räknas inte droppstänk.

•Inhalation av varma eller frätande gaser kan ge svullnad i luftvägar. Inhalation av varm gas ska misstänkas vid ex svedda näshår, sotiga näsborrar. Om symtom som hosta eller heshet föreligger råder en klar misstanke om hot av luftvägen och dessa får röd varningssymtom.

•Högspänningsolycka: >1000 V (vanlig hushållsel är 230 V) kan ge livshotande skador som arytmier och inre brännskador. Blixtnedslag räknas som högspänningsolycka.

•Storlek: på barn är det lättast att använda barnets hand (handflata inkl fingrar) som referens. Handen är 1% av hudens storlek. Generellt har barnet proportionerligt större huvud och bål, och mindre armar och ben än en vuxen har så 9%-regeln fungerar inte fullt ut.

•Cirkumferent brännskada går runt en hel kroppsdel (exempelvis arm, ben, thorax). När brännskadan svullnar kan cirkulationen till delen, eller andningen om det är thorax, bli påverkad och man kan behöva avlasta akut med eskariotomi. Svullnaden kan ske fort och barnet får orange varningssymtom.

•**Åtgärd**: burnfree på brännskador (ej på kemiska skador). Smärtlindring. Förbered ev infart. Överväg traumalarm.

Sökorsak i Elvis
BRÄNNSK
ELOLYCKA

NYHETER i denna version

- Tabeller borttagna kring smärtlindring, hänvisas till ePed för detta.

A.8 Triage guide

Sökorsak **Rimlig tidsåtgång för triage**	**Kontroller/åtgärder** **SUBFI och vikt på alla barn**	**Begränsad anamnes**
Allergi **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Ge Aerius, ev få ordination på Betapred och Adrenalin IM. Emla för infart vid allvarlig reaktion	Debut, duration, vad, symtom (urtikaria, dyspné, buksymtom), tidigare reaktioner
Andningsbesvär **10 minuter**	Pox, puls, AF, KÅ, temp, vikt Inhalera VB	Debut, duration, andningsarbete, tidigare besvär
Bett av människa/ hund/katt/kanin **Bett av orm** **10 minuter**	Temp, vikt Inspektera och tvätta sår Pox, puls, AF, KÅ, BT, temp, vikt Emla för infart, immobilisera i högläge	När och var När och var, vilken orm
Bröstsmärta **10 minuter**	Pox, puls, AF, KÅ/BT, EKG, temp, vikt	Debut, duration, något som förvärrar/förbättrar, palpömhet
Bråck **5 minuter**	Vikt, ev smärtlindring	Debut, duration, kvar eller reponerat
Brännskada **10 minuter**	Vikt, temp, Burnfree på blöta kompresser (max 5% av kroppsytan), smärtlindring i maxdos, inte Ipren om >15% brännskada	Vad har barnet bränt sig på, vid skållskada, var det något i vätskan ex. mjölk i kaffet
Buksmärta **10 minuter**	Pox, puls, AF, KÅ/BT, Temp rektalt!, vikt Smärtlindring vb	Debut, duration, något som förvärrar/förbättrar, kräkning/avföring
Diabetes **10 minuter**	Pox, Puls, AF, KÅ/BT, temp, vikt Emla x3 för infart samt på magen	Symtomdebut
Diarré **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Vätskeersättning	Debut, duration, frekvens blod/slem?
Dålig viktuppgång **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Om ammas, amningsvikt, PMM	Äter, kissar, bajsar?
Elolycka **- hushållsel** **-högspänning** **10 minuter**	Puls, frekvens och rytm-askulteras	Hushållsel, medvetslöshet, kramper, vatten inblandat, brännskador?

