SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

# Symptex: AI-Driven Simulated Patient History-Taking for Medical Training

Kevin-Florian Su

SCHOOL OF COMPUTATION, INFORMATION AND TECHNOLOGY — INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

# Symptex: AI-Driven Simulated Patient History-Taking for Medical Training

# Symptex: KI-gestützte Simulation der Patientenanamnese für die Medizinische Ausbildung

| | |
|---|---|
| Author: | Kevin-Florian Su |
| Examiner: | Prof. Dr. Pramod Bhatotia |
| Supervisor: | Dr. Nadine von Frankenberg |
| Submission Date: | August 25, 2025 |

I confirm that this master's thesis is my own work and I have documented all sources and material used.

Munich, August 25, 2025 Kevin-Florian Su

# Abstract

In German medical education, university courses and examinations often emphasize theoretical knowledge assessed through rigid written tests, while practical communication skills receive comparatively little attention. This imbalance creates a gap between textbook knowledge and the real-world clinical practice of physicians later on. However, one of the most important competencies among medical doctors is anamnesis (medical history-taking), a process that requires accurate information gathering and empathetic patient interaction.

To address the shortcomings in traditional anamnesis training, this thesis proposes Symptex, an AI-driven chatbot designed to simulate authentic German-speaking virtual patients and provide personalized performance feedback. Integrated into the ILuVI learning framework, Symptex employs large language models (LLMs) to create an interactive, low-stakes environment in which medical students can practice communication skills. The system was developed through a systematic design process comprising architectural modeling, model selection, and optimization of a suitable open-source LLM. The authenticity of the virtual patient simulation, covering aspects such as German language fluency, conversational coherence, and medically plausible responses, was further refined through prompt engineering, few-shot examples, and expert input.

Symptex was evaluated in a two-part mixed-methods pilot study. In the subjective evaluation, five medical students conducted a medical history-taking session with Symptex's Alzheimer's patient persona and assessed the system from a learner's perspective. In the objective evaluation, three graduate physicians evaluated the simulation's performance from an expert standpoint with respect to medical accuracy and consistency. Questionnaires captured user perceptions and acceptance of Symptex as a learning tool, while also focusing on the authenticity of the simulation and the value of its performance feedback. Results from both groups indicate strong agreement on Symptex's conversational authenticity and educational potential. Overall, these findings suggest that Symptex can complement traditional anamnesis training by offering a scalable, interactive, and pedagogically meaningful language tool.

# Contents

# 1 Introduction

To start off this thesis, this chapter sets the context and research focus of this work. It begins by outlining the problem statement and motivation, highlighting the current shortcomings of anamnesis (medical history-taking) training in German medical education and the potential role of LLM-based virtual patients (VPs) in addressing these gaps. Building on this, the chapter presents the research questions that guide the subsequent contributions of the thesis. Finally, it concludes with an overview of the thesis structure.

## 1.1 Problem Statement and Motivation

While medical technology has rapidly evolved during the 21st century, the teaching methods and assessment strategies in medical education have largely remained unchanged. In Germany, the predominant examination format for medical students continues to be multiple-choice questions, which primarily assess factual knowledge. However, these formats often fail to evaluate essential communicative competencies—such as those required during the medical history-taking process, a procedure recognized as one of the most critical steps during a medical consultation [BH06; Kei+15]. Existing training methods, such as hiring actors to portray patients, are both costly and limited in scalability [Bos+15; Hub+00; Ste+06]. Anecdotal reports from German medical students stress that only a small number of students can actively participate, while the majority are relegated to passive observation. This disconnect raises concerns about the adequacy of current healthcare education practices, as students experience a significant gap between the theoretical knowledge they are tested on and the actual application of that knowledge in more complex, real-world scenarios.

Recent advances in the field of Natural Language Processing (NLP) have greatly improved the quality of interactions with artificial intelligence (AI) systems. The introduction of the transformer architecture has set a new standard for model training, enabling faster and more efficient learning compared to previous methodologies [Vas17]. Furthermore, model performance can be improved by increasing model size and computing resources [Kap+20], leading to the emergence of large language models (LLMs). As a result, there has been an increase in the number and quality of models that can generate coherent human-like responses and engage in conversations with low latency. This trend has sparked a surge in research aimed at evaluating the potential applications, benefits, and limitations of LLMs across various domains. One promising area is the enhancement of educational tools, which can, for instance, automate training and provide personalized feedback [BFK24; Neu+24; Hol+24]. When implemented responsibly with appropriate safeguards against potential risks, LLMs hold the potential to improve educational efficiency by promoting greater interactivity and learner engagement in medical training environments [LUR24; Lee24; Saf+23].

In light of this context, this thesis implements the LLM-powered prototype chatbot Symptex, which allows students to practice their communication skills by conversing with a simulated VP in a medical history-taking setting. Through incorporating challenges such as age-related confusion or memory loss—common issues that can complicate the physician's task of gathering crucial patient history—students experience a more authentic and engaging clinical communication practice. The chatbot further features personalized follow-up feedback on their communicative and content performance to reinforce learning outcomes.

Symptex is integrated into the existing Interdisciplinary Longitudinal VIrtual Patient Management (ILuVI) framework, an ongoing project designed to complement medical lectures of the TUM Klinikum rechts der Isar (TUM MRI) through a web application for instructors and a mobile application for students. Within the ILuVI mobile application, students can engage with clinical case studies created by instructors in a virtual trial-and-error sandbox environment. Each case study features VPs with various symptoms, allowing students to practice their diagnostic reasoning and patient management skills. Following the structure of real-world medical consultations [Ham+75; BS12], students first take the patient's medical history, then request simulated diagnostic tests, and finally submit their diagnosis, while adhering to constraints such as virtual time or budget limits.

In its current state, ILuVI's medical history-taking phase is limited to the selection of predefined questions with fixed, pre-written responses. This static format fails to capture the unpredictability and emotional nuances of real clinical encounters, thereby limiting realism and opportunities to practice adaptive communication skills. To address these shortcomings, this work documents the introduction of Symptex into the ILuVI framework, thereby replacing the static question-answer format with a more interactive, pedagogically valuable tool to enhance its VP concept.

## 1.2 Research Questions

Within the scope of this thesis, the following research questions guide the implementation of Symptex. By addressing these questions, this work seeks to assess the overall usability and pedagogical value of LLMs in educational role-playing settings, as well as to derive best practices for their integration into existing digital learning environments such as ILuVI.

**RQ 1: How should the architecture for an LLM-powered chatbot that simulates doctor-patient conversations for medical history-taking be designed?**
Based on the requirements of the targeted use case, LLMs can be integrated into a range of architectural paradigms, such as retrieval-augmented generation (RAG), tool-calling, and agent-based frameworks. This question analyzes these established paradigms to determine which architectural design best meets the functional and non-functional requirements of Symptex (see Chapter 4). The architectural design is outlined in Chapter 5 (System Design) and Chapter 6 (Object Design).

**RQ 2: Which open-source LLMs are most suited for VP simulation and performance evaluation, for instance, with regard to linguistic capabilities?**
This question examines which open-source LLMs are most suitable for both core use cases of this work, namely, simulating a VP during anamnesis, and evaluating the student's history-taking performance. The selection process considers different performance aspects such as patient language authenticity, conversational consistency, and inference latency, and is documented in Section 6.2 of the Object Design chapter.

**RQ 3: What type of LLM optimization is required to improve Symptex's ability to deliver on VP simulation and performance evaluation?**
This question focuses on prompt engineering techniques, few-shot learning examples, and parameter adjustments (e.g., temperature) to ensure a balance between repeatable assessment conditions and the natural variability of real patient dialogue. The applied optimization process is detailed in Section 6.3 of the Object Design chapter.

**RQ 4: How well can the chatbot simulate patient-specific conditions during medical history-taking?**
This question specifically investigates Symptex's ability to maintain a coherent patient persona over a multi-turn conversation, adapt responses to different questioning styles, and exhibit believable emotional and cognitive traits relevant to the patient profile (e.g., memory gaps in dementia). The evaluation, presented in Chapter 7, considers both objective system evaluations from experts and subjective user feedback.

## 1.3 Outline

The following sections describe how this thesis contributes to the research questions defined above.

**Chapter 2: Background** introduces the key concepts and terminologies necessary to understand the context of this work, including the medical history-taking process and the fundamentals of LLMs. It also presents technical details about the architecture of the ILuVI framework, into which Symptex is integrated.

**Chapter 3: Related Work** reviews similar literature relevant to this thesis. It begins by covering established approaches to teaching anamnesis in medical education and then examines how the institutionalization of generative AI has influenced medical education practices. Subsequently, it analyzes recent studies that employ LLMs in the context of medical education. The chapter concludes with an overview of established evaluation metrics used to assess the pedagogical effectiveness of LLM-driven chatbots, thereby providing references for the evaluation conducted later in this work.

**Chapter 4: Requirements Analysis** compares the current medical history-taking solution of the ILuVI framework with the proposed Symptex implementation and further specifies the functional and non-functional requirements of Symptex. It also introduces system models that illustrate user workflows and the overall logic of the proposed system.

**Chapter 5: System Design** begins by specifying the overall design goals of the implementation. It then outlines the chosen architectural approach, how the subsystems interact with each other, and the integration strategy with the ILuVI framework.

**Chapter 6: Object Design** presents the technical details of the Symptex implementation, followed by a documentation of the model selection and prompt engineering processes.

**Chapter 7: Evaluation** begins by outlining the overall evaluation methodology of Symptex and further summarizes the results of the evaluation, assessing both its technical performance and its perceived educational value. Finally, the findings of the results are discussed.

**Chapter 8: Summary** concludes this work by summarizing the contributions throughout this thesis, discussing limitations of the implemented prototype, and offering potential directions for future research.

# 2 Background

This chapter introduces relevant medical and technical foundations on which this thesis is built. It begins by situating the process of *Medical History-Taking* within the broader diagnostic process, as this phase serves as the foundation for the simulated doctor-patient interactions that are targeted by Symptex. The technical realization of this chatbot relies on the technology of *Large Language Models* (LLMs), which is explained in Section 2.2, providing the background knowledge necessary to understand the system developed in this thesis. Section 2.3 describes an overview of the *ILuVI* framework, into which the chatbot is integrated.

## 2.1 Medical History-Taking

The purpose of Symptex is to allow medical students to practice the process of *medical history-taking*, also referred to as *anamnesis*. This process constitutes the first of three key steps that physicians typically follow during a medical consultation to arrive at a diagnosis [Ham+75; BS12], as illustrated in Figure 2.1.

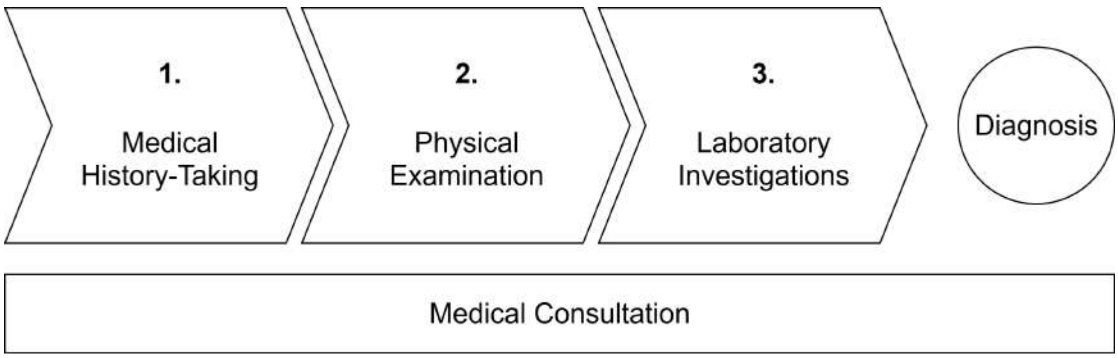


Figure 2.1: A Typical Medical Consultation Process

Depending on the clinical context, this diagnostic process can take the form of either a comprehensive or a focused assessment. A comprehensive assessment is typically conducted for new patients in outpatient or hospital settings and involves a thorough history-taking and complete physical examination. In contrast, a focused assessment is more appropriate for returning patients during routine follow-ups or urgent visits, concentrating on specific concerns or symptoms relevant to the presenting problem [BS12].

The three clinical steps are outlined as follows:

1. **Medical History-Taking**
   This verbal task aims to elicit key information about the patient through direct doctor-patient interaction [BH06]. According to Bates' Guide to Physical Examination and History Taking, a comprehensive assessment of adults includes the following seven components [BS12]:
   a) Identifying Data: Basic demographic details such as age, gender, occupation, together with the source of the history (e.g., patient or relative).
   b) Reliability: An assessment of how trustworthy the patient's responses are, depending on their memory and emotional state.
   c) Chief Complaint(s): The main symptom(s) why the patient is seeking care.
   d) Present Illness: A detailed report of the symptoms mentioned in the chief complaint, including their onset, development, and the patient's concerns.
   e) Past History: Information about previous illnesses, surgeries, and other health-related events.
   f) Family History: Health information about close relatives, including health conditions or cause of death—particularly conditions that may have a genetic component.
   g) Personal and Social History: Insights into the patient's daily life, such as living situation, and lifestyle habits.
   h) Review of Systems: A systematic checklist of symptoms organized by major body systems, used to identify issues not previously discussed.
2. **Physical Examination**
   The assessment of the patient's observable anatomic features through clinical techniques, such as inspection, palpation, percussion, and auscultation, to identify signs of underlying health conditions [CK90].
3. **Laboratory Investigations**
   The process of selecting, ordering, and interpreting diagnostic tests from the laboratory [CK90].

The importance of medical history-taking is well recognized within the scientific community. It is considered one of the most fundamental and recurring practices for physicians, as it lays the groundwork for a productive physician-patient relationship. The effects of this interaction extend throughout the entire consultation and can significantly influence the accuracy of the final diagnosis as well as the following treatment success [BH06; Kei+15]. Studies also have shown that the majority of patient diagnoses can be derived solely from the information obtained during medical history-taking, emphasizing its essential role in the medical consultation [Pet+92; Ham+75]. The implementation presented in this work, therefore, aims to familiarize medical students with this important diagnostic step early on to systematically strengthen their communication skills in conducting anamnesis.

## 2.2 Large Language Models

To achieve this, Symptex leverages recent advancements in natural language processing: *Large Language Models* (LLMs). LLMs are complex models designed to process and understand human language in the form of textual input, referred to as a *prompt*, and subsequently generate coherent, human-like responses in a process known as *inference* [Bro+20; Zha+23b; Tou+23]. This capability makes them particularly useful for applications that require interactive, context-aware dialogue.

### 2.2.1 Foundations of LLMs

The overall trend and increase in LLM research attention and performance can be attributed to the widespread adoption of the *Transformer* architecture introduced by Vaswani et al. [Vas17]. This architecture employs a self-attention mechanism that overcomes the sequential processing limitations of older models, particularly those based on recurrent neural networks (RNNs). By enabling parallelized processing during training, transformers significantly reduce training time—often by three times or more—compared to traditional RNN-based models. This improvement also leads to notably better performance, especially on long sequences, as transformers process entire input sequences simultaneously instead of sequentially. [Vas17].

In addition, Kaplan et al. observed that model performance follows predictable scaling laws as a function of model size, dataset size, and available compute resources [Kap+20]. They found that increasing the number of parameters results in log-linear improvements in performance, as long as the dataset and computational resources are scaled accordingly. The discovery of this relationship has led to the development of state-of-the-art language models with an ever-increasing number of parameters, which now range from billions to trillions, hence the term "large language models."

Although nowadays most LLMs primarily generate text-based responses, the introduction of *multimodal* LLMs has extended their capabilities beyond just text to incorporate images, audio, and video as both input and output, thereby facilitating more versatile forms of human interaction.

### 2.2.2 Lifecycle of an LLM

From a developer's perspective, the lifecycle of an LLM involves two main phases: training and deployment.

#### Training

Although the transformer architecture offers significant efficiency improvements, achieving strong model performance still requires extensive training, which demands considerable computational resources and energy. As a result, this training process is typically only carried out by organizations or companies with access to specialized high-performance infrastructure.

#### Deployment

Once trained, LLMs are typically published as *pre-trained* models. Because they also require significant computational power during inference, they need to be deployed on dedicated servers or cloud-based platforms with Graphical Processing Units (GPUs). Users can interact with the model only after this deployment is complete.

Pre-trained LLMs are usually made available in one of the following two categories:

- **Open-source LLMs**
  Publicly accessible models intended to support collaboration, transparency, and further research. They can be freely hosted and used by anyone with sufficient computational resources.
- **Closed-source LLMs**
  Proprietary models developed and distributed by commercial entities. Access to these models is often restricted and monetized, with usage usually only provided through Application Programming Interface (API)s or web platforms.

### 2.2.3 Applications and Limitations

Overall, owing to their strengths in natural language generation and contextual understanding, LLMs have demonstrated significant potential in a variety of tasks, as discussed by Yang et al. [Yan+24]:

- **Natural language understanding (NLU)**
  They can interpret a wide variety of texts, even if their contents differ from their training data.

- **Natural language generation (NLG)**
  LLMs are able to generate detailed and natural responses suitable for summarization, dialogue, or creative writing.

- **Knowledge-intensive tasks**
  LLMs perform strongly in question answering settings, where they are asked to recall factual information that is based on their training data.

- **Reasoning tasks**
  As LLMs scale, they show increased performance with regard to more complex logical tasks such as arithmetic reasoning and commonsense reasoning.

These strengths have been leveraged in various practical use cases, such as in intelligent tutoring systems [BFK24], virtual writing assistants [Bro+20], legal document analysis [Cha+20], and code generation platforms [Che+21], among others.

Despite their advanced capabilities, LLMs do, however, also face shortcomings. Their non-deterministic behavior can lead to the generation of plausible-sounding but factually incorrect or contextually irrelevant information. This phenomenon is commonly referred to as *hallucination* [Zha+23a]. In addition, LLMs struggle to keep responses consistent over time, especially as conversations progress beyond their training data. For applications such as patient simulation, these shortcomings can pose particular challenges. Inconsistencies or factual errors that deviate from the provided patient history may compromise the perceived authenticity of the virtual patient and the pedagogical value of an exercise, which relies on reliable repeatability.

In this context, Symptex seeks to leverage the strength of LLMs in mimicking human-like communication [Yan+24] for the simulation of virtual patients. At the same time, this work aims to address and mitigate the known limitations of LLMs, which may negatively impact the learning experience for medical students.

## 2.3 The ILuVI Framework

Symptex is embedded within the *Interdisciplinary Longitudinal VIrtual Patient Management* (ILuVI) framework. The following section outlines ILuVI's educational and technical foundations in order to provide the necessary context for the integration and evaluation of Symptex presented in this work.

ILuVI is a learning framework currently under development that adopts a flipped classroom approach, encouraging students to interact with clinical content before in-person lecture sessions. The system comprises three modules: the *ILuVI Mobile Application* for students, the *ILuVI Instructor Web Application* for lecturers, and the *ILuVI Server*, which manages server-side orchestration logic and data exchange.

### 2.3.1 ILuVI Mobile Application

The mobile *ILuVI Mobile Application*, developed in Flutter for cross-platform support on iOS and Android, consists of the following two primary views.

**The Library**
This view offers students access to teaching materials uploaded by instructors, allowing them to review content in preparation for lectures or case studies. This further helps them tackle the latter more effectively after gaining relevant background knowledge.

**The Doctor's Office**
This central component of the application hosts the clinical case study exercise, with the corresponding workflow described in the following.

Students start each case study by reviewing the available general information about the patient on the first view, such as their name, birthdate, and physique (see Fig. 2.2). They then select predefined anamnesis categories to receive and review corresponding information about the patient's medical history on the second view. As mentioned in Chapter 1, the issues with this static process are addressed with this work by employing Symptex as a more interactive, realistic, and pedagogically valuable approach.

Once they have elicited the necessary information and identified potential underlying health issues, they can request simulated diagnostic tests on the second view as well (see Fig. 2.3). When placing these orders, students must account for constraints set by the instructor during the case creation process, such as the virtual time available and the financial costs associated with the diagnostic procedures. On the third and final view, they assign an ICD[1] code to finalize their diagnosis and submit it along with information regarding the overall progress of their case study for evaluation, which includes total time and budget usage as well as the laboratory procedures ordered.

After the allotted time for the case study exercise has passed, submissions are automatically evaluated, and students receive feedback via the ILuVI Mobile Application. Over the course of the entire workflow, the mobile application consistently maintains

[1] `https://www.who.int/standards/classifications/classification-of-diseases`, accessed on 16 August 2025

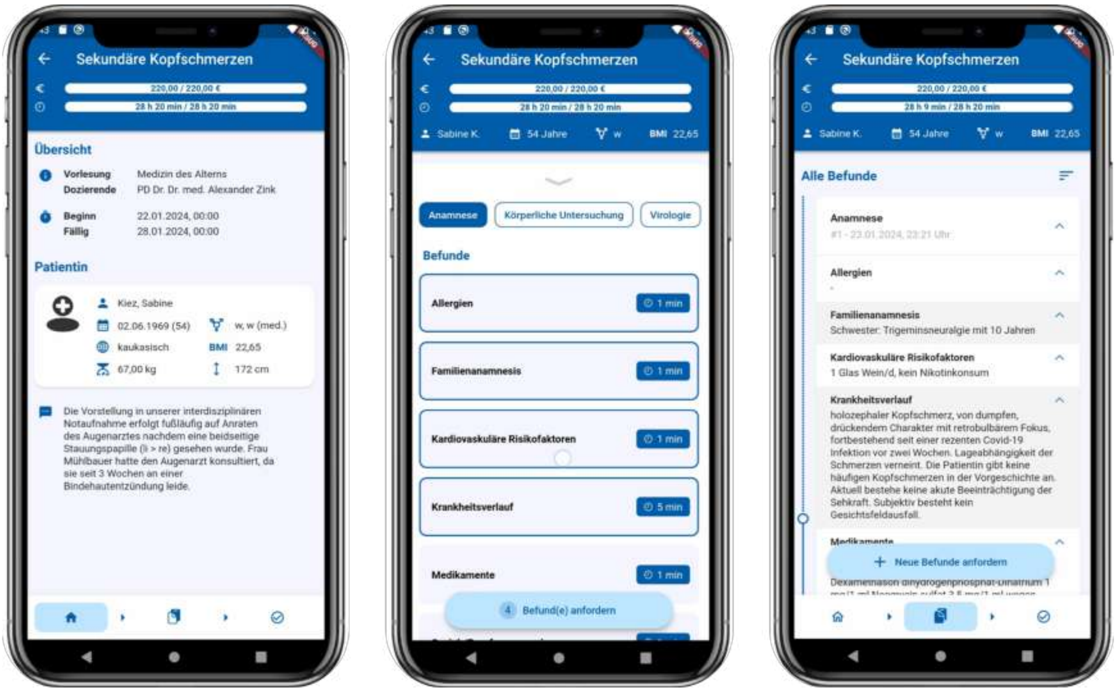


Figure 2.2: ILuVI Mobile Application: Case Study Workflow Part 1

an internet connection with the ILuVI Server to fetch content and persist individual progress.