	Pox, puls, frekvens och rytm, AF, KÅ/BT, temp, vikt, AT, brännskador, nedsatt känsel	
Sökorsak **Rimlig tidsåtgång för triage**	**Kontroller/åtgärder** **SUBFI och vikt på alla barn**	**Begränsad anamnes**
Extremitetsskada – misstänkt fraktur **10 minuter** - klämskada	Vikt, smärtlindring (viktigt för bra rtg och gips) Vikt, smärtlindring, Emla för finger-/tå-basblokad	Skademekanism, när Skademekanism, ytter/innerdörr, handtags eller gångjärnssida, när
Fallolycka/skallskada **10 minuter**	Pox, puls, AF, KÅ/BT, temp, pupiller, vikt Smärtlindring, Ge något att dricka eller äta	Medvetslös, amnesi, kräkning
Feber med oklart fokus **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Passa med mugg för urinsticka	Debut, duration, omgivningsfall
Främmande kropp i näsa/öra/öga **5 minuter**	Vikt	Vad, när, smärta
Främmade kropp i extremitet **5 minuter**	Temp, vikt, Emla lokalt	Vad, när, smärta
Gipskomplikation/omläggning **0 minuter**	Direkt till team	Skada, när gips/omläggning lades, hur länge ska det sitta
Gråtande barn **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Ev smärtlindring, PMM	Debut, duration, något som förvärrar/förbättrar
Haltande barn (hälta utan trauma) **10 minuter**	Temp, vikt	Debut, duration, infektion nyligen?
Hematemes/melena **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt	Debut, duration, frekvens, färskt eller gammalt blod, ammas
Hjärtsjukdomar **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt, EKG	Debut, duration, något som förvärrar/förbättrar
Huvudvärk **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Smärtlindring, ev Emla för infart	Debut, duration, karaktär
Intoxikation **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt, Ring giftinfo! Övriga kontroller/åtgärder enligt dem	Vad, mängd, när, syfte

Kramper **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt, EKG	Debut, duration, karaktär
Sökorsak **Rimlig tidsåtgång för triage**	**Kontroller/åtgärder** **SUBFI och vikt på alla barn**	**Begränsad anamnes**
Lokalinfektion **10 minuter**	Temp, vikt Febernedsättande vb, ev rita runt rodnad	Debut, duration Obs på snabbt tilltagande rodnad/gasbildning i huden
Misshandel **5 minuter**	Vikt, fort in på rum/lugnare del av akuten	INTE ta anamnes, räcker att berätta för EN person dvs läkaren
Nausea **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Vätskeersättning, ev febernedsättande, Ondansetron	Debut, duration, Kräkning utan diarré alltid träffa läkare!
Neurologiska besvär **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt	Debut, duration, tidigare besvär
Postoperativa besvär **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt	Debut, duration, när och vad opererad
Psykisk ohälsa -ätstörning **10 minuter** - övrigt	Pox, puls, AF, KÅ/BT, temp, (vikt), EKG, Emla för infart -	Debut, duration, kissat/bajsat Debut, duration
Shunt **10 minuter**	Pox, puls, AF, BT, temp, vikt Smärtlindring, ev Emla för infart	Debut, duration
Skrotal smärta **5 minuter**	Temp, vikt	Debut, duration, påverkad gång
Smärta rygg/nacke **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Smärtlindring	Debut, duration, karaktär, någon tänkbar orsak till smärtan
Trauma thorax/buk **10 minuter**	Pox, puls, AF, KÅ/BT, vikt Bukomfång, rita var du mätt Smärtlindring	Vad hände, när, palpera genom buk/thorax, något som förvärrar/förbättrar
Underlivsbesvär **10 minuter**	Temp, vikt, urinsticka Smärtlindring, Xylocain lokalt + Alvedon/Ipren	Debut, duration, möjlig orsak till besvären
Urinbesvär **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt, urinsticka Smärtlindring	Debut, duration, tidigare besvär

Ögon – infektion **10 minuter** **-trauma**	Pox, puls, AF, KÅ/BT, temp, vikt Om ethmoidit misstanke, Emla för infart Vikt, smärtlindring	Debut, duration, smärta, var är det svullet Rör ej om penetrerande skada Skölj om frätande ämnen
Sökorsak **Rimlig tidsåtgång för triage**	**Kontroller/åtgärder** **SUBFI och vikt på alla barn**	**Begränsad anamnes**
Öron **10 minuter**	Pox, puls, AF, KÅ/BT, temp, vikt Om mastoidit misstanke Emla för infart	Debut, duration, smärta, var är det svullet