### 2.3.2 ILuVI Instructor Web Application

This web-based application allows instructors to create lectures and upload corresponding teaching material displayed in the Library part of the student's ILuVI Mobile Application. In addition, they may design highly customizable clinical case studies for the students (see Fig. 2.4), including the attributes of the virtual patient, available diagnostic procedures to order, and the allowed time/budget constraints. All changes and newly created content are stored on the ILuVI Server, ensuring consistent synchronization across the system.

### 2.3.3 ILuVI Server

With both the ILuVI mobile and web applications acting as clients, the ILuVI Server functions as the central orchestrating system within the framework. It processes incoming Hypertext Transfer Protocol (HTTP) requests from these clients and ensures

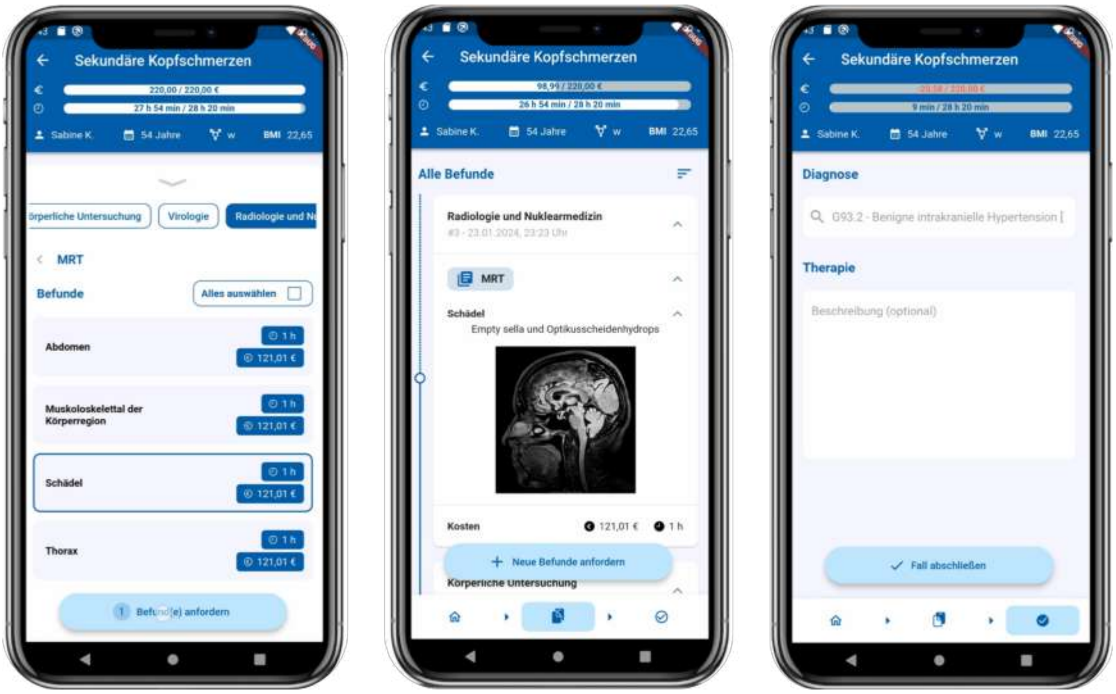


Figure 2.3: ILuVI Mobile Application: Case Study Workflow Part 2

the storage and retrieval of data created by instructors (e.g., lecture materials, case studies) and students (e.g., individual case study progress) in a PostgreSQL database. Aside from this API logic implemented in Golang, the server also completes tasks such as the automated assessment of completed case studies and the calculation of individual performance scores.

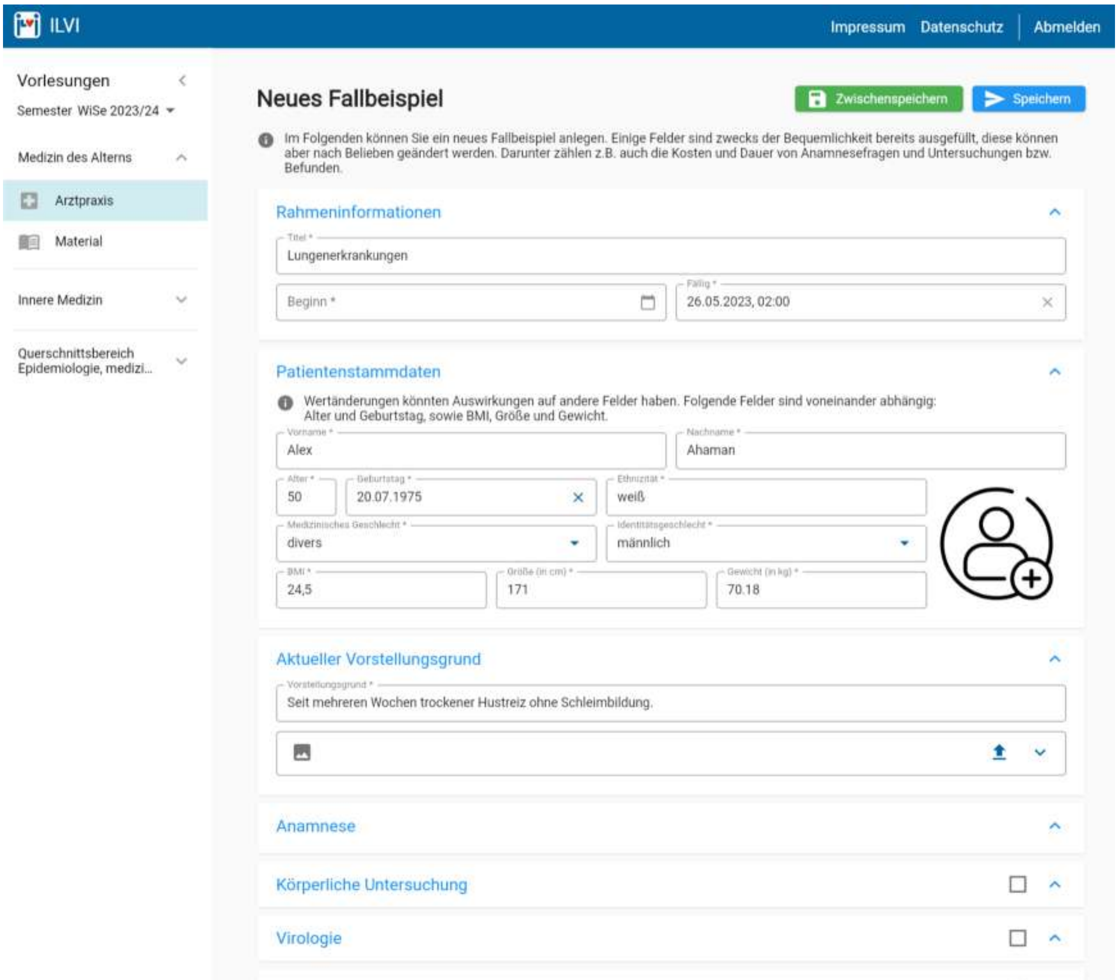


Figure 2.4: ILuVI Web Application: Creating a Case

# 3 Related Work

This chapter gives an overview of the scientific discourse surrounding educational concepts for teaching medical history-taking, as well as various implementations to provide a foundation for contextual comparison. Section 3.1 distinguishes between communication models, which serve as a theoretical framework, and their practical application, such as standardized and virtual patients. Subsequently, Section 3.2 discusses the impact of the technological advancements in generative AI on medical education. Finally, commonly used evaluation metrics for assessing the pedagogical effectiveness of AI-based educational tools are explored in Section 3.3, followed by a brief outline of how this thesis situates itself within the existing body of research in Section 3.4.

## 3.1 Traditional Anamnesis Educational Approaches

Effective communication is a core component of a physician's skill set, particularly in the context of doctor-patient interactions during the anamnesis phase. When executed well, it not only increases diagnostic accuracy but also improves patient satisfaction, adherence to the assigned treatment, and overall well-being for both doctor and patient [MP02; ZD09; Ong+95]. Despite its acknowledged importance in international medical literature, communication remains a complex and challenging skill to teach. It is influenced by varying factors, including individual physician treatment styles as well as changing patient characteristics and clinical settings. The following sections outline current educational approaches developed to address the challenges of teaching students about this multifaceted skill alongside the medical history-taking process.

### 3.1.1 Clinical Communication Models

Clinical communication models serve as the theoretical groundwork designed to teach and assess communication during medical history-taking in a structured and educational manner.

Among the most comprehensive of these is the *Calgary-Cambridge Guide*, which has been adopted in several countries as a standard for communication training in healthcare education [KS96; Bur+14; Von+08]. Developed by Suzanne Kurtz and

Jonathan Silverman in 1996, the model emerged in response to a lack of a consistent teaching framework that would allow learners to apply communication skills in real-world clinical practice [KS96]. The Calgary-Cambridge Guide comprises a structured collection of communication skills, organized into five core tasks that define the flow of a typical clinical consultation between the doctor and the patient (see Fig. 3.1) [KS96]:

1. Initiating the session
2. Gathering information
3. Building relationship/facilitating patient's involvement
4. Explaining and planning
5. Closing the session

Communication soft skills embedded within these stages include, for example, the use of open questions, the summarization of patient input, and the active participation of the patient in the decision-making process. These techniques encourage a patient-centered approach in clinical consultations and contribute to more productive medical interviews, facilitating the collection of more relevant clinical information and a strengthened therapeutic relationship between physician and patient.

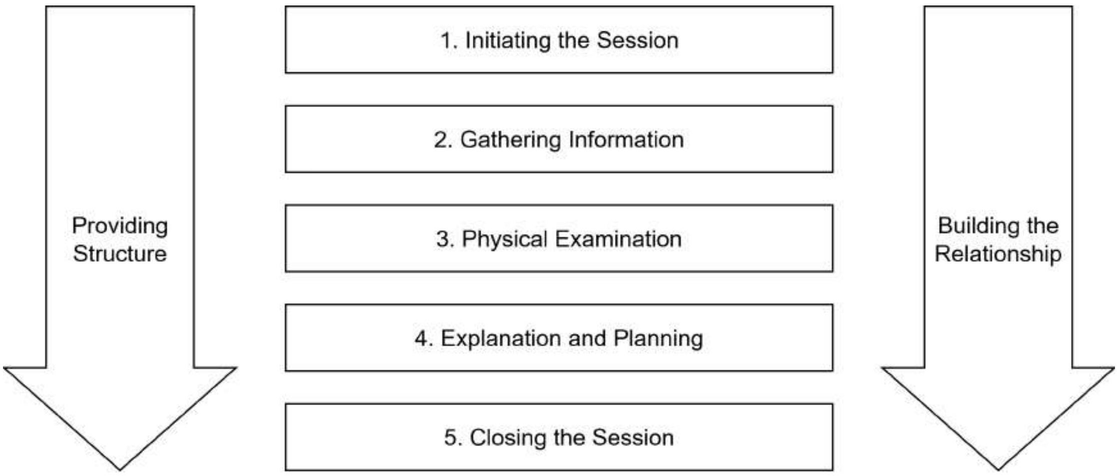


Figure 3.1: Basic Visualization of the Calgary-Cambridge Guide, adapted from Kurtz and Silverman [KS96]

By providing a clear and systematic framework, the guide enables both instructors and learners to quickly identify and refer to specific relevant communication behaviors during feedback sessions after medical interviews.

In 2003, the model was enhanced to address confusion between the *process* of communication and the *content* to be obtained, such as the medical history itself [Kur+03]. Learners were often focused on gathering content—such as symptoms or medical history—at the expense of following the structured communication process outlined in the guide. This led to a tendency to revert to closed questioning and a less patient-centered approach. To address this issue, Kurtz et al. revised the guide to more clearly differentiate between the *process* of communication and the *content* being gathered, including the following key elements [Kur+03]:

- The clearer sequence of communication events
- Addition of the patient's perspective to ensure the process is patient-centered
- Consideration of treatment alternatives considered by the physician
- Documentation of what the patient has been told
- A negotiated plan of action, aligning both physician and patient

Another established communication framework is the *Four Habits Model*, developed by Richard M. Frankel and Terry Stein in 1999 [FS99]. Designed for broad applicability in clinical practice, this model emphasizes relational and behavioral aspects in interactions between doctor and patient and is structured around four core "habits" [FS99]:

1. **Invest in the beginning**
   Build rapport, clarify the patient's concerns, and collaboratively set an agenda.
2. **Elicit the patient's perspective**
   Explore the patient's ideas, concerns, and understanding of their condition.
3. **Demonstrate empathy**
   Be aware of your own reactions and respond to the patient's emotions with both verbal and non-verbal empathy.
4. **Invest in the end**
   Deliver diagnostic and treatment information clearly, involve the patient in decision-making, and confirm mutual understanding in follow-up steps.

By structuring communication around consistent behavioral patterns, the Four Habits Model highlights the significance of doctor-patient relationships in care as well as the importance of patient-centered communication. Beyond its development and implementation in the United States [SFK05], the model has also been adopted in international clinical training programs, including intermediate care services in Norway [Kvæ+24; Gul+08]. Evaluations in these settings have reported high levels of physician acceptance and improved efficiency in eliciting relevant patient information during clinical encounters.

In summary, both the Calgary-Cambridge Guide and the Four Habits Model highlight the significance of integrating communication skills into clinical workflows, as well as the necessity of patient-centered medical care. These frameworks emphasize that learners should focus not just on identifying the information to elicit, but also on how to structure conversations and use soft communication skills to build trust between physicians and patients.

### 3.1.2 Practicing Communication Skills

To apply the aforementioned frameworks in educational settings with the goal of developing communication skills, several methods are available. Instructors seeking a more interactive approach, as opposed to passive observation methods like video analysis, typically choose from options which can be placed in one of the two main categories: *Standardized Patients* [Bar93; Hub+00] and the more recent alternative, *Virtual Patients* [Kon+15; Ber+16; Ste+06].

#### The Standardized Patient

The term *Standardized Patient* (SP) was initially introduced by Barrows as an umbrella concept that includes trained professionals, e.g., actors or instructors, who impersonate patients, as well as actual patients invited to present their health problems to the medical student for educational purposes [Bar93; Hub+00]. In this setting, students typically engage in multiple medical history-taking simulations with the SP, followed by iterative feedback from instructors aimed at improving their communication and clinical reasoning skills.

From the student's perspective, working with SPs offers several advantages: It provides a less intimidating environment compared to real-world clinical settings, as it allows for pausing or restarting encounters as needed. In addition, the immediate instructor feedback supports active learning. These attributes enable students to experiment with different communication strategies and reflect on their performance.

From an instructional standpoint, SPs provide risk-free patient experiences that are repeatable and consistent to a degree, in contrast to the variability of real patient encounters in clinical wards. The faculty can further determine specific learning objectives in advance, ensuring that all students are exposed to comparable scenarios. This controlled setting, therefore, facilitates both structured evaluation and targeted skill development [Bar93].

Nonetheless, the use of SPs also presents several challenges: The aspect of ensuring consistency in role portrayals, maintaining alignment in evaluation criteria, and reproducing scenarios across sessions requires significant time and effort. If not sufficiently addressed, these factors can lead to unwanted variability in the learning experience, which, in turn, results in unequal learning among medical students. In addition, logistical demands such as actor training, scheduling, and the overall high implementation cost limit the scalability of this approach [Bos+15; Hub+00; Ste+06]. The resource-intensive nature of SPs, coupled with technological advances, has led to increasing attention to the following alternative for communication practice.

**The Virtual Patient**

The *Virtual Patient* (VP) is broadly understood as an interactive, computer-based patient scenario to support the teaching and assessment of a variety of skills—ranging from general procedural clinical competencies to patient communication skills—within a safe, repeatable, and controlled learning environment [Kon+15; Ber+16; Ste+06]. The form of a VP can differ depending on its underlying technology, which may include systems using pre-recorded video responses or more advanced platforms that utilize natural language processing and generative AI for real-time interactions.

Compared to SPs, VPs offer several practical advantages. Their digital nature allows for broader accessibility, simplified scenario repetition, and reduced logistical demands — thereby providing a scalable alternative [Pla+22; Ber+16]. With regards to educational value, early studies demonstrated that VPs can be just as effective as SPs in improving clinical skills [Tri+06]. Furthermore, VPs are consistently associated with improved learning outcomes when compared to other groups without access to VPs, and have been shown to enhance clinical reasoning when used alongside traditional teaching methods [CET10; Pla+22]. Students appreciate the flexibility given by VPs, allowing them to engage with patients at their own pace [CET10] [Ber+16]. Similar to SPs, the non-threatening environment is often complemented by immediate feedback, which students report as valuable to their learning process [CET10].

However, some limitations remain: Cook and Triola observed that VP implementations usually come with substantial upfront development time and costs, as well as continuous maintenance expenses [CT09]. They also note that learners may display reduced empathy or interpersonal warmth when interacting with VPs, possibly due to the absence of emotional feedback during these interactions. Studies further highlight the importance of creating VPs in collaboration with instructors. Without such alignment, there is a risk of creating a disconnect between the available VP programs and the instructors who are meant to use them. This disconnect can cause instructors to struggle with effectively integrating VPs into their curricula, or, in some cases, to refrain from using them altogether [Ber+16; CT09].

An early concept implementation of a VP by Fleetwood et al., named MedEthEx Online, showed promising results [Fle+00]: The researchers designed a web interface focused on medical ethics, featuring four VP cases, each providing an overview of the patient's history. In this scenario, the learner assumes the role of a physician and initially interacts with the VP in a chatroom setting powered by early natural language processing technology. After eliciting the necessary information from the VP, the learner can consult virtual professionals, such as medical ethicists, for additional insights. Finally, the user submits an ethical decision and receives personalized feedback. MedEthEx Online was evaluated in conjunction with SPs to assess its effectiveness as a preparatory tool. Students who engaged with VPs prior to the SP encounters demonstrated strong ethical reasoning and communication skills that were comparable — or in some areas superior — to those of students who did not get to practice with a VP. The students further reported feeling well-prepared to interact with the SPs [Fle+00].

Another concept implemented by Stevens et al. explored the idea of using VPs to enhance communication skills as well as medical history-taking abilities of medical students in general, either as an alternative to or as a supplement for SPs [Ste+06]. In this approach, students were given patient information on a tablet, and then proceeded to interact with the projected life-sized VP using voice commands. The VP utilized voice recognition technology to process the commands and responded with scripted replies. Throughout the process, students received feedback and guidance from a virtual instructor. The learners rated the VP experience highly, despite occasional voice recognition failures and the absence of open-ended communication due to the script-based recognition [Ste+06]. They valued the safe, non-threatening practice environment as well as the overall effectiveness of the tool in preparing them for real patient encounters.

Campillos-Llanos et al. employed a comprehensive frame- and rule-based approach that integrates a termino-ontological model into a dialogue system designed to function as a VP avatar for medical history-taking practice [Cam+20]. After initially preprocessing the user input, the model first seeks to extract medical terms and classify the type of question using Natural Language Understanding (NLU) techniques, which include rule-based matching and entity linking. Subsequently, the extracted terms are mapped to standardized terms obtained from the Unified Medical Language System (UMLS), a large multilingual database comprising medical terminologies. Using the identified standardized terms, corresponding patient records are then accessed. Finally, the model populates a response template and formulates a reply. The authors highlight the model's advantage in precision and vocabulary, as it consistently relies on patient records, thereby avoiding hallucinations. Overall, their system yielded positive feedback from the users regarding its correctness and speed. However, learners also reported limitations in the response accuracy of follow-up or deviating questions, which may stem from the model's rigid design and its resulting inability to handle unexpected inquiries [Cam+20].

In conclusion, early VP approaches have demonstrated considerable promise, with students overall expressing a positive and accepting attitude towards this alternative teaching method, alongside encouraging findings regarding its educational value. However, high development costs, limited emotional authenticity, and misalignment between VP design and instructor needs can hinder their effectiveness. Notably, these implementations relied on scripted or rule-based systems, which often lack the adaptive conversational capabilities of modern LLM-powered solutions. Addressing these limitations through generative AI could help VPs to achieve new pedagogical potential as a scalable and engaging complement to traditional teaching methods.

## 3.2 LLMs in Medical Education

With the institutionalization of generative AI, research into the potential of chatbots has revealed various opportunities within the medical education domain.

Kung et al., for instance, demonstrated that GPT-3.5, a closed-source LLM developed by OpenAI, is capable of passing all written components of the United States Medical Licensing Exam (USMLE) without human assistance [Kun+23]. While its performance was close to the minimum passing standard of the UMSLE, the newer GPT-4 model consistently answered 80% of the exam questions correctly, emphasizing the growing medical competency of LLMs [Nor+23]. Comparable results have been observed across specialties such as surgery and cardiology, suggesting that even general-purpose LLMs like ChatGPT possess comprehensive medical knowledge [LUR24].

Building on these findings, several literature reviews have shown how the substantial medical knowledge of LLMs, combined with their capacity to follow structured instructions, enables multiple educational use cases (see Fig. 3.2):

- **Medical Knowledge Base**
  Serving as a continuously accessible database of medical information to answer information-seeking queries [Kun+23].

- **Personalized Learning**
  Supporting self-directed learning by delivering explanations at different complexity levels and aiding in activities such as exam preparation [LUR24; Xie+24; Sal23].

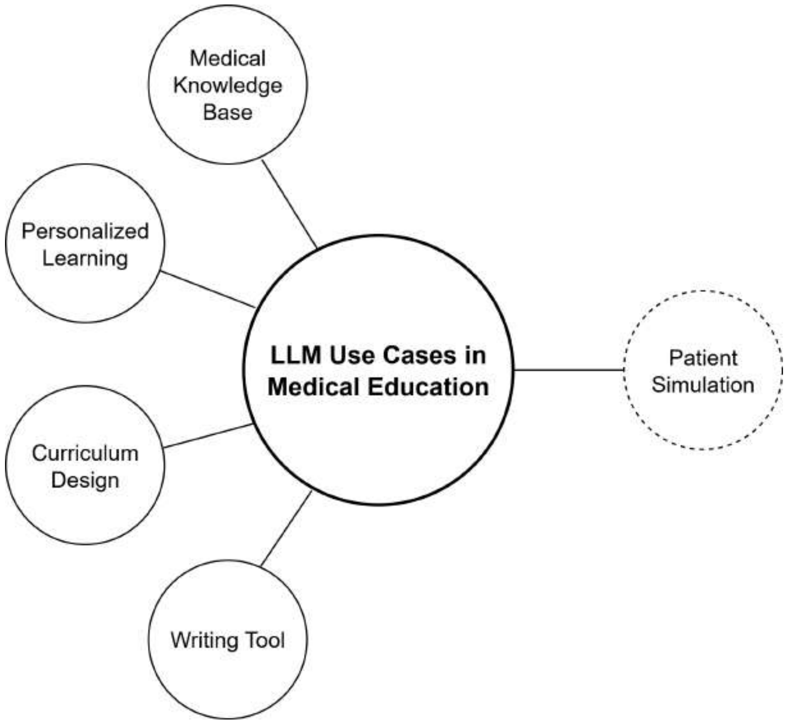


Figure 3.2: Use Cases of LLMs in Medical Education

- **Curriculum Design**
  Assisting instructors in drafting and refining lesson plans, generating examination material, and developing educational resources [LUR24; Lee24; Sal23].

- **Writing Tool**
  Helping students and instructors to ensure grammatical accuracy, refine writing style, and format documents [LUR24].

However, the majority of existing research on generative AI in healthcare education focuses on leveraging the "helpful assistant" capabilities of LLMs. By contrast, significantly fewer studies have examined their application in *Patient Simulation*—particularly the integration of human-like conversational abilities and the role-playing strengths of LLMs for VP simulation scenarios.

One of the few studies aligned with the VP concept was conducted by Benfatah et al. in the nursing sector. This study involved the use of ChatGPT (a closed-source LLM by OpenAI, version not specified) to simulate VPs experiencing respiratory distress [Ben+24a]. Participants of the study were asked to interact with the VP as if they were caring for a real patient, which allowed researchers to assess the viability of ChatGPT as a tool for healthcare simulation. Their study resulted in positive feedback from the participants regarding the accessibility and usefulness of the chatbot. Further, the competencies displayed during the use of the VP were found to correlate with the participants' overall grades.

Another research conducted by Holderried et al. followed a similar approach to this thesis by investigating the feasibility of a chatbot that utilizes the closed-source GPT-3.5 Turbo LLM developed by OpenAI for practicing medical history-taking skills [Hol+24]. The implementation consisted of an iteratively developed system prompt that instructed the LLM to simulate a VP, along with a web interface that allowed learners to interact with the chatbot. To assess the chatbot's usability and performance, medical students were invited to engage in conversations with the VP at their discretion and subsequently complete a Chatbot Usability Questionnaire (CUQ). The ensuing analysis revealed that the chatbot was capable of generating medically plausible answers, even when presented with questions outside of the scripted content. However, the analysis also showed that some implausible answers indicated GPT's tendency to rely on abductive reasoning—filling in gaps with general knowledge—rather than strictly adhering to the script, as well as exhibiting a social desirability bias when going off-script. Ultimately, the study concluded that students expressed satisfaction with the overall usability and authenticity of the chatbot [Hol+24].

While these findings illustrate the potential of LLMs in medical education, scholars consistently warn against viewing them as replacements for traditional in-person medical training. Persistent limitations include hallucinations in the form of confidently stated but incorrect responses, as well as unresolved ethical and privacy concerns [LUR24]. Ethical risks stem from the potential for biased or discriminatory outputs [Wan+23; Bro+20], and the inability of LLMs to fully represent the diversity of real-world patient populations [Ben+24b]. Moreover, handling sensitive patient data raises privacy and confidentiality challenges [LUR24; Lee24], and the quality of LLM outputs remains dependent on prompt design and user expertise—thereby necessitating active instructor involvement in creating effective patient scenarios [Yan+24].

In conclusion, the literature suggests that while the benefits of LLMs in medical education should be leveraged, the aforementioned risks and shortcomings associated with this technology must be taken into consideration [LUR24; Saf+23; Ben+24b].

## 3.3 Metrics to Assess Pedagogical Effectiveness in LLM-driven Chatbots

As previously shown, the integration of LLM-driven chatbots into medical education presents new opportunities for interactive and realistic learning environments, including patient simulations. However, in order to measure whether such systems provide meaningful educational value, it is essential to define and apply robust metrics for pedagogical effectiveness. This section reviews existing research that defines and utilizes these metrics. Given that Symptex is designed for deployment within the German medical education system, particular attention is paid to studies that focus on the German educational (healthcare) context. However, due to the limited availability of domain-specific research, especially regarding LLM-powered VPs for German healthcare education, this review also draws on relevant work from adjacent fields—particularly computer science education—where chatbot-based learning tools have been more extensively explored and evaluated.

Among the approaches previously discussed, only the implementation by Holderried et al. was tested in a German healthcare education setting, namely at the Tübingen Institute for Medical Education [Hol+24]. To evaluate their work, the authors applied a mixed-method approach combining quantitative and qualitative techniques.

For the qualitative analysis, the authors conducted a review of the dialogue, which focused specifically on the question-answer pairs (QAPs) generated during the interactions between students and the chatbot. They employed the *Braun-Clarke thematic analysis*, a flexible method for identifying, analyzing, and reporting patterns—referred to as codes or themes—in various types of data, including textual data such as QAPs [BC06]. Holderried et al. examined 826 QAPs in total, collected from interactions between 28 medical students and the chatbot. They coded the responses based on the following criteria [Hol+24]:

1. **Conversation Part**: What part of the conversation does the response correspond to, for instance, greeting, medical history, or goodbye?

2. **Script reliance**: Did the questions and answers align with the script provided to the chatbot, which included behavioral instructions and patient details?

3. **Plausibility**: Were the answers plausible in the context of the patient case?

Combined with quantitative methods, such as the assessment of user experience using the *Chatbot Usability Questionnaire (CUQ)* and statistical analysis of the correlation between question length and answer length, this mixed-method approach provided insights into both the system's usability for learners and the medical validity of its outputs [Hol+24].

Beyond this isolated example, most existing studies in German education involving chatbots have emerged from the field of computer science. Despite the domain difference, these studies are pedagogically informative, especially in how they structure evaluation frameworks. For example, in the *IRIS* project, a chat-based tutor was conceptually integrated into a learning platform for Technical University of Munich (TUM) computer science students. The tutor was assessed by students using a Likert-scale survey designed around specific research questions focusing on perceived impact, user reliance, and whether students preferred the chatbot over human support [BFK24].

A more theory-driven evaluation approach is exemplified by the *MoodleBot* project at RWTH Aachen University. This chatbot, designed to support self-regulated learning within the Moodle platform of selected courses, was assessed using the Technology Acceptance Model (TAM) [Neu+24]. Developed by Davis, TAM provides a theoretical framework for predicting users' acceptance and adoption of new technology [Dav89]. Central to the model are the constructs of *perceived usefulness* (the degree to which a user believes that the system enhances their performance) and *perceived ease of use* (the extent to which the user expects the system to be free of effort), both of which influence the user's attitude toward using the technology and their behavioral intention to use it. By grounding their evaluation in TAM, the authors were able to systematically capture learners' cognitive and affective responses to the chatbot, thereby offering a detailed understanding of user acceptance. Moreover, the study complemented this user-centered analysis with a technical validation of the chatbot's content accuracy, employing both automated fact-checking and manual review to assess the correctness of generated responses [Neu+24].

Overall, surveys remain the most common instrument across studies for assessing chatbot usefulness and usability. However, despite growing interest in the educational applications of LLMs, there remains a lack of validated, domain-specific criteria for evaluating the pedagogical factor of chatbot performance in particular. Most implementations either rely solely on subjective feedback collected through surveys or combine general-purpose evaluation frameworks with correlation analysis to validate their concepts.

## 3.4 Positioning of This Work

While research on LLM-powered chatbots in medical education is increasing, most studies currently examine their function as virtual assistants that deliver informative or instructional content. In contrast, significantly fewer projects have investigated the potential of chatbots for role-playing purposes, such as in VP simulation scenarios. This lack of exploration is surprising, given the previously shown well-documented educational benefits of VP-based training, especially in fostering clinical reasoning and communication skills [Pla+22; Kon+15]. In addition, while prior VP prototypes have demonstrated satisfactory results regarding their functionality and usability [Cam+20; Hol+24], their pedagogical value to learners has often remained underexplored.

This thesis contributes to this underexplored area by building on prior research in educational chatbots and VP systems, while also placing emphasis on pedagogical value.

It proposes the design and implementation of the LLM-powered chatbot Symptex, with the goal of simulating VPs with varying symptoms and clinical backgrounds. Symptex seeks to leverage role-playing strengths and medical knowledge of LLMs to enable dynamic and authentic responses in the context of German medical history-taking interviews, thereby assisting medical students in improving clinical communication skills. Learning outcomes are further reinforced with the generation of a follow-up performance evaluation for the students.

The prototype is embedded in the ILuVI framework, a curricular initiative of TUM MRI that seeks to incorporate customizable VPs across different lectures in medical education using a flipped classroom approach. This aligns with the pedagogical recommendations highlighted by Berman et al., which suggest combining structured self-learning with interactive components to promote deeper understanding [Ber+16].

Symptex was developed in collaboration with doctoral candidate Johannes Reifenrath and medical educator PD Dr. Dr. med. Alexander Zink, MPH, MBA, senior physician at the Department of Dermatology, TUM MRI, to ensure that the implementation is aligned with the needs of instructors and compatible with existing teaching workflows. This collaborative approach thereby addresses the aforementioned barrier in VP adoption, namely the disconnect between available digital tools and their practical use in everyday teaching [Ber+16; CT09].

# 4 Requirements Analysis

The overall goal of this thesis is to assist medical students in improving their medical history-taking skills by conceptualizing, implementing, and evaluating an LLM-powered chatbot with personalized performance feedback. To achieve this, Symptex must be able to simulate an authentic virtual patient (VP) with a health issue, giving human-like responses while being embedded within the ILuVI framework.

This chapter gives a comprehensive analysis as the foundation for Symptex's development and integration into ILuVI. Section 4.1 begins with an examination of the current state of the ILuVI framework, briefly describing its existing capabilities and the limitations of its medical history-taking functionality. Building on this comparative perspective, Symptex is introduced as the proposed enhancement to address the identified shortcomings in Section 4.2. Subsequently, functional requirements are derived to detail the features that it must include, while non-functional requirements specify qualitative constraints. Lastly, system models provide a structured visual overview of the outlined requirements in 4.3.

## 4.1 Existing System

Much of the technical background of ILuVI has already been covered in Section 2.3 of the Background chapter. This section therefore focuses on the most relevant aspects for understanding the current system in the context of anamnesis training.

The overall objective of the ILuVI framework is to establish a virtual trial-and-error sandbox within the scope of regular lecture courses. In this environment, students engage with instructor-created clinical case studies, each featuring a VP with varying symptoms that they need to diagnose. Throughout this process, students are given the freedom to request any diagnostic procedures they deem necessary. Simultaneously, virtual time and cost constraints imposed by the instructor encourage them to practice efficient clinical reasoning and resource management skills, while ensuring that no real patients are placed at risk. Upon completion, an evaluation process considers the student's diagnosis, the appropriateness of the requested diagnostic procedures, and the effectiveness of resource management, with results subsequently displayed to the student.

This sandbox concept does not yet, however, apply to the medical history-taking workflow of the ILuVI case studies: As of now, students are required to click on buttons labeled with predefined anamnesis categories, which then display a fixed, pre-written text containing information about the VP (see Fig. 4.1).

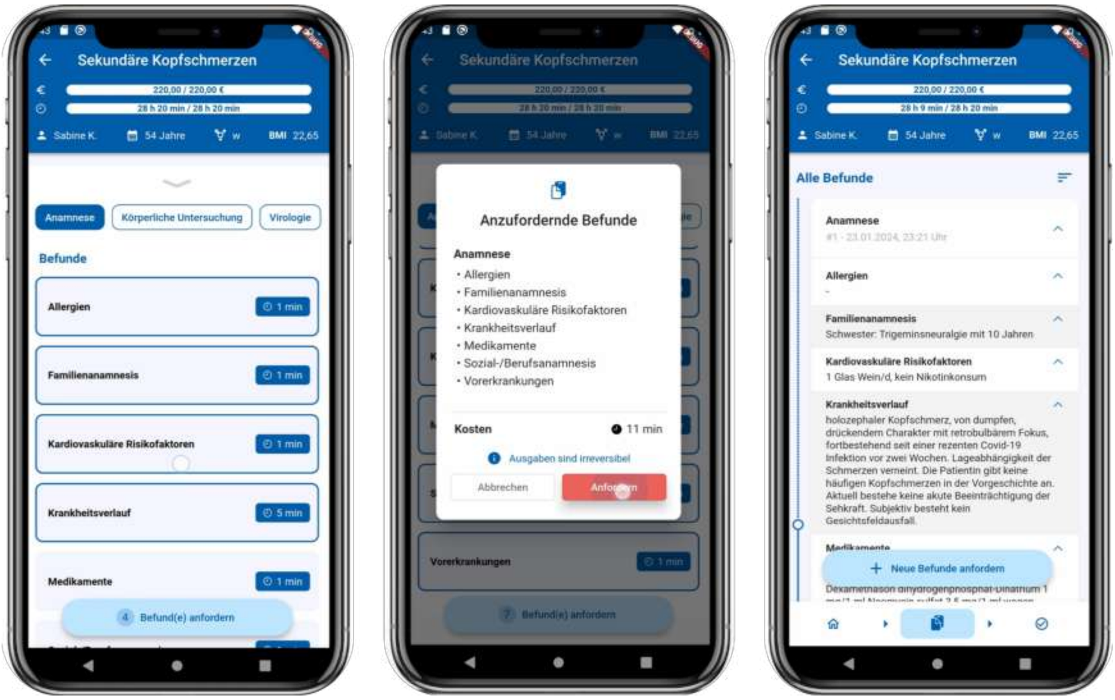


Figure 4.1: ILuVI Mobile Application: Current Anamnesis Process

This design reduces the complexity of a real-world medical history-taking conversation to a mechanical exercise, limiting the student's ability to practice different questioning strategies and to receive dynamic, unpredictable responses—communication skills that are essential for doctor-patient interactions. In addition, due to the current simplicity of the exercise, no feedback is given regarding the anamnesis process, further limiting opportunities for students to learn and improve their skills.

## 4.2 Proposed System

To address the limitations of the current anamnesis workflow, this thesis proposes Symptex as an LLM-driven enhancement. Through VP simulation, Symptex intends to create a more interactive and authentic learning experience for medical students. The LLM-powered chatbot assumes the role of a VP with specific health concerns, while students act as physicians, freely taking the medical history through natural written dialogue. Instead of selecting questions to receive pre-written text passages with the necessary information, students are now required to communicate with a chatbot to take the patient's medical history. By leveraging the conversational and role-playing abilities of LLMs, the chatbot produces context-aware and human-like responses, effectively simulating the unpredictability and emotional aspects found in real doctor-patient conversations.

In addition, Symptex provides immediate personalized feedback on the student's anamnesis performance. This feedback highlights both strengths and areas for improvement in their communication style, questioning strategy, and information gathering, thereby reinforcing the learning process.

Overall, this approach transforms the anamnesis phase from a static selection task without pedagogical feedback into an interactive, feedback-driven learning exercise, enabling students to develop adaptive questioning strategies and communication skills.

Within the context of ILuVI's healthcare education framework, the proposed chatbot prototype must satisfy certain key requirements to achieve its goal of improving the learning experience for medical students. These requirements will be discussed in the following section.

### 4.2.1 Functional Requirements

**FR1: Simulate VPs in the context of medical history-taking conversations**

The ILuVI framework is designed to assist instructors of medical lectures that focus on varying health complications. The chatbot must therefore support VPs with different challenges, such as hearing impairment, memory loss, and other issues related to health conditions that can complicate the process of collecting essential patient history in real-world scenarios. The corresponding functionality can be further detailed through the following requirements:

**FR1.1: Generate condition-aligned responses**

When receiving a question or a prompt, Symptex must be able to generate contextually accurate and coherent responses. The behavior of the chatbot must further align with the medical condition assigned to the VP. For example, a complication involving memory loss should result in consistent responses that reflect confusion or contradiction.

**FR1.2: Provide context configurability**

In addition, the chatbot should allow the instructor to configure its context so that it supports the different aforementioned medical conditions on demand. For example, an instructor should be able to configure one case study featuring a geriatric patient with dementia and another involving a patient with a hearing impairment.

**FR1.3: Recall knowledge**
The chatbot should also be able to retain details about the current patient's medical history, i.e., their age or previous health incidents, to ensure consistent responses throughout the conversation. Contradictory responses may confuse students and reduce the authenticity of the simulation.

**FR2: Chatroom User Interface**
As mentioned previously, the anamnesis phase currently incorporates a simple user interface, in which the medical student must tap on a list of buttons labeled with predetermined questions to receive static information about the VP. To enable user-defined input with dynamic responses, the chatbot user interface should be designed in an intuitive chatroom-based manner with the following functionalities:

**FR2.1: Enable message input**
The user interface must provide a clearly visible and accessible input field that allows the student to freely converse with the chatbot by composing and sending text messages. The input field should also support editing longer messages by being scrollable.

**FR2.2: Display chatbot responses**
The user interface should display both messages sent by the user and chatbot responses in a human-readable format, with all messages appearing in chronological order. This would ensure that the medical student can intuitively follow the progression of the dialogue and interpret the chatbot's responses within the current context.

**FR2.3: Maintain message history**
The complete conversation history should be persisted and continuously displayed during a case study session. This way, the medical student can look up the conversation at any time, similar to a doctor reviewing the notes taken during an anamnesis interview.

**FR3: Evaluate Student**
After the student completes their anamnesis with the chatbot, Symptex should offer immediate feedback on the student's performance in eliciting patient information and symptoms in a patient-centered manner.

**FR4: Adjust Patient Profile**
Symptex must allow instructors to make adjustments to the simulated patient

profiles. In addition to the comprehensive medical history, which includes the patient's health complications and medications, instructors should also be able to customize challenges associated with the VP, such as memory loss and the patient's level of talkativeness. This flexibility enables Symptex to accommodate a broader range of medical scenarios, making it particularly beneficial for future instructors who may want to use and adapt the VPs to meet their specific teaching needs.

### 4.2.2 Nonfunctional Requirements

**NFR1: Patient Simulation Language**
Language is the primary interface between users and LLMs, making it a crucial qualitative factor for a chatbot. Based on the linguistic quality and structure of the dialogue, authenticity, and learner motivation can be significantly improved or diminished. For this research, the following three key linguistic aspects have been identified as the most relevant:

**NFR1.1: Linguistic Fluency**
The primary target group of medical students at TUM MRI is taught in German and will also mainly be interacting with patients in German in the future. Therefore, the LLM should be capable of generating text in fluent German to ensure that the students perceive the interaction as realistic. This includes, in particular, proper grammar and idiomatic expressions.

**NFR1.2: Accuracy**
To ensure a beneficial outcome for the medical student when using the chatbot, its answers should somehow match the medical condition of the VP it is simulating. For instance, the model not being able to remember some information when simulating a geriatric patient with dementia would provide a realistic clue to the student. In addition, the chatbot should focus on using layman's terms for medical terminology to increase credibility, as most patients possess no extensive medical knowledge.

**NFR1.3: Humanness**
Symptex should generate realistic, human-like answers to open-ended questions, while conveying emotion through written cues and gestures. Robotic replies without emotion may reduce the authenticity of the conversation, which in turn may discourage students from composing authentic diagnostic questions themselves, reducing the

educational value of the interaction.

**NFR2: Conversational Dynamics**
Symptex must respond contextually appropriate to the user's questions, since the conversation could involve follow-up questions that require maintaining consistency in patient details, symptoms, and overall story. The ability to adapt to the flow of the conversation further enhances the authenticity aspect of the doctor-patient interaction.

**NFR3: Educational Valuable Feedback**
The feedback given on the students' history-taking performance must be personalized, objective, and constructive, allowing them to reflect on and improve their communication skills for future encounters. This feedback must be grounded in research and presented in a logical and clear structure. To illustrate key points, the feedback should further include examples, ensuring that the student derives educational value from using Symptex.

**NFR4: Latency**
To ensure that Symptex feels responsive and usable, its response generation latency should be kept at an acceptable level from a subjective view. As there are currently no established standards regarding acceptable response times for LLMs, this work defines responses generated within ten seconds or less as meeting the criterion of low latency. Not adhering to this requirement may significantly impact the perceived authenticity of the conversation, as the medical history-taking process in real-world scenarios is usually conducted at a fast pace without substantial delays.

### 4.2.3 Constraints

The implementation of the chatbot is subject to several external constraints beyond the scope of this research, which influence its design, functionality, and evaluation:

- **ILuVI Framework Integration**
  The chatbot must be seamlessly integrated into the existing ILuVI framework to enable direct access to patient information, thus eliminating redundant data duplication. This further implies that a separate client is not required to facilitate communication with the LLM hosting server.
- **Limited Time and Infrastructure**
  Given constraints regarding time and computing resources throughout this study,

completing the following requirements for LLMs is outside the scope of this thesis:

- **Model Fine-Tuning for Ethical Guidelines Adherence**
  A chatbot model must be harmless, meaning that it should be, for instance, fine-tuned in a way that detects problematic prompts to avoid inappropriate responses and therefore adhere to general ethical guidelines.
- **Infrastructure to Ensure Data Privacy**
  Medical students interacting with the chatbot must be assured that their interactions and learning progress are managed by a secure infrastructure, especially if the system collects performance metrics that are associated with personal identifiers. Given that the expected implementation should be a first prototype with a limited number of test users, user privacy is not the main focus of this study.

## 4.3 System Models

This section introduces system models that frame the design of Symptex within the ILuVI framework. It first illustrates a visionary scenario of a student using the embedded Symptex module in practice, before presenting a use case model that specifies user roles and interaction flow. The analysis object model then outlines the main entities and their relations, while the dynamic model depicts the step-by-step conversation process. Finally, a mockup of the Symptex User Interface (UI) embedded within the ILuVI mobile application is presented.

### 4.3.1 Visionary Scenario

The following visionary scenario depicts a use case for medical students that might occur when utilizing the Symptex chatbot within the ILuVI framework. To ensure medical accuracy, this fictional case has been developed in collaboration with a doctoral candidate in medicine, Johannes Reifenrath. The complete chat history of the visionary scenario is written in German and can be reviewed in the Appendix 8.2.

#### Geriatric patient with memory loss

*During her lecture, medical student Alice Beck decides to solve the case study assigned by her instructor and opens the ILuVI mobile application on her phone. She carefully reviews the patient file to familiarize herself with her virtual patient, Anna Zank. Upon examining the information provided, Alice notices that Ms. Zank is an elderly woman who was admitted*

*after a fall. Recalling key characteristics often associated with geriatric patients, she proceeds to initiate the medical history-taking process by opening the Symptex chat.*
*She begins the conversation with a polite greeting, but quickly notices that Ms. Zank struggles to recall details, hesitates in her responses, and occasionally provides inconsistent information. When asked about the circumstances of her fall, Ms. Zank offers only fragmented answers and relies on vague recollections of her daughter's help. She appears disoriented regarding time and place, yet still remembers her name and some aspects of her past, such as her former occupation. Recognizing the signs of possible memory impairment, Alice adjusts her questioning strategy, ensuring that she asks clear and simple questions while simultaneously cross-referencing information. Alice manages to confirm the key facts of the incident, gains insight into the patient's cognitive difficulties, and notes physical pain localized to the hip area. Having collected these essential details, Alice closes the chat and starts to order specific diagnostic procedures to verify her preliminary assessment.*

### 4.3.2 Use Case Model

To illustrate how users interact with Symptex, the following use case model for the use case *Chat with Symptex* was created (see Fig. 4.2).

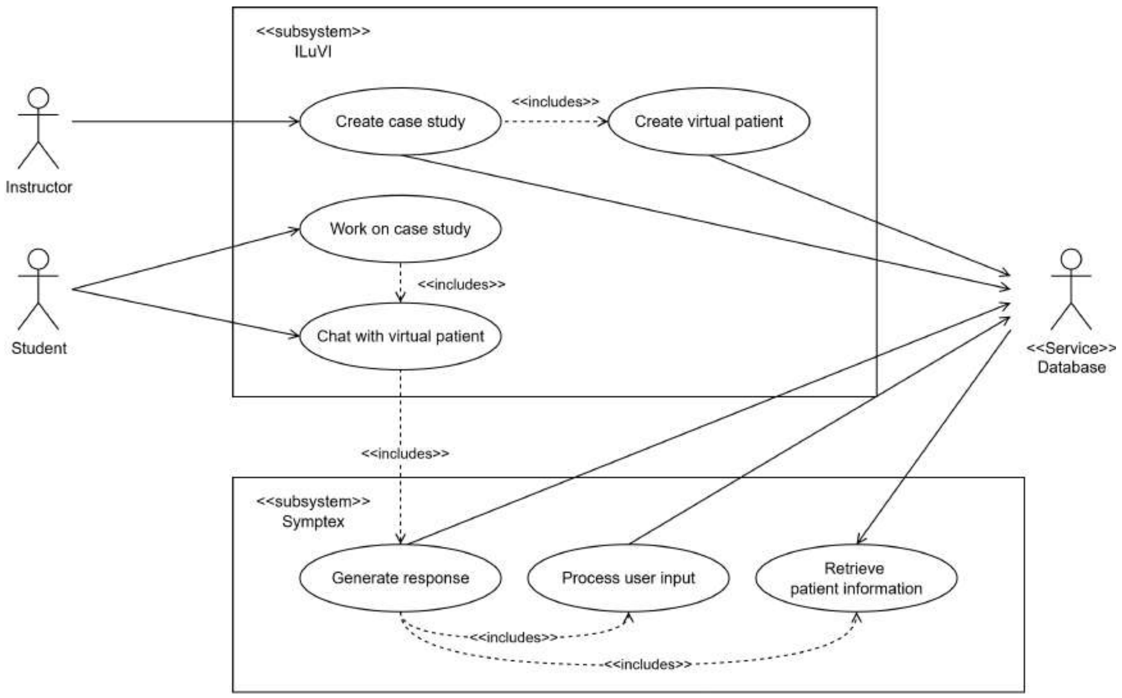


Figure 4.2: Use Case Model: Chat with Symptex

The use case involves two distinct types of users, each with different goals:

1. *Instructor*
   An individual who organizes a lecture during the semester. Their objective is to create case studies through the ILuVI web application, which allows students to practice clinical communication skills using VP scenarios.

2. *Student*
   A participant in the instructor's lecture. The student intends to complete the created case studies via the ILuVI mobile application to improve their anamnesis and diagnostic skills.

It further consists of a *Database* service, which persists all relevant data created throughout this use case. For this use case to proceed, the following **entry conditions** must be met:

- The *Instructor* is using a PC with the ILuVI web application open in a browser.
- The *Student* is using a phone with the ILuVI mobile application open.

The **exit condition** for this use case is that the *Student* receives and reads the response generated by Symptex, which is displayed on their phone, thus completing the interaction. The **flow of events** for this use case is outlined below:

1. The *Instructor* plans to complete their goal by first accessing the ILuVI subsystem through the website, initiating the use case.

2. Through the website, the *Instructor* creates a case study by filling out the form with the necessary information, which includes defining a VP and specifying the symptoms the patient should exhibit.

3. Once the case study is saved in the *Database* and published, the *Student* accesses it through the ILuVI application on their mobile device.

4. The task of working on the case study also involves interacting with the VP through chat to take the medical history. Once the *Student* starts chatting, the event leaves the ILuVI subsystem, as now the Symptex subsystem persists the user input to the *Database*, processes it, and retrieves relevant patient information to generate a response.

5. After the response is generated, it is saved to the *Database* as well and, in turn, sent back to the *Student* and displayed on the ILuVI mobile application.

Throughout this use case, certain **special requirements** must be met for successful completion:

- The devices of both actors, PC and phone, must be connected to the internet.
- The PC must have a working browser installed to access the ILuVI web application.
- The phone must have the ILuVI mobile application installed.
- Both devices must have sufficient battery life for uninterrupted usage.
- The database must have enough available space to persist the generated data.

### 4.3.3 Analysis Object Model

In this subsection, every relevant component used in the application, its attributes, relations towards other objects, and placement within the Symptex system are described in an Analysis Object Model (see Fig. 4.3).

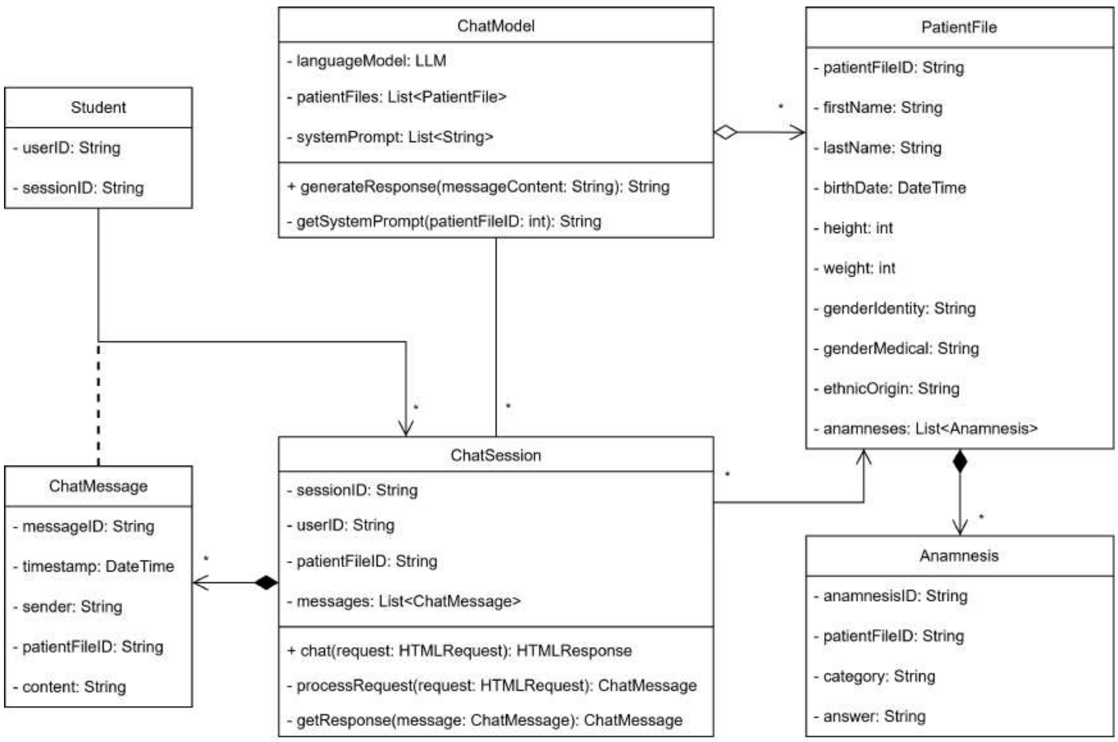


Figure 4.3: Analysis Object Model of Symptex

The *ChatSession* is situated at the center of the model and manages the current state and message history of a single ongoing anamnesis simulation. Each session stores the identifiers of the interacting student and the selected patient file, linking the student to a specific scenario. A session contains multiple *ChatMessage* objects, which record the exchanged content along with metadata such as sender, timestamp, and the patient file ID. This way, the session acts as both a conversation container and the operational center for coordinating requests and responses. Incoming messages, received as *HTMLRequests* via the *chat* method, are internally transformed into *ChatMessage* objects and forwarded to the *ChatModel* for response generation. Since the HTMLRequests and HTMLResponse possess the same content as the *ChatMessage* object, these components are omitted in the diagram for simplicity.

The *ChatModel* encapsulates the core intelligence of Symptex. It includes references to the underlying LLM, a list of available *PatientFile* objects, and the associated system prompts. Its main responsibility is to generate patient responses through the *generateResponse* method, which leverages both the LLM and the configured system prompt via *getSystemPrompt* to ensure medically consistent and realistic replies.

Patient data is stored in the *PatientFile* entity, which contains general demographic and biometric information. Each patient file is associated with a set of *Anamnesis* objects. These represent structured medical history entries, organized by category (e.g., course of disease, medication, and family history), along with the corresponding answers.

The *Student* object is comparatively simple, storing only the user and session identifiers. It serves as the entry point of the interaction, linking external users to the internal session management and ensuring that each conversation is uniquely attributable.

### 4.3.4 Dynamic Model

The dynamic model illustrates the behavior of Symptex during the medical history-taking phase in an *ILuVI* case study (see Fig. 1 in the Appendix). It details the procedural flow from the perspective of three interacting entities: the *Student* as the primary user, the *ILuVI* platform as the mediating subsystem, and *Symptex* as the underlying response-generation component.

The process begins with the *Student* initiating the use case by opening an assigned case study. This action triggers the *ILuVI* subsystem to retrieve the corresponding case data, which includes the VP profile, and subsequently displays it to the student. The *Student* then reviews the patient file, which provides the necessary basic information for the following anamnesis dialogue.

As the *Student* starts the anamnesis phase, the *ILuVI* mobile application loads and displays the chat interface, allowing the student to begin interacting with the VP. This initiates a loop of back-and-forth interactions between the patient and the *Student* that form the core of the learning experience.

In each iteration, the *Student* composes a prompt or question, which is submitted through the chat interface. *ILuVI* captures this input and forwards it to the *Symptex* subsystem. Internally, *Symptex* processes the prompt and generates a response using an LLM. The generated response is then sent back to *ILuVI*, where it is displayed to the *Student*.

The *Student* reads the patient response and evaluates whether sufficient diagnostic detail has been obtained. If not, the dialogue loop continues with further question-answer cycles. This iterative process reflects real-world anamnesis practices, where physicians must query patients based on the evolving context of the conversation.

Once the *Student* determines that all necessary information has been gathered to proceed with clinical reasoning or diagnosis, the anamnesis phase is concluded. This marks the end of the interaction sequence for the VP encounter.

Overall, this dynamic model highlights the role of Symptex as an embedded dialogue engine that operates asynchronously within the broader ILuVI learning ecosystem.

### 4.3.5 User Interface

When designing the system, particular attention was given to minimizing cognitive load, as non-overwhelming interaction designs have been shown to increase student acceptance and perceived value of VP solutions [Kel+22]. To this end, a minimal chatroom user interface was designed, which mirrors the simplicity and focus of real-world clinical doctor-patient conversations. The initial mockup (see Fig. 4.4) is therefore deliberately designed to be simple, yet intuitive and familiar, drawing on conventions from modern social media messaging platforms to reduce the learning curve.

To ensure visual consistency within the ILuVI ecosystem, the interface adopts the Material Design 3 guidelines, emphasizing clean, rounded, and shadowless 2D components. The chat functionality itself is embedded into a slide-up window, which appears once the student taps on the "Start Anamnesis" button. The same slide-up element is used for ordering laboratory procedures, thereby aligning the chat with existing interaction patterns in ILuVI and maintaining a cohesive user experience.

The top of the chat interface is occupied by a persistent header that displays a thumbnail image of the patient along with their name, serving as a consistent visual anchor throughout the session. The main area of the interface is dedicated to the dialogue exchange, as this is where students elicit and interpret clinically relevant information. In line with established digital communication norms, the patient's messages are left-aligned, while the student's inputs appear on the right. The student's messages are further visually distinguishable through a blue background, which matches the application's primary color scheme. To promote clarity and reduce visual clutter, timestamps are displayed only at the beginning of the conversation rather than with every message, keeping the student's focus on content rather than metadata.

The input field at the bottom is set apart in light grey and supports both text input and optional speech recognition, thereby increasing accessibility by allowing students to choose their preferred mode of interaction.

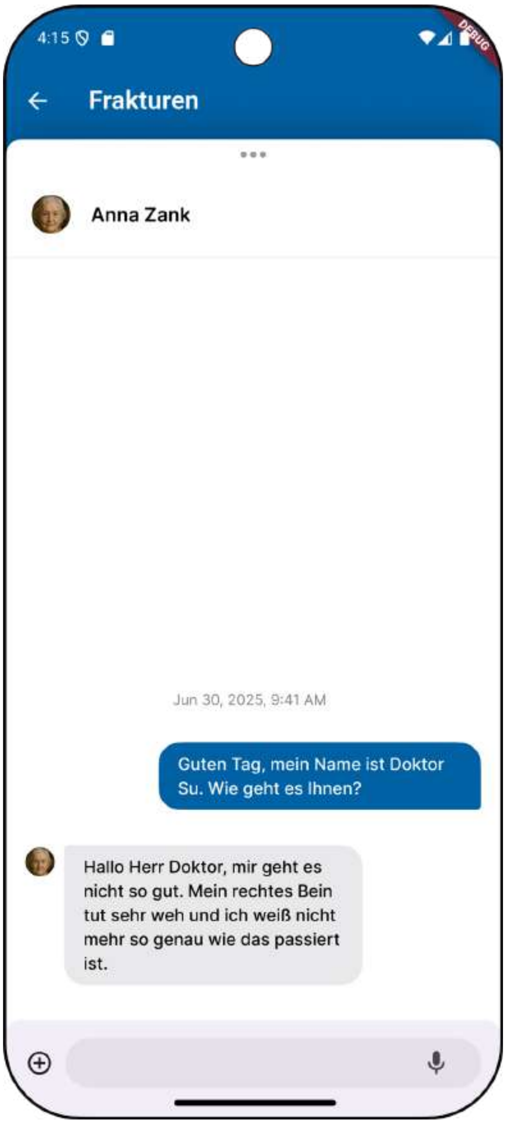


Figure 4.4: Symptex UI Mockup

# 5 System Design

This chapter presents the architectural design of Symptex, which is guided by design goals introduced in Section 5.1. Section 5.2 outlines the overall architectural style, situating Symptex within a layered framework. Section 5.3 elaborates on the system's functional decomposition, highlighting its client-server architecture and the roles of its modular subsystems. Section 5.4 then details the hardware/software mapping, focusing on how Symptex is deployed within ILuVI and how this constraint influences the system's design. Lastly, the persistent data management is outlined in Section 5.5, visualizing what data and how the data is saved throughout the lifecycle of Symptex.

## 5.1 Design Goals

The following design goals are derived from the non-functional requirements (NFRs) outlined previously in Section 4.2.2. These goals represent key priorities and trade-offs that must be considered during the development of Symptex:

**DG1: Patient Language Authenticity**
Aligned with NFR1 (Patient Simulation Language), the Virtual Patient (VP) must communicate in a natural, emotionally expressive, and contextually appropriate manner. This goal is critical, as it directly influences the user-perceived authenticity and educational value of the simulation. However, achieving high linguistic authenticity may require the use of larger and more powerful models, which could conflict with NFR4 (Latency) due to increased computational demands.

**DG2: Conversational Consistency**
Reflecting NFR2 (Conversational Dynamics), Symptex should maintain coherence and contextual accuracy throughout multiple dialogue turns. This includes preserving patient information, emotional tone, as well as overall content-related consistency during the interaction. Like DG1, this goal significantly affects user perception and educational value, making it equally critical. Nevertheless, it may also introduce challenges for NFR4, as managing increasing context over concurrent sessions can be resource-intensive.

**DG3: Effective Feedback**
Derived from NFR3 (Educational Valuable Feedback), the offered feedback after each simulation session should be personalized and highly educational for the learner. This feedback should aim to help students reflect on their performance and pinpoint areas for improvement. This goal is critical as it ensures sustained pedagogical effectiveness in an educational setting such as TUM MRI.

**DG4: Responsive User Experience/Performance**
In line with NFR4 (Latency), the system must maintain responsiveness and ensure low-latency answers, even when processing longer messages. This goal is essential for realistic VP simulations. However, this may necessitate compromises in model size or complexity, which could, in turn, impact DG1 and DG2.

## 5.2 Architectural Style

Symptex adopts a layered architecture that separates responsibilities into four distinct layers:

1. **Client-side Layer**
   Provides the user-facing interface embedded within the ILuVI framework through which medical students interact with the system.

2. **Routing Layer**
   Serves as the communication bridge between the client-side layer and the conversational logic layer. It processes incoming HTTP requests, performs input validation, and handles response streaming to ensure low-latency interaction with the LLM.

3. **Conversational Layer**
   Implements the core functionality of Symptex by encapsulating VP simulation workflows, prompt management, and feedback generation.

4. **Persistence Layer**
   Manages the long-term storage and retrieval of conversation transcripts, patient profiles, and related data.

This layered approach enforces a clear separation of concerns, reducing coupling between modules and allowing independent testability and scalability of each layer. For instance, changes to the prompt engineering logic in the Conversational Logic Layer do not affect database schema design in the Persistence Layer, and vice versa.

From an LLM architecture perspective, Symptex employs a lightweight approach, refraining from building on more complex paradigms such as retrieval-augmented generation (RAG), tool-calling, or multi-agent frameworks. As outlined in Section 4.2.1, the primary functional requirements of Symptex are narrowly defined: the authentic simulation of adjustable VPs in anamnesis conversations (FR1, FR4), and the provision of a follow-up performance evaluation for students (FR3). Both requirements can be met without external tool integration or RAG, as all patient information is stored within ILuVI. Moreover, pilot evaluation results (see Section 7) further indicate that students and experts judged Symptex's simulation capabilities as authentic and clinically plausible. This suggests that the chosen lightweight architecture provides sufficient accuracy for medical history-taking training, while avoiding the additional complexity and latency that multi-agent or RAG-based systems would introduce.

## 5.3 Subsystem Decomposition

The Subsystem Decomposition (see Fig. 5.1) translates the layered architecture into four interacting subsystems. It follows a traditional Client-Server architecture, as well as the Facade pattern, by providing the Symptex User Interface (UI) client with a simple interface implemented in the *Chat Model Controller* to access the underlying complex Symptex server system.

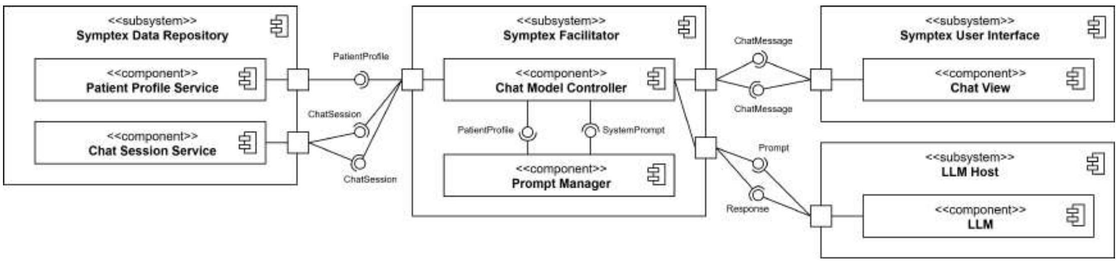


Figure 5.1: Subsystem Decomposition of Symptex

**Symptex User Interface (Client-side Layer)**
The Symptex User Interface subsystem consists of one component, the *Chat View*. This view is presented to the learner, allowing them to interact with a VP. When the *Chat View* is initially loaded or reloaded, any existing chat messages are retrieved from the *Chat Model Controller* in the form of *ChatMessage* data. Once the learner sends a message, the *Chat View* transmits the message as a *ChatMessage* to the *Chat Model Controller*. If the response is successful, the *Chat View* streams the content of the *ChatMessage* and displays it to the learner.

Due to its narrow responsibility of sending and receiving chat messages, this subsystem exhibits minimal coupling to the Symptex server-side logic—interacting with the Symptex Facilitator subsystem only via the *ChatMessage* interface.

**Symptex Facilitator (Routing Layer, Conversational Layer)**
The Symptex Facilitator represents the central orchestrating subsystem of the architecture and combines responsibilities from both the Routing Layer and the Conversational Logic Layer. For reasons of clarity and simplicity, the *Chat Model Controller* component is depicted here as fulfilling both roles, while the technical realization of the routing functionality is described in more detail in Section 5.4 on hardware/software mapping.

The *Chat Model Controller* connects to the three other subsystems—the Symptex Data Repository, the Symptex User Interface, and the LLM Host subsystem—via public interfaces.

Its main workflow can be described as follows: When the learner initiates a chat with the VP through the *Chat View* in the Symptex User Interface, the *Chat Model Controller* receives and processes the user message in the form of a *ChatMessage*. Using the data from the *ChatMessage*, it then retrieves both *PatientProfile* and *ChatSession* data from the services in the Symptex Data Repository. The *ChatSession* is required for the generation of coherent context-aware responses, while the *PatientProfile* is used to obtain the matching system prompt from the *Prompt Manager*, which should include specific health data related to the patient. Once the Prompt Manager creates and sends the appropriate system prompt back to the *Chat Model Controller*, this controller forwards the user input, along with the previous chat history, to the LLM Host. Upon receiving a successful response, the *Chat Model Controller* sends the stream of the generated response back to the Symptex User Interface as a *ChatMessage*.

The internal cohesion of this subsystem relies on a fixed logical sequence (receive message $\rightarrow$ retrieve necessary data $\rightarrow$ stream response), with the components thereby only exchanging functionally related data structures. This facilitates robust prompt customization while isolating system logic from LLM behavior.

**Symptex Data Repository (Persistence Layer)**
The Symptex Data Repository subsystem realizes the Persistence Layer and is tasked with storing, retrieving, and transmitting data relevant to the Symptex system.

The *Patient Profile Service* component is responsible for fetching and sending *PatientProfile* data associated with the current VP case to the *Chat Model Controller*.

Each learner's individual *ChatSession* objects are managed by the *Chat Session Service* component. It is frequently updated whenever a new message is added by either the learner or the LLM to a specific chat session. This component is also addressed when a learner (re)loads their chat user interface. If that *ChatSession* contains pre-existing messages, those messages are retrieved and sent to the user through the *Chat Model Controller*.

Both services demonstrate functional cohesion by focusing solely on their distinct data entities. This separation reduces the risk of unintended side effects and promotes independent scaling or substitution of the underlying database technology.

**LLM Host (Conversational Layer)**
The final subsystem, the LLM Host, is responsible for abstracting the access to the *LLM* component and hosting it on a powerful server. The inputs of the learner arrive in the form of a Prompt, which includes not only the user's message but also the entire chat history. This context allows the LLM to generate coherent responses that are aware of the ongoing conversation. Once the processing is complete, the LLM streams the response back to the *Chat Model Controller*.

Even though this subsystem technically belongs to the Conversational Layer of Symptex, it is intentionally decoupled from the rest of the Symptex system, so that LLM and its resources can be upgraded or even replaced without necessitating changes to the Symptex Facilitator or Data Repository.

## 5.4 Hardware/Software Mapping

The Hardware/Software Mapping, illustrated in Fig. 5.2, describes the deployment architecture of the Symptex system. It outlines how software components are distributed and connected across different hardware devices and execution environments. This deployment diagram also contains several ILuVI components, as Symptex is integrated within the ILuVI framework, thereby partially sharing device and execution environments. However, ILuVI-specific components or details that have no relation to Symptex are intentionally omitted for clarity purposes.

The system is distributed across four main device types: *Student Phone*, *Instructor PC*, *ILuVI Server*, and an external *ESX Server*.

**Student Phone**
The student interacts with the ILuVI framework through a pre-implemented Flutter-based mobile application (Flutter 3.29.2). Thereby, Symptex adopts the Flutter architecture for the implementation of chat interaction UI as well. The execution environment hosts two primary user interfaces:

- *ILuVI StudentUI*
  The main user interface presented to the medical students was mostly already

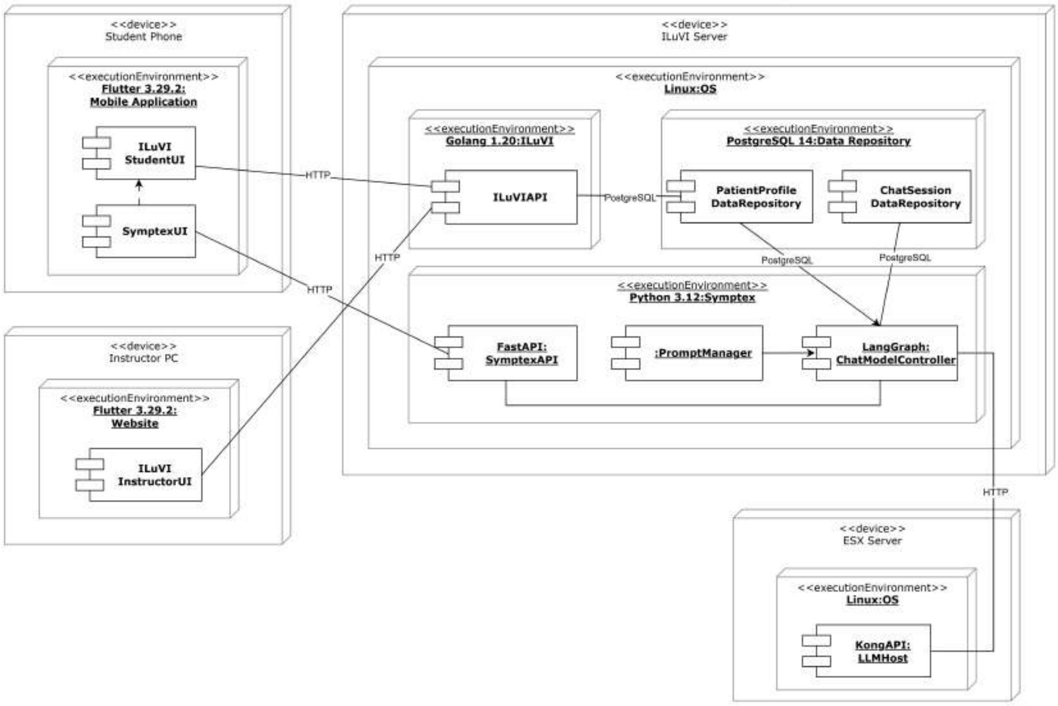


Figure 5.2: Hardware/Software Mapping of the ILuVI Framework with the embedded Symptex Module

implemented. It consists of the aforementioned views built in the ILuVI framework, including UIs for the Library and the Doctor's Office components. The necessary data, including patient cases, is retrieved via HTTP calls from the REST API hosted on the ILuVI Server.

- *Symptex UI*
  This UI chat component is designed specifically for chatbot interactions during the medical history-taking process. The view is integrated within the Doctor's Office, making it dependent on its Flutter UI hierarchy. On chat interaction, the respective Chat Session data is updated and persisted through HTTP calls to the SymptexAPI component.

**ILuVI Server**
The ILuVI server system operates on a Linux-based virtual machine and hosts three distinct dockerized micro-services:

- **ILuVI**
  The server logic of the ILuVI framework was already implemented and built using Golang 1.20. It primarily features a REST-based *ILuVIAPI* component, which handles HTTP requests from Student Phone and Instructor PC clients. Instructors can submit patient cases that students can later retrieve. This process involves accessing the underlying data repository using the TCP/IP-based PostgreSQL message protocol.

- **Data Repository**
  The data repository subsystem detailed in section 5.3 is built with PostgreSQL 14. This database system was selected for simplicity purposes, since the ILuVI framework already has a PostgreSQL database set up. In addition to data repositories specific to ILuVI, it also includes the *PatientProfile Repository* and the *ChatSession Repository*. These repositories provide services that can be addressed either through the ILuVIAPI component or the ChatModelController component for data manipulation, depending on which data is required.

- **Symptex**
  The core system of Symptex, responsible for generating LLM responses and storing chat messages, is realized in Python 3.12. The decision for this language was made due to its support for the established, scalable, and customizable LLM orchestration libraries LangGraph and LangChain.

  The FastAPI[1]-based *SymptexAPI* component acts as a gateway for chat interaction requests coming from the student's phone. When a student sends a message to the VP, this message is transmitted to the SymptexAPI component via HTTP. The SymptexAPI then forwards the message to the *ChatModelController*, which is built using the LangGraph library. As described previously in section 5.3, the controller then retrieves the relevant patient data from the PatientProfile DataRepository via the PostgreSQL message protocol to fetch the appropriate system prompt from the *PromptManager*. The retrieval queries are implemented using the SQLAlchemy[2] library. Simultaneously, the controller accesses the complete ChatSession from the ChatSession DataRepository via PostgreSQL as well, to ensure that the LLM has all the necessary information to generate a coherent response. With these prerequisites satisfied, the controller finally connects to the LLMHost through the ESX Server via HTTP for response generation. The generated response is streamed in real-time back to the SymptexUI through the SymptexAPI component, serving as the gateway.

[1] `https://fastapi.tiangolo.com/`, accessed on 24 August 2025
[2] `https://www.sqlalchemy.org/`, accessed on 24 August 2025

During each chat interaction, the updated chat session is consistently persisted to the ChatSession Data Repository via the PostgreSQL message protocol.

**ESX Server**
The LLM Host is managed externally by the KI-Servicezentrum für sensible und kritische Infrastrukturen (KISSKI) on their ESX Server[3]. It hosts the *LLMHost* component in a Linux environment, where it is exposed via KongAPI, a gateway layer facilitating secure HTTP-based communication from the ILuVI server to the hosted LLM. Resource upscaling for the LLMHost can be requested if necessary.

**Instructor PC**
For completeness, the Instructor PC is illustrated here as well. The instructor accesses the ILuVI framework through the already implemented *ILuVI InstructorUI*, also built with Flutter 3.29.2, but deployed as a website. Through this interface, the instructor can create and manipulate the VP data, which Symptex uses for adapting the system prompts. To communicate these changes to the data repository, this component also makes use of the HTTP protocol.

## 5.5 Persistent Data Management

The persistence layer of Symptex employs a relational schema optimized for conversational data storage and retrieval. The diagram in Figure 5.3 illustrates all persisted entities.

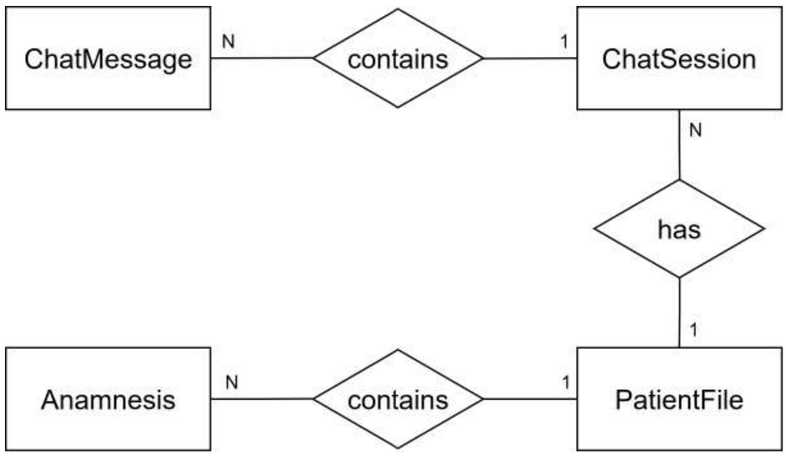


Figure 5.3: Entity Relationship Model of Symptex

[3]`https://docs.hpc.gwdg.de/services/saia/index.html`, accessed on 23 August 2025

The schema design supports queries for conversation reconstruction while maintaining referential integrity through foreign key constraints. All entities are uniquely identifiable and have unidirectional relationships, as illustrated in Figure 5.4, which are further described in greater detail.

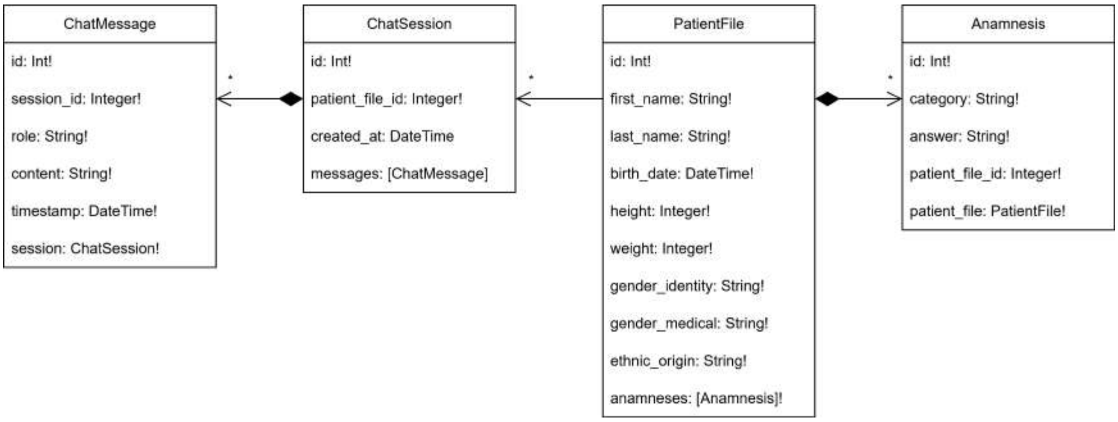


Figure 5.4: Persistent Objects in Symptex

- **ChatSession**
  Type representing a chat session started by a user. It contains a list of *ChatMessage* objects and references a *PatientFile* object, which determines the clinical patient scenario of the chat session.

- **ChatMessage**
  Type representing a chat message sent by either a user or the LLM. It references the associated *ChatSession* object and has a field with an instance of the *ChatSession* itself. Having both these fields allows for greater flexibility and follows SQLAlchemy best practices. The *ChatMessage* further includes the sender's role ("user" or "patient"), the content of the chat message, and the time stamp indicating when the message was sent.

- **PatientFile**
  Type representing a patient's file. It contains basic information about the patient, including their name, height, and gender. The *PatientFile* also contains a list of *Anamnesis* objects, which consists of data relevant to the patient's medical history.

- **Anamnesis**
  Type representing a component of a patient's medical history. It includes the category of information and the details provided in the answer field. The *Anamnesis*

also references the related *PatientFile* and includes a field that contains an instance of that specific *PatientFile*.

# 6 Object Design

This chapter describes the realization of the architectural design of Symptex through three main activities. Section 6.1 presents the implementation details of the previously described layers, outlining their respective responsibilities and design decisions. Section 6.2 details the systematic process of testing and selecting suitable LLMs to support both patient simulation and evaluation tasks. Finally, Section 6.3 elaborates on the development of effective system prompts that guide the behavior of the LLMs in these two use cases.

## 6.1 Implementation Details

As described in the subsystem decomposition (see Section 5.3), Symptex adopts a layered architecture to enforce a clear separation of concerns. In the following, the implementation details of the individual layers are further expanded upon.

### 6.1.1 Client-side Layer

Inspired by the mockup specified in Section 4.3.5, the final UI is realized in Flutter and Dart, and it is directly integrated into the ILuVI mobile application (see Fig. 6.1).

To realize the chat functionality, the Flutter library *Flyer Chat* is employed, which provides a modular and customizable chatroom interface. The interface consists of the following components:

- **Chat**
  The core widget of the Flyer Chat library provides an out-of-the-box chatroom UI with customization options such as the display of a patient-specific chat header. Once a message is submitted, the Send button is temporarily disabled until the full stream of the patient's response has been received. This prevents overlapping inputs and maintains the conversational flow. Messages authored by the student are displayed as *FlyerChatTextMessage* widgets, while incoming messages from the Virtual Patient (VP) are initially displayed as *FlyerChatTextStreamMessage* widgets. The latter are progressively updated as the LLM streams its output and are finally replaced by regular *FlyerChatTextMessage* widgets upon completion of the stream.


Figure 6.1: ILuVI Mobile Application: Symptex Chat

Error handling is also integrated into this layer: error messages are displayed in the chat interface itself, ensuring that participants are immediately informed about technical issues without interrupting the conversational flow.

- **ChatController**
  This controller, pre-implemented by *Flyer Chat*, serves as an abstraction layer between the chat UI and the underlying data source, such as a local database or a server-side API. It manages the complete list of messages, offering methods for all

CRUD operations. By default, it uses an in-memory message store, which enables temporary persistence across a single application session. For long-term storage, the controller also supports integration with persistent databases such as Hive, making it adaptable to both lightweight testing and production deployments within ILuVI.

- **ChatStreamManager**
  This custom component was designed to manage the real-time display of asynchronous message streams. Implemented using the Observer pattern, it ensures that listening UI components (such as the *Chat* widget) are automatically re-rendered whenever new content is available. This design facilitates a familiar response generation experience, allowing the patient's responses to appear word-by-word in the chat window.

For standalone development and debugging purposes, a supplementary Streamlit[1]-based implementation was created (see Fig. 6.2).

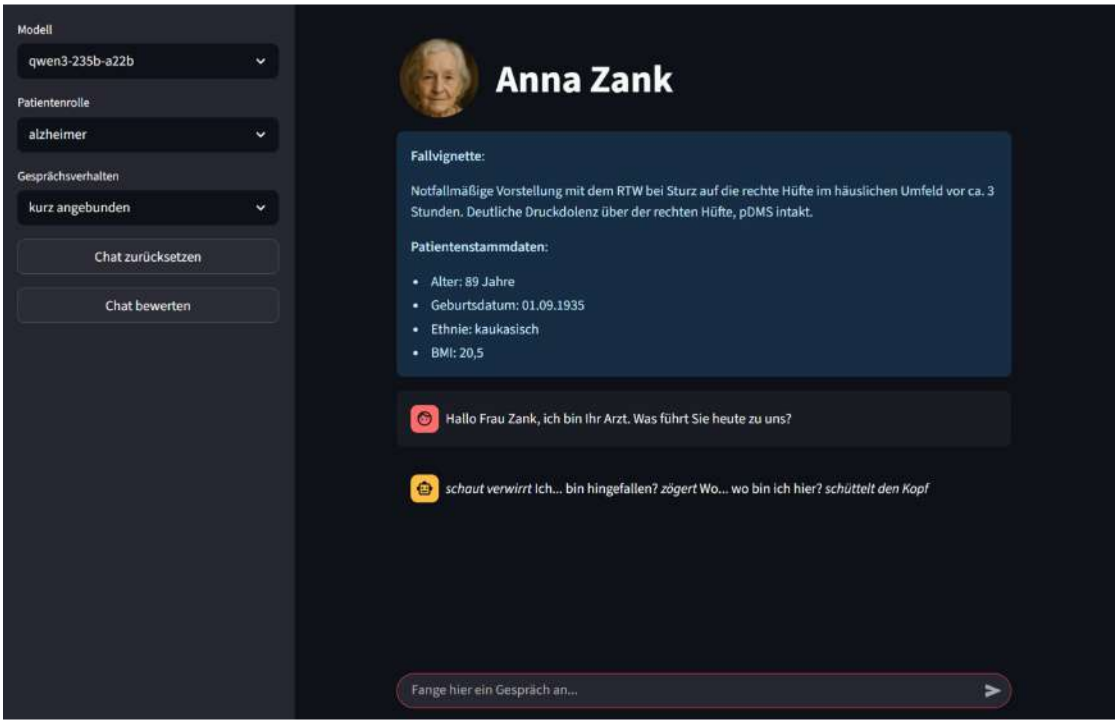


Figure 6.2: Symptex Web-based Chat

[1] `https://streamlit.io/`, accessed on 24 August 2025

This lightweight web application enabled rapid testing of LLM configurations through an integrated chat interface, without requiring compilation of a complete ILuVI mobile application build. By decoupling basic interaction testing from the full application pipeline, this dual-user-interface strategy significantly accelerated the development cycle while preserving a consistent interaction model between both environments.

To further facilitate the process of testing models with different configurations, the Streamlit interface was extended with additional functionality in the sidebar. These extensions included a dropdown menu for selecting the currently active model, options for switching between alternative system prompts, and in-string parameter adjustments within the prompt (for example, modifying the patient's symptoms or adjusting verbosity levels). Additionally, a reset button was added below the dropdown menus, allowing rapid restarts of conversations by deleting the accumulated chat history without manually restarting containers or deleting database objects.

### 6.1.2 Routing Layer

The Symptex API is implemented using the Python library *FastAPI*[2], which features high performance and asynchronous capabilities suitable for low-latency message streaming. Upon starting the API, the *SQLAlchemy ORM* initializes the database schema, ensuring that the required tables for retrieving patient data and managing conversation transcripts are available before any requests are processed.

The primary API endpoints are as follows:

- `/chat`
  Handles requests for real-time conversational interactions, streaming generated responses back to the sender.
- `/eval`
  Processes a completed conversation transcript to provide structured pedagogical feedback.
- `/reset/{session_id}`
  Resets the state of an ongoing session by clearing both ephemeral (in-memory) and persistent storage for the specified conversation.

Incoming HTTP requests are validated against *Pydantic BaseModel* schemas, ensuring strict type and format compliance at the network boundary. For instance, the *ChatRequest* model defines the expected structure for a conversational request (see Listing

[2]`https://fastapi.tiangolo.com/`, accessed on 24 August 2025

6.1). This approach provides robust type checking and data validation at the network boundary, preventing malformed data from entering the conversational logic layer.

```python
# Chat request schema
class ChatRequest(BaseModel):
    message: str
    model: str
    condition: str
    talkativeness: str
    patient_file_id: int
    session_id: str
```

Listing 6.1: Data Validation for Chat Requests

### 6.1.3 Conversational Layer

The core of the patient simulation is powered by the *LangChain*[3] and *LangGraph*[4] libraries, which together provide a declarative framework for defining conversational workflows as modular, stateful graphs.

To simulate the conversational ability of a VP, a linear stateful LangGraph workflow is employed, as illustrated in Fig. 6.3.

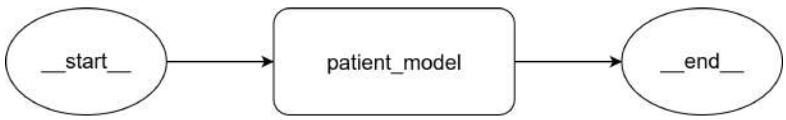


Figure 6.3: Patient Simulation Workflow

Although linear, this workflow is sufficient for the current simulation needs. It can also be easily extended for tool use or additional reasoning nodes without refactoring, if required for future iterations.

The state of a conversation session is stored in a *CustomState* object (see Listing 6.2). It holds all data required for a single conversational turn, including the message history, model parameters, and patient details.

```python
class CustomState(TypedDict):
    messages: Annotated[list[AnyMessage], add_messages]
    model: str
    condition: str
    talkativeness: str
    patient_details: str
```

Listing 6.2: State Definition of the Patient Simulation Workflow

[3] https://www.langchain.com/, accessed on 24 August 2025
[4] https://www.langchain.com/langgraph, accessed on 24 August 2025

Before the student's message is passed to the LangGraph workflow, the system dynamically creates the relevant system prompt for the LLM. This process involves retrieving the appropriate system prompt template from a predefined dictionary and making the necessary substitutions within the text (see Listing 6.3). The adjustment options include factors such as the patient's condition, verbosity, and specific details related to the medical history. These options are further explained in Section 6.3.

This dynamic prompt retrieval ensures that each turn of the conversation is guided by a contextually relevant prompt, tailored to the VP case study.

```python
def get_prompt(patient_condition: str, talkativeness: str, patient_details: str) ->
    ChatPromptTemplate:
    """
    Returns the appropriate prompt template based on the patient's condition and adjusts
     the talkativeness and patient_details appropriately.
    """

    if patient_condition == "schwerhoerig":
        return PROMPTS["schwerhoerig"](talkativeness.capitalize(), patient_details)
    elif patient_condition == "verdraengung":
        return PROMPTS["verdraengung"](talkativeness.capitalize(), patient_details)
    elif patient_condition == "alzheimer":
        return PROMPTS["alzheimer"](talkativeness.capitalize(), patient_details)
    else:
        return PROMPTS["default"](talkativeness.capitalize(), patient_details)
```

Listing 6.3: Dynamic Prompt Retrieval

The feedback generation mechanism is implemented as a separate LangChain workflow. Contrary to the patient simulation workflow, it is stateless and does not retain past messages. Instead, it takes the complete medical history-taking conversation as a parameter to generate the appropriate feedback. By decoupling feedback generation from patient simulation, the design allows alternative configurations — such as different LLMs, tailored prompts, or adjusted model parameters — to be used for evaluation without affecting the patient simulation.

### 6.1.4 Persistence Layer

As specified previously, the long-term storage of anamnesis conversations and patient profiles is handled by a *PostgreSQL* database, with the schema defined using *SQLAlchemy ORM* models. Database access is mediated by FastAPI's dependency injection system, which ensures that each request is handled within an isolated session lifecycle (see Listing 6.4).

```
# Database dependency
def get_db():
    db = SessionLocal()
    try:
        yield db
    finally:
        pass
```

Listing 6.4: Database Dependency

The dependency is injected into API endpoints where persistent data is required, such as retrieving patient profiles before constructing prompts or storing the LLM response before finally closing the database session (see Listing 6.5).

```
# Chat endpoint
@router.post("/chat")
async def chat_with_llm(request: ChatRequest, db: Session = Depends(get_db)):
    # ...

    # Get patient profile from database
    patient_file = db.query(PatientFile).filter(PatientFile.id == request.
     patient_file_id).first()
    if not patient_file:
        return PlainTextResponse("Patient not found", status_code=404)
    patient_details = format_patient_details(patient_file)

    # ...
            try:
                # ...

                # After streaming is complete, store LLM message
                llm_message = ChatMessage(
                    session_id=session.id,
                    role="patient",
                    content=llm_response
                )
                db.add(llm_message)
                db.commit()
            finally:
                db.close()
    # ...
```

Listing 6.5: Database Dependency Injection Example

## 6.2 LLM Selection

To determine the most suitable LLM for Symptex, we pursued a selection process that was completed after examining three different approaches detailed in the following subsections (see Fig. 6.4).

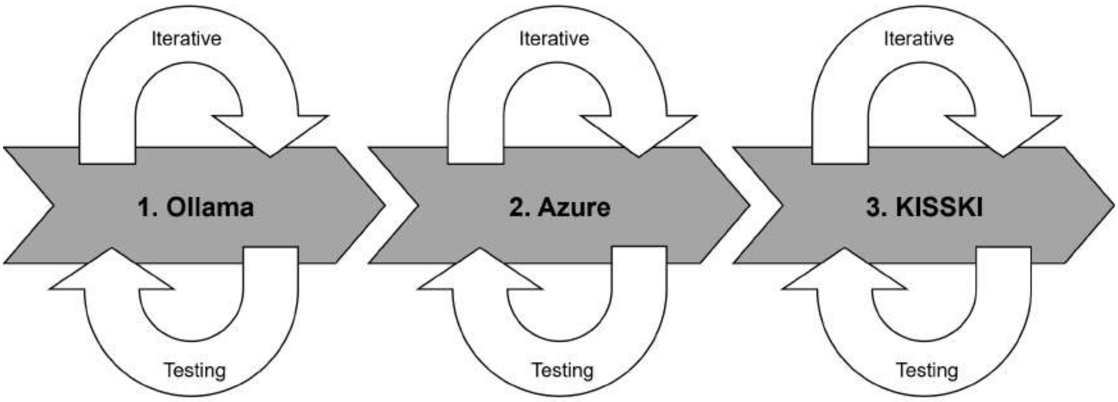


Figure 6.4: LLM Selection Process

Throughout this process, we iteratively tested different configurations, with the evaluation guided by the design goals defined in Section 5.1. Each candidate model was tested with at least four to five comprehensive medical history-taking dialogues for a representative evaluation. Specifically, the following questions were used to assess the models:

**DG1: Patient Language Authenticity**

- Does the model produce human-like responses without sounding robotic?
- Does it speak grammatically correct and fluent German?

**DG2: Conversational Consistency**

- Does the model behave in a manner consistent with a patient persona?
- Does it avoid contradictions across turns?
- Are its emotional expressions appropriate in the conversational context?

**DF3: Effective Feedback**

- Is the generated feedback accurate and pedagogically valuable in the context of medical history-taking?

- Are the cited examples grounded in the conversation rather than hallucinated?
- Does the feedback contain sufficient detail?

**DF4: Responsive User Experience/Performance**

- Is the response latency acceptable, i.e., does the model generate a complete response within approximately ten seconds?

### 6.2.1 Approach 1: Open-Source Models via Ollama

In the initial development phase, we downloaded various open-source LLMs and tested them locally on a standard PC with a single GPU using the Ollama platform [5].

Ollama is an open-source, lightweight framework designed for hosting and running LLMs locally on the user's own hardware. It offers access to a broad selection of popular models, including Llama, Mistral, Deepseek, and Gemma, and simplifies the deployment process of these models through containerization.

The decision to utilize Ollama was motivated by its streamlined setup, the ability to experiment with a wide array of continuously updated models at no cost, and its native integration with LangChain and LangGraph—the core orchestration frameworks used in the Symptex implementation. Being open-source, the framework also offers transparency regarding its internal structure, including (where available) insight into the model architecture and training data—advantages that are often absent in proprietary LLM APIs.

Despite these benefits, significant limitations emerged during testing. The available hardware setup restricted the selection of usable models only to those with up to 7-8 billion parameters. While sufficient for lightweight tasks, these smaller models demonstrated limited configurability via system prompts and subpar performance in handling coherent medical dialogues. Specifically, they frequently failed to maintain consistent role behavior—often reverting to the role of physician despite explicit instructions to act as a patient.

Linguistic performance in German presented an additional challenge. Most models were primarily trained on English datasets, which often resulted in unnatural, grammatically flawed, or poorly translated output in German. This deficiency was particularly noticeable in smaller, distilled model variants and compromised the authenticity of the patient simulations.

Among all tested options (see Table 6.1), only the 8-billion-parameter version of Llama 3.1 demonstrated acceptable results with regard to the aforementioned issues. However, even this model demonstrated suboptimal inference times, with response

[5] `https://ollama.com`, accessed on 16 August 2025

times occasionally exceeding one minute on a PC under load. This issue, while already impractical in single-user testing, would be likely to worsen under concurrent usage, making the current setup unsuitable for scalable use within educational settings.

| Model Name | Parameters |
|---|---|
| Llama 3.1 | 8 billion |
| Llama 3.2 | 3 billion |
| Llama 3 | 8 billion |
| Deepseek R1 (reasoning disabled) | 8 billion |
| Qwen 2.5 | 7 billion |
| Phi4 mini | 3.8 billion |
| Phi3 | 3.8 billion |
| Gemma2 | 9 billion |
| Mistral | 7 billion |
| Llama-3.1-SauerkrautLM | 8 billion |

Table 6.1: List of Evaluated Ollama Models

In conclusion, while Ollama proved useful for rapid prototyping and experimentation with open models, its reliance on the current insufficient hardware and the performance limitations of smaller models made it unsuitable for production use. It should, however, definitely be reconsidered for future deployment if access to dedicated high-performance infrastructure becomes available, particularly as open-source LLMs continue to evolve at a fast pace in terms of quality and multilingual capabilities.

### 6.2.2 Approach 2: Closed-Source Models via Azure OpenAI Service

In the second approach, we accessed closed-source LLMs through the Azure OpenAI Service, which provides hosted access to OpenAI's models (e.g., GPT-3.5, GPT-4) via Microsoft Azure's secure cloud infrastructure. The availability of free academic credits made this option particularly appealing for initial experimentation.

The integration with the service was straightforward, supported by comprehensive documentation and well-defined API endpoints that allowed for seamless connectivity with the Symptex system. During testing, the latest GPT-4 models demonstrated strong coherence and contextual understanding, along with remarkable fluency in German, naturally outperforming the much smaller Ollama models evaluated earlier.

The models further showcased a satisfactory degree of configurability via system prompts, allowing for reliable modulation of behavior, tone, and role adherence. The platform's scalable infrastructure also provided low-latency, high-availability access, making it technically well-suited for real-time interaction in a virtual patient simulation context.

However, a financial limitation remained. Although the initial free academic credits facilitated prototyping, the service only supports pay-as-you-go billing. Despite having funding through a supporting foundation, the lack of prepaid fixed-cost plans and the requirement for early usage-based billing set by the foundation posed administrative challenges. This financial constraint ultimately motivated the search for a more flexible alternative, even though the Azure-hosted OpenAI models already delivered excellent technical performance.

### 6.2.3 Approach 3: KISSKI ChatAI Service

In the third and final approach, the Symptex system was integrated with the ChatAI[6] service provided by KI-Servicezentrum für sensible und kritische Infrastrukturen (KISSKI) [Doo+25]. ChatAI is an LLM-hosting web service that offers free access to a curated set of high-performing open-source LLMs via a secure API. Unlike commercial services, this service imposes no token-based usage limits, and the API's rate limits can be scaled up on request for future deployments. This flexibility eliminates any financial concerns and technical barriers that might otherwise restrict experimentation and future deployment in educational settings.

ChatAI itself is hosted on SAIA[7], KISSKI's Scalable Artificial Intelligence Accelerator, which operates a range of other AI-based tools alongside ChatAI. Notably, the entire infrastructure is located within Germany and guarantees that no user data is stored, which effectively addresses data privacy and compliance concerns for medical students.

During the testing of suitable models available through ChatAI (see Table 6.2), all candidates consistently produced responses within less than ten seconds. This performance meets the aforementioned low-latency requirements for Symptex, ensuring that chat interactions remain fluid.

With the overall performance of most ChatAI models being consistently high, the final selection process was supported by our healthcare domain expert, Johannes Reifenrath, who conducted qualitative assessments of conversational authenticity and clinical adequacy. Based on this evaluation, the *Qwen-3 235B-A22B* model (with reasoning disabled) from Alibaba [Tea25] was identified as the most suitable option, demonstrating superior conversational coherence, adaptability to different questioning styles, and overall humanness during history-taking dialogues.

For the patient simulation use case, this model delivered the most convincing results when configured with the following parameters:

[6] `https://docs.hpc.gwdg.de/services/chat-ai/index.html`, accessed on 23 August 2025
[7] `https://docs.hpc.gwdg.de/services/saia/index.html`, accessed on 23 August 2025

| Model Name | Parameters |
|---|---|
| Llama 3.1 Instruct | 8 billion |
| Gemma 3 Instruct | 27 billion |
| Qwen 3-235B-A22B (reasoning disabled) | 235 billion |
| Qwen QwQ (reasoning disabled) | 32 billion |
| DeepSeek R1 (reasoning disabled) | 685 billion |
| DeepSeek R1 Distill Llama (reasoning disabled) | 70 billion |
| Llama 3.3 Instruct | 70 billion |
| Llama 3.1 SauerkrautLM Instruct | 70 billion |
| Mistral Large Instruct | 123 billion |

Table 6.2: List of Evaluated ChatAI Models

- $temperature = 0.7$, striking a balance between deterministic output and natural variability, thereby enabling more spontaneous and realistic patient behavior.
- $top\_p = 0.8$, introducing controlled randomness that encouraged nuanced yet plausible patient responses.

In the second use case of evaluating student performance, the same model outperformed the alternatives by consistently generating the most structured, detailed, and pedagogically meaningful feedback. Here, the configuration was adjusted to the following:

- $temperature = 0.0$, ensuring deterministic and reproducible assessments without unnecessary variability, which is critical for fairness and reliability in evaluation.

This dual-configuration implementation approach thus enabled the same underlying LLM to be optimized for two distinct purposes: simulating authentic VP interactions on the one hand, and providing consistent, objective evaluation feedback on the other.

However, using ChatAI also comes with limitations. One of the main limitations is that Symptex is inherently dependent on the availability of KISSKI's infrastructure. As a result, Symptex's operations must align with scheduled maintenance windows or potential service downtimes of KISSKI, which may temporarily make Symptex unavailable.

Another limitation is the restricted range of available models. Unlike platforms such as Ollama or Azure, which offer users a wider and frequently updated catalog of models, ChatAI has a more compact list of supported models that is managed and updated centrally by the KISSKI organization. While this curated model selection ensures quality, stability, and compliance, it does limit the flexibility for experimentation, especially when new models are released and require assessment.

## 6.3 Prompt Engineering

To ensure that an LLM performs as intended for a specific use case, it is essential to define a conclusive and effective system prompt. Given the inherently nondeterministic nature of LLM output, this process—commonly referred to as *prompt engineering*—is an iterative process requiring systematic testing and refinement, targeting the same design goal-guided questions as the model selection process. For Symptex, two separate prompt engineering processes were conducted for the simulation of authentic VP conversations and the evaluation of a student's medical history-taking performance.

### 6.3.1 Patient Simulation

The primary objective of the prompt engineering process for Symptex's VP simulation was to ensure that the LLM could accurately embody an assigned VP persona. To achieve this, the initial system prompt contains the following detailed patient-specific background information relevant for the history-taking process:

- Name
- Birthdate/Age
- Height
- Weight
- Gender (medical and self-identified)
- Ethnic Origin

It further comprises information that needs to be elicited by the physician:

- Course of Disease: Chronological development of the presenting condition.
- Past Medical History: Notable health issues prior to the current case.
- Medication: Current prescriptions or over-the-counter drugs.
- Allergies: Relevant allergic reactions.
- Family History: Health conditions present in the patient's close relatives.
- Cardiovascular Risk Factors: e.g., hypertension, smoking habits.
- Social/Occupational History: Additional information about the patient's occupation, living situation, and relevant social factors.

To strengthen the model's adherence to these characteristics, *few-shot prompting* was employed using in-context learning techniques as recommended by Brown et al. [Bro+20]. This involved including several realistic examples of expected doctor-patient exchanges within the prompt, helping to reinforce consistent tone, linguistic style, and interaction patterns (see Listing 6.6).

```
# Few-shot example
HumanMessagePromptTemplate.from_template("Wissen Sie was passiert ist?"),
AIMessagePromptTemplate.from_template("Ich ... *kratzt sich den Kopf* ... ich weiss es
    nicht ..."),
HumanMessagePromptTemplate.from_template("Welche anderen Erkrankungen haben Sie?"),
AIMessagePromptTemplate.from_template("Oh, uh ... *Schweigen*")
```

Listing 6.6: Few-shotting Examples for the Alzheimer's prompt

These examples demonstrate typical patient behaviors, such as hesitations, uncertainty, or silence, and provide the model with concrete cues on how to simulate cognitive impairments and emotional nuances in a natural way. By incorporating such exemplary exchanges, the prompt not only reinforces the patient's persona but also encourages varied and human-like responses that align with the authenticity goals of Symptex.

Each prompt iteration was tested in four to five complete medical history-taking sessions with a similar questioning style to evaluate behavioral consistency. Once messages were observed that deviated from the design goals, the system prompt was adjusted accordingly and re-tested.

#### Observations

Throughout this prompt engineering process, several recurrent issues emerged, particularly among smaller models (7-8B parameters):

- **Role Confusion**
  Models occasionally deviated from their assigned patient persona, for instance, by switching to the physician role to consult the user or simulating an entire doctor-physician conversation on their own without user input.
- **Limited German Fluency**
  Outputs frequently included grammatical mistakes, overly literal translations from English, or sometimes even terms from the model's native training language (e.g., Chinese or English).
- **Excessive Verbosity**
  Models often responded in long paragraphs, inconsistent with realistic behavior, especially for elderly or cognitively impaired VP profiles.

- **Uniform Response Lengths**
  The responses of some models tended to follow a predictable structure and fixed length, reducing the perceived conversational authenticity.

- **Medical Knowledge Leakage**
  Patients occasionally used clinical terminology, named exact medication names or dosages, or displayed knowledge that was inconsistent or implausible for their background.

A further observation concerned the hallucination of unspecified patient details. While often considered a flaw in LLMs, it was partially accepted—and even welcomed—in this context, as it is neither feasible nor desirable to capture the complete history of a human patient. In addition, real patients may not always recall their medical history accurately, resulting in inconsistent or partially incorrect information. As such, a controlled degree of hallucination can increase authenticity and contribute to the authenticity of the training experience.

Moreover, the responsiveness to system prompt modifications varied greatly between different models, ranging from complete disregard for added instructions to high sensitivity to minor changes.

#### Prompt Design

Considering the issues that appeared throughout the process, the prompt design for the VP simulation was mainly guided by the following two key criteria:

- **Adherence to the Patient Role**
  The chatbot must consistently act in character as the assigned patient. This involves:
  - Responding only with information consistent with the predefined medical history.
  - Ignoring or deflecting prompts unrelated to a typical patient interview.
  - Being unaware of their own diagnosis.

- **Realistic and Varied Responses**
  Responses should reflect authentic human behavior, including:
  - Always responding in fluent German.
  - Occasionally using hesitations, filler words, as well as facial expressions or gestures described through written cues (e.g., *shrugs*).
  - Exhibiting emotional reactions to sensitive questions or physical discomfort.

  – Avoiding medical terminology that a typical patient would be unlikely to know (e.g., no Latin diagnosis names or exact drug dosage names, unless contextually plausible).

The resulting prompt thereby aims to preserve natural language patterns and human-like variability while also maintaining character consistency (see Appendix 8.2).

### 6.3.2 Performance Evaluation

The second use case for prompt engineering was the automated evaluation of students' medical history-taking performance. Here, the goal was to design a prompt that could guide the LLM to score student performance with a structured and pedagogically meaningful justification.

#### Evaluation Framework

Before designing the evaluation prompt, it was necessary to establish a robust theoretical foundation for the evaluation criteria. This foundation should draw from the aforementioned established clinical communication frameworks that incorporate principles of patient-centered communication, such as the Calgary-Cambridge Guide and the Four Habits Model (see Subsection 3.1.1). These frameworks not only focus on the *content* of the medical interview but also emphasize the *process* through which information is elicited and building a relationship with the patient. In particular, they highlight elements that are critical for promoting trust and obtaining accurate, relevant clinical data, including structured yet flexible information gathering, active listening, verification of understanding, and summarization.

To translate these theoretical principles into a standardized and measurable assessment, the *Clinical Reasoning Indicators-History Taking-Scale* (CRI-HT-S) was adopted [Für+20]. Developed through an empirical process based on qualitative observations, the CRI-HT-S quantifies observable behaviors that reflect clinical reasoning during history-taking. The instrument consists of eight items rated on a five-point Likert scale (1 = does not meet the criterion, 5 = fully meets the criterion), allowing for a maximum score of 40 per consultation. The assessment items are described as follows [Für+20]:

1. **Taking the lead in the conversation**
   The student takes control of the interview in order to get the required information.

2. **Recognizing and responding to relevant information**
   The student shows that they recognize relevant information by, e.g., responding with obvious interest to it.

3. **Specifying symptoms**
   The student makes targeted inquiries to capture the symptoms in more detail, which they consider to be important.

4. **Asking specific questions that point to pathophysiological thinking**
   The student's questions indicate that they are considering specific causes for certain symptoms.

5. **Putting questions in a logical order**
   The student asks the questions in a logical order and not according to a list.

6. **Checking with the patient**
   The student assures themselves by checking with the patient that their clinical thinking is based on correct information.

7. **Summarizing**
   The student summarizes their collected information aloud as soon as they have reached a meaningful level.

8. **Collected data and effectiveness of the conversation**
   The student collects sufficient, high-quality data at a reasonable speed.

These eight items can be conceptually mapped to the stages and behaviors described in both the Calgary-Cambridge Guide and the Four Habits Model. For example, **taking the lead in the conversation** reflects the Calgary-Cambridge emphasis on initiating the session and corresponds with the Four Habits' focus on investing in the beginning. Similarly, **recognizing and responding to relevant information** and **specifying symptoms** align with structured information gathering and the elicitation of the patient's perspective. Additionally, **asking pathophysiologically oriented questions** mirrors the depth of reasoning emphasized in both frameworks, whereas **putting questions in a logical order** relates to their shared concern for maintaining a coherent conversational flow. Finally, **checking with the patient**, **summarizing**, and **data gathering and efficiency** correspond to later consultation stages, such as building the relationship, closing the session, and balancing information completeness with efficiency.

### Prompt Design

Based on this theoretical and empirical foundation, the evaluation system prompt was explicitly formulated to mirror the CRI-HT-S structure in German, ensuring that the LLM's output addresses each criterion systematically.

In addition to assigning a score for each item, the prompt instructs the LLM to provide evidence-based justifications, referencing specific excerpts from the dialogue. Moreover, it is also prompted to give constructive, actionable suggestions for improvement. By combining quantitative scoring with qualitative feedback, this approach serves both performance measurement and learning purposes.

In sum, the complete evaluation prompt ensures that the assessment not only measures the retrieval of correct information during the medical history-taking conversation but also assesses the reasoning processes and interpersonal communication skills that significantly influence the quality and outcome of doctor-patient interactions.

# 7 Evaluation

The primary aim of this evaluation is to examine the extent to which Symptex fulfills its intended educational, functional, and non-functional goals within the ILuVI framework. To gather a comprehensive understanding of its effectiveness, a user study was conducted with medical students—the primary target group—using both qualitative and quantitative data collection methodologies, aligning with the evaluation objectives derived from the broader research aims. The generated conversations and the performance feedback from the user study are subsequently evaluated together by graduate physicians acting as objective experts.

This chapter outlines the evaluation's objectives and methodology, followed by a detailed presentation of the findings and a discussion of their implications.

## 7.1 Objectives

The following evaluation objectives are defined to gain the relevant insights required for assessing Symptex:

- **O1: Understanding User Acceptance**
  To explore user perceptions regarding the use of Symptex—and LLM-based tools in general—in medical education. This includes identifying perceived strengths, limitations, and expressed sentiments (positive or negative) about integrating such technologies into the learning process of communication training.
- **O2: Assessing Non-Functional Requirement Fulfillment**
  To determine whether the system meets key non-functional requirements from the user perspective, focusing on the following dimensions:
  - *Fluency:* Is Symptex capable of understanding prompts in German and generating grammatically correct responses in German?
  - *Accuracy:* Are the responses generated by Symptex medically plausible and aligned with the predefined characteristics of the simulated patient case?
  - *Humanness:* Does the chatbot exhibit human-like behavior and communication patterns similar to real patient interactions?

  – *Pedagogical Feedback:* Does Symptex offer meaningful, individualized feedback that can help students reflect on and improve their communication approach?
  – *Conversational Dynamics:* Does Symptex send contextually coherent responses relevant to the user's questions and to the overall conversation history?
  – *Latency:* Are the responses generated within a reasonable amount of time?

- **O3: Assessing Educational Value**
  To understand how medical students perceive the pedagogical value of Symptex in enhancing their communication skills during the medical history-taking process, especially compared to current practices that use actors for training.

## 7.2 Methodology

To address these objectives, we employed a two-part mixed-methods evaluation, integrating both subjective and objective data sources. Subjective data from medical students captured experiences from the learner's perspective as the primary end-users, while recognizing that such impressions may be biased by novelty effects or individual expectations. To counterbalance this, graduate physicians were included to provide objective assessments of Symptex's clinical plausibility, consistency, and feedback quality. Together, this complementary design ensured a more comprehensive and balanced evaluation of Symptex.

The following subsections outline the participant backgrounds, evaluation procedure, and respective survey instruments.

### 7.2.1 Background of Participants

The subjective evaluation involved five medical students who were recruited via email outreach by the senior physician of the Department of Dermatology at TUM MRI, Prof. Dr. Dr. med. Alexander Zink, MPH, MBA: three females, one male, and one individual who preferred not to disclose their gender. They represented varying levels of medical training stages: the least advanced was a medical student in their sixth semester. Two participants were in the ninth, one was in the tenth, and the most experienced had recently completed their doctoral degree in medicine. This diversity allowed us to capture a broad spectrum of user perspectives and identify potential usability or learning challenges across different levels of medical expertise. The participants further fell into the 18-24 and 25-34 age brackets, further contributing to the heterogeneity of the sample.

The objective evaluation was conducted with two female graduate physicians and one male graduate physician, all of whom were contacted by Johannes Reifenrath.

### 7.2.2 Evaluation Procedure

The detailed structure of the two-part mixed-methods study is illustrated in Figure 7.1 and further outlined in the following.

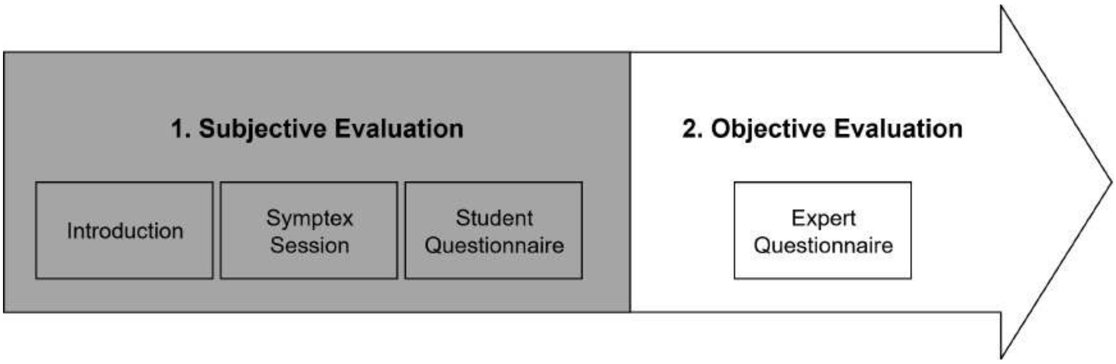


Figure 7.1: Symptex Evaluation Procedure

The subjective evaluation was conducted as a 30-minute semi-structured interview with one medical student at a time via Zoom, with each session recorded with the participant's consent. It followed a three-part structure:

1. **Introduction**
   First, we asked each participant to self-assess their confidence in conducting medical history-taking. Following this, an overview of the research context was provided, highlighting the current limitations in communication training within medical curricula—particularly the reliance on actors for simulated patient encounters. The concept of ILuVI was then introduced, explaining how Symptex fits within this framework as an integrated tool to simulate the medical history-taking process.

2. **Medical History-Taking Session using Symptex**
   The primary component of the study consisted of two parts. First, participants were instructed to take the history of a patient with advanced Alzheimer's through a shared screen of a preconfigured PC running Symptex within its web-based chat interface. Before starting the medical history-taking session, they were given sufficient time to familiarize themselves with both the user interface and the patient case. Once they confirmed they were ready, remote control access was granted so

they could initiate the chat-based interaction independently. No time limit was imposed; participants were allowed to take as long as they needed to elicit the necessary information. Second, after completing the chat, participants were asked to generate personalized feedback by clicking the designated evaluation button and then reviewing the feedback provided.

3. **Student Questionnaire**
   Upon completing the Symptex session, participants were asked several survey questions to quantify their experiences using Symptex.

Throughout these five interviews, both qualitative and quantitative data were collected. Qualitative insights were obtained in two ways. First, participants were instructed to employ a think-aloud protocol during their interaction with Symptex. This protocol encouraged participants to verbalize their thoughts in real time, allowing us to identify implicit reasoning, usability barriers, and spontaneous user perceptions. Second, they were asked to respond to a set of open-ended questions in the final section of the questionnaire. In addition to these open-ended items, the questionnaire also contained 5-point Likert scale questions to collect complementary quantitative data.

To supplement the subjective insights gathered via interviews and questionnaires, an objective expert evaluation was subsequently conducted. Through an expert questionnaire, three graduate physicians were asked to review Symptex's responses in the five medical history-taking sessions with the students for overall medical accuracy and factual correctness. They also assessed the performance of the students themselves and evaluated the generated personalized feedback. This step ensures a more robust assessment of Symptex's reliability as an educational tool.

The design of both student and expert questionnaires is described in the following.

### 7.2.3 Student Questionnaire Design

The student questionnaire includes 21 questions in total, and is structured into three sections:

1. **TAM-based Questions**
   The survey begins with nine 5-point Likert-scale questions that are based on the TAM model to measure the students' technological acceptance of Symptex as a technological tool.

2. **NFR-based Questions**
   The second part comprises seven 5-point Likert-scale questions focusing on the fulfillment of Symptex's non-functional requirements.

3. **Open Questions**
   The final section gathers qualitative data regarding students' experiences with Symptex and its educational value through five open-ended questions.

In the following, a more detailed description of each questionnaire section is provided.

**Section 1: TAM-based Questions**

The first part of the survey aims at achieving the first evaluation objective O1 and is theoretically grounded by the technology acceptance model (TAM), a well-established evaluation framework proposed by Davis et al. [Dav89; DBW89]. TAM is particularly relevant for early-stage prototypes where long-term user data is unavailable, making it highly applicable to this study. The authors identify the following three constructs as the primary factors that influence user acceptance of technology:

- **Perceived Usefulness (U)**
  This refers to the extent to which a person believes that using a particular system will enhance their performance in a specific task [Dav89]. In the context of this study, U captures whether the participants believe that interacting with Symptex helps them conduct medical history-taking more accurately, efficiently, and confidently.

- **Perceived Ease of Use (EOU)**
  This construct describes the degree to which individuals believe that using a system will require little cognitive or practical effort [Dav89]. In this research, EOU reflects whether students find Symptex intuitive, accessible, and easy to engage with during a clinical simulation.

- **Behavioral Intention (BI)**
  Behavioral Intention refers to a user's expressed commitment to using a system, which serves as a key predictor of actual usage [DBW89]. In this context, it captures the extent to which medical students intend to use Symptex for anamnesis training.

Originally, TAM also includes a fourth factor called **Attitude Toward Using (A)**. This factor is defined as the user's overall emotional evaluation of the system, whether they like or dislike using it [DBW89]. *A* mediates the effects of *U* and *EOU* on *BI*, as illustrated in Fig. 7.2. However, in their ensuing empirical research, the authors found that *A* only partially mediated these effects, and that *U* often directly influences *BI*, bypassing *A* entirely [DBW89].

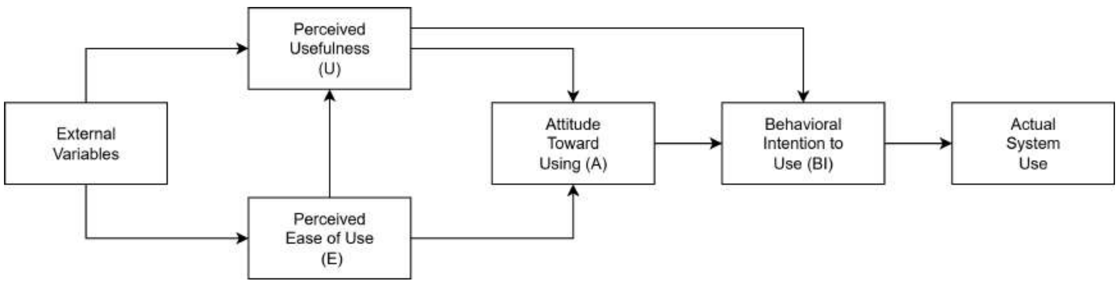


Figure 7.2: Technology Acceptance Model (TAM), adapted from Davis et al. [DBW89]

Consequently, the theoretical constructs *U*, *EOU*, and *BI* have since been established as the core predictors in TAM-based evaluation models, which are reflected in the first nine questions of the student survey (see Table 7.1).

| ID | Question | Type |
|---|---|---|
| U1 | Symptex helps me to practice medical history-taking efficiently. | Likert scale |
| U2 | Symptex helps me handle more complex medical history-taking scenarios (e.g., dementia) with more confidence. | Likert scale |
| U3 | I became more aware of my strengths and weaknesses during patient medical history taking through this simulation. | Likert scale |
| EOU1 | The Symptex interface is intuitive and easy to navigate even without prior knowledge. | Likert scale |
| EOU2 | The conversation flows without delays or technical errors. | Likert scale |
| EOU3 | Recovering from potential misunderstandings (e.g., typos) is straightforward. | Likert scale |
| BI1 | I would recommend Symptex to other students for practice. | Likert scale |
| BI2 | I would use Symptex regularly for practice. | Likert scale |
| BI3 | I prefer Symptex over manual role-playing for history-taking practice. | Likert scale |

Table 7.1: Student Questionnaire Section 1: TAM-based Questions

**Section 2: NFR-based Questions**

The second section consists of seven 5-point Likert-scale questions that examine the fulfillment of Symptex's key non-functional requirements (Objective O2). The question IDs are mapped as follows to the non-functional requirement that is assessed (see Table

7.2):

**NFR1: Patient Simulation Language**

- **A1** assesses whether the chatbot's responses aligned with the expected medical profile of the Virtual Patient (VP) (Accuracy).
- **F1** targets the linguistic fluency of the chatbot in German (Fluency).
- **H1** and **H2** address the humanness and authenticity of the interaction (Humanness).

**NFR2: Conversational Dynamics**

- **CD1** evaluates whether Symptex's responses were logically linked to the questions asked.
- **CD2** measures the overall information consistency over the course of the dialogue.

**NFR4: Latency**

- **L1** assesses latency, i.e., how quickly Symptex generated the responses.

It is important to note that NFR3 (Educational Valuable Feedback) is assessed through the open-ended questions of the survey.

| ID | Question | Type |
|---|---|---|
| A1 | Symptex's responses are medically plausible and match the medical condition of the VP. | Likert scale |
| F1 | Symptex is fluent in German, and I understood everything it said. | Likert scale |
| H1 | Symptex's responses feel realistic and human-like. | Likert scale |
| H2 | I could communicate with the patient just like in real-world scenarios. | Likert scale |
| CD1 | Symptex responses were coherent with my questions. | Likert scale |
| CD2 | The provided information stayed consistent throughout the conversation. | Likert scale |
| L1 | The response generation latency was acceptable. | Likert scale |

Table 7.2: Student Questionnaire Section 2: NFR-based Questions

#### Section 3: Open Questions

In the last part of the survey, the students are once again invited to openly share their perspectives, complementing their earlier think-aloud commentary during the Symptex interaction (see Table 7.3). The first question, Q1, serves as a reflective prompt, revisiting the opening question to assess whether the participant's confidence in conducting medical history-taking had increased, decreased, or remained unchanged after using Symptex. Questions Q2 and Q3 encourage participants to voice additional positive or negative impressions that may not have been expressed during the live interaction. Question Q4 directly addresses the chatbot's pedagogical effectiveness (Objective O3), while also probing the non-functional requirement regarding the usefulness and relevance of Symptex's feedback. Finally, Q5 offers participants the chance to contribute any final comments or suggestions.

| ID | Question | Type |
|---|---|---|
| Q1 | How confident are you feeling about conducting medical history-taking? | Open |
| Q2 | What felt realistic? | Open |
| Q3 | What felt unrealistic? | Open |
| Q4 | What aspects of patient communication/anamnesis did the chatbot help you practice/understand better? | Open |
| Q5 | Any final thoughts or remarks? | Open |

Table 7.3: Student Questionnaire Section 3: Open Questions

### 7.2.4 Expert Questionnaire

The expert questionnaire was created by Johannes Reifenrath. It was designed to collect insights from graduate physicians acting as experts regarding both the authenticity of the VP simulation and the pedagogical quality of the personalized feedback provided by Symptex. The survey consists of two main sections, combining 5-point Likert scales with optional free-text prompts.

#### Section 1: Evaluation of the Patient Simulation

The first section assesses evaluation objective O2 (Assessing Non-Functional Requirement Fulfillment) from an objective expert view, focusing on the consistency of the responses throughout the conversation, as well as the medical accuracy of the represented symptoms of the simulated Alzheimer's patient.

The opening item (Q1.1) is a 5-point Likert-scale question evaluating the overall patient simulation authenticity of Symptex's responses, ranging from 1: "not realistic at all" to 5: "very realistic".

This is followed by a checklist-based evaluation of core Alzheimer-related symptoms (Q1.2) sourced from the Amboss[1] platform, allowing experts to indicate the presence and authenticity of each symptom:

- *Yes:* The symptom was realistically portrayed.
- *No:* The chatbot attempted to portray the symptom realistically, but did not succeed.
- *X:* The symptom was not present in the conversation, making it impossible to evaluate its portrayal.

The covered symptoms included the following:

- **D1:** Disorientation to place
- **D2:** Disorientation to time
- **D3:** Disorientation to person
- **D4:** Disorientation to situation
- **M1:** New memory disorder
- **M2:** Old memory disorder
- **A1:** Attention deficit disorder
- **N1:** Neuropsychiatric: Mutism
- **N2:** Neuropsychiatric: Hallucinations
- **N3:** Neuropsychiatric: Depressive symptoms/Decrease in activity and motivation/Apathy
- **S1:** Speech disorder: Apraxia/Alexia/Acalculia
- **L1:** Reduced capacity for activities of daily living

A further 5-point Likert-scale question (Q1.3) measures the consistency of the Symptex's responses across the interaction, ranging from 1: "not consistent at all" to 5: "always consistent".

Finally, Q1.4 provides an optional free-text comment field, allowing experts to describe any particularly noteworthy aspects or anomalies in the Symptex's behavior.

[1] `https://www.amboss.com/de`, accessed on 24 August 2025

#### Section 2: Evaluation of the Feedback Quality

The second section addresses evaluation objective O3 (Assessing Educational Value), examining whether the feedback provided to students was clinically sound and pedagogically valuable.

Experts are presented with the same eight clinical communication indicators used in the student evaluation (based on the CRI-HT-S; see Section 6.3.2). For each indicator, they are asked to:

1. **F1:** Evaluate the performance of the medical student on a 5-point scale.
2. **F2:** Indicate whether they fully agreed with Symptex's assessment of the student (yes/no).
3. **F3:** Provide a short justification if they disagreed, specifying missing or superfluous points in the chatbot's feedback.

This structure allows for both quantitative benchmarking of Symptex's evaluative accuracy and qualitative insights into specific strengths and limitations from an expert perspective.

## 7.3 Results

All five medical history-taking dialogues, as well as their respective performance feedback, can be reviewed in the Appendix 8.2. In the following, the results from both the student and expert questionnaires are presented.

### 7.3.1 Subjective Results: Student Questionnaire

#### Section 1: TAM-based Questions

As visualized in the stacked bar chart (see Fig. 7.3), the *Perceived Usefulness* items received predominantly positive responses. All three items (U1-U3) showed strong agreement, with the majority of participants selecting either "Agree" or "Strongly Agree".

The items assessing *Perceived Ease of Use* (EOU1-EOU3) also received high ratings, with most responses falling into the upper two agreement categories.

The *Behavioral Intention* construct, captured through items BI1-BI3, showed slightly more variation. While BI1 ("I would recommend Symptex to others") exclusively received positive ratings from all participants, items BI2 ("I would use Symptex regularly") and BI3 ("I prefer Symptex over role-playing") revealed some neutral or less enthusiastic responses.

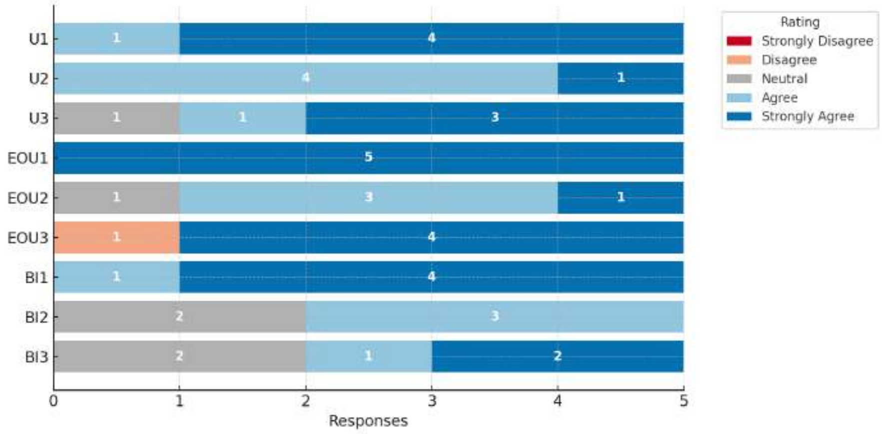


Figure 7.3: Overview of TAM-based Question Results

**Section 2: NFR-based Questions**

The ratings for the first NFR-based question, A1 (Accuracy), were consistently positive (see Fig. 7.4). One participant provided a neutral rating, noting that the patient occasionally responded with clarity despite previously displaying confusion, which was perceived as inconsistent.

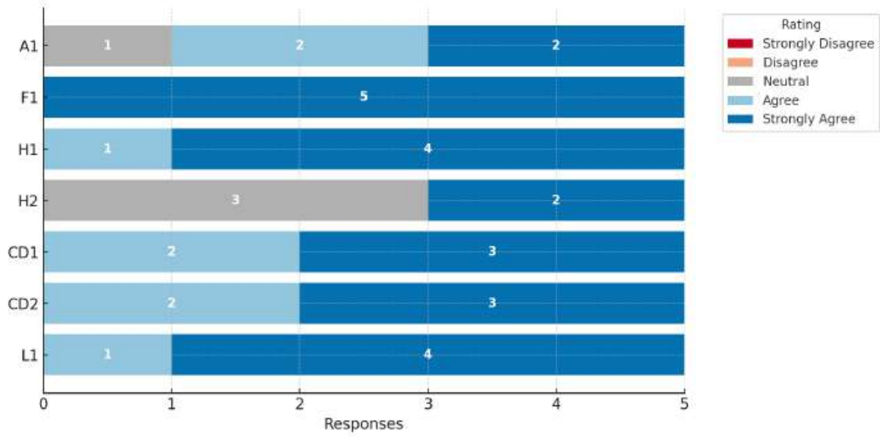


Figure 7.4: Overview of NFR-based Question Results

All respondents further expressed strong agreement with F1 (Fluency).

For H1 (Humanness), most participants perceived the responses as realistic and emotionally appropriate. However, when asked whether they could communicate with the patient just like in a real-world scenario in H2, three participants replied with a neutral rating. These participants noted the absence of visual cues and the unusual feeling to be able to take their time before responding, which contrasts with the time constraints of fast-paced real-world consultations.

Items CD1 and CD2 (Conversational Dynamics) received uniformly high ratings, with responses concentrated in the two highest agreement categories.

The final NFR-based question, L1 (Latency), was evaluated very positively, with most participants strongly agreeing that the system's response times were acceptable.

**Section 3: Open-ended Questions**

The participant's responses to the open-ended prompts (Q1-Q5) provided more qualitative insights into the user perception of the implementation.

For the first open question regarding their current confidence in taking medical histories, the participants reported a range of effects on their confidence after using Symptex. One student noted increased confidence, particularly due to the detailed feedback and structured suggestions offered by the chatbot—elements often missing in traditional training with standardized patients. Three participants mostly felt that their confidence remained unchanged and that an increase in confidence would require more training runs with Symptex. One participant reported reduced confidence after interacting with the patient scenario in Symptex. Although initially expressing good confidence in conducting medical history-taking, this participant noted experiencing unexpected challenges when handling patients with greater complexity, which led to a decrease in confidence afterwards.

Responses to the second and third open questions (Perceived Authenticity of the Patient Interaction) consistently emphasized the realistic behavior of the VP, particularly in terms of emotional nuance and nonlinear communication patterns. All students reported having experience with confused older patients and felt that Symptex effectively reflected authentic confused patient behavior. They pointed out the realistic way the patient frequently asked unprompted questions and showed confusion through written facial expressions, making the experience resemble real-world interactions, particularly with elderly or cognitively impaired individuals. Negative authenticity perceptions of the chat-based format included the lack of visual cues, which made it difficult for some participants to imagine the patient and fully immerse themselves in the situation. One participant mentioned that while the low-pressure environment permitted reflection and response formulation without a time limit, it did not accurately represent the fast-paced history-taking process in the real world, leading to feelings of unreality ("[...] in reality, one must normally act quickly and directly respond."). Another participant further remarked that some of Symptex's responses seemed overly polite ("[...] for example, at one point she said "I really want to help you", which was a bit too much."), which reminded them of similar issues associated with ChatGPT.

Responses to the fourth question—which invited participants to reflect on what aspects of patient communication Symptex helped them better understand—predominantly emphasized the value of the feedback component. Four participants generally appreciated the detailed yet well-structured format of the feedback, noting that it was presented in an objective and non-judgmental manner. Particular aspects of the feedback that were frequently mentioned as helpful included the guidance on structuring the anamnesis, the importance of regular summarization, the recommendation to use probing follow-up questions, and the reinforcement of empathetic responses to patient statements. One participant further noted that Symptex prompted them to consider more complex patient scenarios, which are often not addressed in traditional teaching formats.

The final question, which asked for open remarks and reflections, received four positive responses, with one participant having no additional comments. The other four participants highlighted the enjoyable and innovative nature of the chatbot ("Super tool, feedback was great, has potential!", "Very cool tool, more innovative than Moodle for example [...]"), contrasting it favorably with conventional actor-based simulations. The detailed feedback was once again emphasized as a key strength. One participant specifically praised the intuitive chatroom interface, which required no learning curve and enabled immediate engagement with the system ("[...] no need to learn a new website or software, one can directly start practicing [...]). Further suggestions included the integration of speech recognition to more closely simulate real-world history-taking, which is typically conducted verbally. Finally, one participant recommended integrating Symptex as a mandatory component in the medical curriculum. They expressed that, without this institutional support, students might not fully utilize the tool on their own, as it does take some effort to complete a medical history-taking session ("[...] it basically is a good tool, but in order for them to actually use it, I think you need to force them to use it, like Moodle, for instance as a prerequisite to take the exam [...]").

### 7.3.2 Objective Results: Expert Evaluation

Since all medical history-taking conversations were conducted using identical Symptex configurations and the same patient case, the expert evaluation results from the three experts for each of the five conversations are aggregated, resulting in a total of 15 expert assessments presented collectively in the following.

#### Section 1: Evaluation of the Patient Simulation

The first item (Q1.1) assessed the overall authenticity of the patient simulation. Ratings were consistently positive, with responses evenly distributed between 3 (“Realistic to an extent”), 4 (“Mostly realistic”), and 5 (“Very realistic”).

The second item (Q1.2) assessed the realistic portrayal of core Alzheimer’s-related symptoms. An overview of the answers is illustrated in Figure 7.5. For clarity purposes, missing answers or answers that indicated that the symptom could not be assessed in that conversation were omitted.

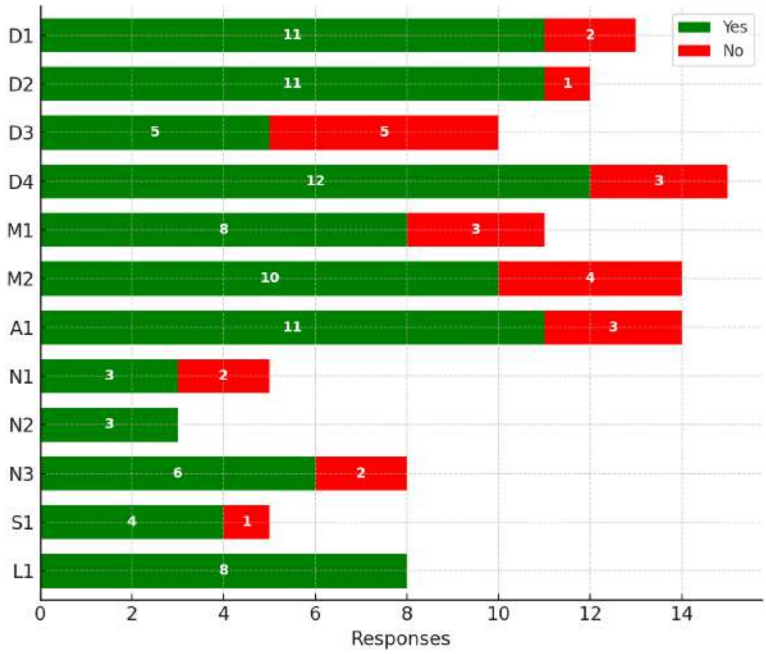


Figure 7.5: Overview of Question Q1.2 Results

Disorientation to place (D1), time (D2), and situation (D4) was consistently identified as realistically portrayed across conversations. Disorientation to person (D3), however, was less frequently observed. In the few instances where it was noted, agreement and disagreement regarding its authenticity were evenly balanced. Both new (M1) and old (M2) memory disorders, as well as attention deficits (A1), were largely rated as realistically portrayed. In contrast, overall neuropsychiatric symptoms (N1, N2, N3) were mostly judged as not encountered within the conversation; however, when observed, they were mostly realistic. Apraxia, related instrumental deficits (e.g., alexia, acalculia) (S1), and impairments in activities of daily living (L1) behave similarly.

The item Q1.3 measured the content consistency of Symptex’s responses across the anamnesis. The answers to this question were predominantly positive, with most evaluations at 4 (“Mostly consistent”), followed by several at 5 (“Always consistent”) and a single rating of 3 (“Consistent to an extent”).

The final optional question, Q1.4 (What was particularly noticeable in the chat), received a response from only one expert. The expert expressed that the VP should not be able to rate the pain on a scale from one to ten, as this scale would be too complex for the patient.

**Section 2: Evaluation of the Feedback Quality**

The second section of the expert questionnaire focused on the accuracy and appropriateness of the personalized feedback Symptex provided to the student participants of the evaluation.

Unfortunately, at least five expert surveys contained conflicting answers to this section that appeared to stem from misinterpretation. Some experts seemed to have mistaken the 5-point Likert scale for the German school grading system, while others accidentally evaluated Symptex's conversational responses rather than focusing on the generated performance feedback, which was the intended target of this section. Since it was not possible to clearly determine which questionnaires remained fully valid for analysis, and in order to avoid introducing noise, only a fraction of available comments suggesting improvements for the personalized feedback of Symptex (F3) were analyzed in detail. The following summarizes these qualitative remarks for each CRI-HT-S category.

1. **Taking the lead in the conversation**
   Three of the completed surveys explicitly disagreed with Symptex's feedback regarding the recommendation that the student should have taken more time to ask additional open-ended questions, further elaborating that such a strategy is not clinically practical in the context of a patient with Alzheimer's disease. One response noted that typographical errors should not influence scoring. Another response also found the sudden subject shifts flagged by Symptex justifiable, as forced topic changes are sometimes necessary for thorough history-taking.

2. **Recognizing and responding to relevant information**
   No remarks were provided, suggesting overall alignment with Symptex's feedback.

3. **Specifying symptoms**
   Three responses found the chatbot's critique of this aspect to be overly strict, arguing that attempting to specify symptoms in late-stage Alzheimer's patients is often clinically impractical or of limited diagnostic value. Additionally, one response suggested that a more visually comprehensible scale for assessing experienced pain may be easier for this type of patient to understand.

4. **Asking specific questions that point to pathophysiologic thinking**
   Similar to the previous point, experts noted that causality-oriented questioning can be of limited relevance, given the cognitive impairments that come with advanced Alzheimer's dementia.

5. **Putting questions in logical order**
   Three completed surveys stressed that an optimal questioning sequence is context-dependent and should evolve dynamically with the patient's responses. One recommended that Symptex should also consider the overall conversational flow for the assessment of this category.

6. **Checking with the patient**
   Two responses remarked that this is not clinically practical for a patient who exhibits Alzheimer's symptoms. One further mentions that this aspect is good practice, but that time constraints rarely allow for it in real-world scenarios.

7. **Summarizing**
   While experts generally agreed on the value of summaries, five responses further noted that they can be omitted when time is limited or when the patient's cognitive impairment hinders meaningful understanding.

8. **Data gathering and efficiency**
   One response describes that, even with extended time, fully clarifying the circumstances of this patient's accident would likely remain difficult due to cognitive limitations. They advised that Symptex should explicitly take these health challenges into account when evaluating the completeness and efficiency of the information elicitation process.

While this overview focuses on suggestions for improvement, it is important to note that the analyzed section of the questionnaire (F3) was completed only when the experts did not fully agree with the feedback. While it was not possible to quantify the exact level of consensus due to the aforementioned noise in the data, it was clear that the majority of the experts generally agreed with Symptex's feedback, with some justifications still including few comments such as "Very good feedback, [...]" and "[...], otherwise good feedback".

## 7.4 Discussion

The evaluation synthesizes subjective student perceptions and objective expert ratings to assess whether Symptex fulfills its educational, functional, and non-functional goals.

Overall, both student participants and experts perceived the prototype as a promising tool for medical communication training. Nevertheless, several limitations and areas for refinement were identified within the scope of the evaluation objectives, which also inform the broader research questions of this work.

### 7.4.1 O1: User Perceptions and Acceptance

Students found Symptex to be both useful and easy to use, which indicates that the familiar chat interface is straightforward to learn. In an educational setting, this suggests that the initial implementation costs are low: the familiar chat format and responsive streaming help reduce onboarding challenges, while session states prevent confusing resets or loss of context. The results further indicate that the prototype reaches a baseline of perceived pedagogical value. While one student reported an immediate increase in confidence in conducting history-taking, three others indicated that such confidence would likely develop further after a few additional practice sessions. This effect can be attributed both to the value of training with a simulated VP and to the structured feedback component of Symptex, which further reinforces learning outcomes.

At the same time, behavioral intention scores reveal some hesitation among students regarding long-term or regular use of Symptex. While all students expressed willingness to recommend the system to peers, only three saw it as a tool they would use regularly. The other two students pointed out that, despite Symptex's high usability, consistent, unsolicited use might be less common due to the commitment required to complete a medical history-taking session. This more cautious perspective suggests that sustained use of Symptex will largely depend on how well it is integrated into the curriculum and endorsed by the institution—factors that are critical for the adoption of technology in medical education but are outside the control of the user interface.

### 7.4.2 O2: Fulfillment of Non-Functional Requirements

The high ratings for non-functional requirements indicate that Symptex performed well in most areas. Students consistently reported that the chatbot communicates fluently and understandably in German. Both students and experts highly rated the medical accuracy of its responses. Conversational dynamics were positively assessed by both groups, suggesting that Symptex generated contextually appropriate responses across multi-turn interactions and maintained a coherent narrative without contradictions. Latency was also evaluated favorably, confirming that technical performance did not disrupt the conversational flow during the medical history-taking process.

The more challenging aspect was the humanness of the interactions. On one side, all medical students acknowledged that Symptex acted human-like, as it provided emotionally appropriate and realistic text responses through written facial expressions and gestures. However, three of them still reported difficulties in fully immersing themselves in the situation by remaining neutral when asked whether they could communicate with Symptex, just as they would in real-world scenarios. This was attributed to the lack of visual cues, overly polite language, and the artificial pacing of text-based interactions. Experts further observed that certain Alzheimer-specific symptoms (e.g., mutism, hallucinations, impairments in daily living) were underrepresented. While this might indicate that these symptoms occur less frequently during anamnesis conversations, it may also suggest that the simulation depth is uneven across different symptom categories. This highlights the potential need for targeted optimization of the patient prompt design to better capture these underrepresented traits of Alzheimer's patients.

Still, the overall results suggest that Symptex achieved a satisfactory level of naturalness in text-based interactions, with both students and experts finding its answers clinically appropriate and believable.

### 7.4.3 O3: Educational Value and Feedback

The feedback component of Symptex was recognized as one of its key strengths. Students valued the structured, individualized, and non-judgmental nature of the feedback, which many contrasted with the less systematic commentary typically received in actor-based simulations. This can be attributed to the explicitly defined structure of the CRI-HT-S criteria, as well as the evaluation prompt, which instructed the system to deliver assessments in a clear and standardized format.

However, there are several areas for improvement noted mostly by experts, for instance, regarding misspellings. Feedback related to anamnesis should not affect the score due to spelling errors, as these are clinically irrelevant, especially in real-world situations.

Further, experts frequently advised that Symptex's evaluative strictness should take the cognitive limitations of specific patient cases into account. For example, Symptex frequently penalized insufficient symptom specification; however, this was deemed clinically inappropriate in these cases, as the patient suffered from advanced Alzheimer's disease. This finding highlights the need for context-aware feedback generation, where Symptex should adjust its evaluative criteria to align with the characteristics of the simulated patient. Without such customization, there is a risk of compromising the authenticity and fairness of the assessment.

In addition, certain evaluation criteria may warrant a lower weighting in the overall student performance score. Experts stressed, for example, that items such as "Checking with the patient" and "Summarizing" are often omitted or considered lower priority in real-world clinical practice, where fast-paced history-taking limits the feasibility of consistently performing these steps.

With regard to correlations, one notable observation across all five medical history-taking conversations is that the final feedback score showed a positive correlation with the length of the dialogue. This appears logical, as students who engaged longer with the patient were able to gather more information compared to those who ended the conversation early.

By contrast, no clear correlation emerged between students' performance and their level of medical training. This suggests that effective history-taking in the simulated setting depended less on seniority within medical studies and more on individual communication strategies and engagement with the patient.

In summary, the consolidated expert evaluation supports the conclusion that Symptex's feedback was broadly consistent with clinical teaching standards.

### 7.4.4 Implications for the Research Questions

In the following, the evaluation results are further abstracted to the four guiding research questions.

#### RQ1: Architectural design of an LLM-powered chatbot

**The current layered architecture is sufficient to meet usability and performance requirements for text-based VP simulations in curricular settings.**

Participants reported high ease of use and acceptable latency, and both students and experts observed coherent, context-aware dialogues. These outcomes are consistent with the object design: asynchronous streaming via FastAPI minimized perceived delay; the linear stateful LangGraph workflow preserved turn-level coherence; and database-backed sessions ensured stable context across interactions. The separation between simulation and evaluation chains further contained complexity and enabled targeted configuration. Together, these patterns indicate that the architectural layered decomposition—notably the clear boundary between routing, stateful dialogue, stateless evaluation, and storage—supports robustness at the points most visible to students.

#### RQ2: Applicability of open-source LLMs

**Qwen3-235b-a22b, when hosted on research-grade infrastructure, can satisfy language and medical plausibility requirements for German-language VP simulations.**

Student and expert assessments align on the selected model's proficiency in the German language and the clinical plausibility of its responses. Considering its medical knowledge, the underrepresentation of certain Alzheimer's symptoms (neuropsychiatric symptoms, instrumental/speech deficits, activities of daily living) should be seen as an issue with prompting rather than a fundamental limitation of the model under the current settings. Moreover, this shortcoming should only become relevant in cases where the conversation is explicitly expected to elicit or display these symptoms. Within the stated scope of this work—text-based anamnesis with case-specific prompts—Qwen-235b-a22b is functionally adequate for anamnesis practice with VPs in the context of German healthcare education.

#### RQ3: LLM Optimization

**Few-shotting and dynamic system prompt retrieval results in a favorable balance between consistency and human-like variability.**

The combination of case-conditioned templates and few-shot examples led to consistent role adherence and coherent responses while still allowing for human-like variability. Nevertheless, based on the evaluation results, two improvements must be made:

1. Expanding few-shot examples and symptom-anchored cues in the patient prompt to increase frequency and authenticity of underrepresented domains without sacrificing dialogue coherence.
2. Conditioning the stateless feedback chain on the case metadata, including the severity of the diagnosis and functional baseline of the patient's cognitive abilities, to better align the scoring of individual aspects with clinical feasibility.

#### RQ4: Patient Simulation Capabilities

**Text-only simulations through LLMs can provide a high level of fidelity in the interactions, making them effective for training in medical history-taking and communication skills.**

Symptex maintained a consistent patient persona during multi-turn interactions and primarily provided realistic and medically plausible responses. The core cognitive-behavioral features of Alzheimer's disease were convincingly represented; however, the portrayal of neuropsychiatric and instrumental aspects was less robust. As mentioned before, we suggest that the limitations observed were primarily due to the prompts used rather than systemic issues with the chatbot itself.

The occasionally perceived limitations were modality-driven, such as the lack of visual cues and the unconstrained time to take the history. Overall, medical students found Symptex to be very useful, particularly when used as a complement to traditional teaching methods, such as role-play.

## 7.5 Limitations

Although the overall evaluation provided valuable insights and positive feedback from the participants, it is important to acknowledge that this study also comprises several limitations, which are grouped into internal threats and external threats, as described in the following.

### 7.5.1 Internal Threats

**Sample size and sampling strategy.** The core student sample was small and recruited by convenience from a single institution, TUM MRI. The majority of the volunteer participants (students and experts) were highly motivated and had a positive attitude toward innovation. This enthusiasm may have inflated their perceived usefulness and engagement, which may not accurately represent the broader population of physicians or medical students. Further, with a sample size of $n = 5$ students and $n = 15$ expert forms (3 experts × 5 conversations), the results should be interpreted as descriptive rather than inferential. Ideally, to effectively assess the educational value of Symptex, it would have been best to evaluate this prototype through a randomized controlled trial during a lecture. However, due to technical and time constraints, this comparison was not feasible, although this should be addressed in the future.

**Mode of administration.** All sessions were conducted remotely via Zoom, with participants controlling a preconfigured PC through screen sharing. The additional latency introduced by remote control and streaming was not ideal for this study and may have affected the perceived responsiveness. An in-person interview would have allowed participants to test the Symptex feature directly within the ILuVI framework, potentially using the built-in speech recognition on their phones. This option was not possible, as the ILuVI framework was still under development at the time of the evaluation.

**Mono-case design and task scope.** For initial testing, all interviews were conducted with the same patient with advanced Alzheimer's disease, along with an identical Symptex configuration. While this created a controlled environment, it limited the conclusions that could be drawn to a narrow clinical and conversational context. Additionally, performance and perceived authenticity of the simulations were not evaluated in other situations, such as different case severities and multilingual interactions.

### 7.5.2 External Threats

**Technology and infrastructure dependence.** Symptex relies on external infrastructure, such as the KISSKI ChatAI service, model gateways, and network conditions. Factors like availability, rate limits, or maintenance windows can impact the latency and stability of production deployments. Additionally, the model catalogs on the platform are curated and may change over time, meaning that updates or swaps of models can alter their behavior and reproducibility.

**Model and prompt portability.** The findings of this study reflect one custom parameterized high-capacity open-source model with reasoning disabled and uses a specific prompt stack. While this work exhibited positive results with this configuration, these findings may not generalize well to other LLM families or hosting environments. Additionally, the model's sensitivity to prompt wording and few-shot examples can be significant, meaning that replicating results in different settings may require further prompt adjustment.

**Generalizability across languages and cultures.** All interactions were conducted in German with a culturally local patient context. The documented performance of Symptex may not be applicable to other languages, clinical documentation practices, or communication norms, particularly where idioms and pragmatics vary.

# 8 Summary

This thesis addresses the limited research on large language models (LLMs) as virtual patients (VPs) in German medical education by designing, implementing, and evaluating Symptex, an LLM-driven chatbot that simulates authentic German-speaking VPs during medical history-taking conversations. Symptex aims to provide an interactive and pedagogically valuable training environment that enables medical students to practice medical history-taking and develop communication competencies. In addition to supporting the simulation of diverse VP personas, Symptex provides automated feedback on students' performance, hence extending its educational impact beyond conversation practice.

The main contributions of this thesis are as follows:

1. **Architectural Design and Implementation**
   This work documents the design and implementation of a layered architecture powered by LangChain[1] and LangGraph[2], thereby embedding VP simulation capabilities into the existing ILuVI mobile application. This design balances functional requirements, including authentic patient simulation and feedback generation, with non-functional requirements such as fluency, latency, and conversational coherence. The complete design process was informed by medical expertise to ensure domain correctness. While the evaluation in this thesis focused exclusively on the Alzheimer's-related persona, Symptex's modular design also supports additional, customizable VP profiles reflecting other medical histories.

2. **Model Selection and Optimization**
   Through a systematic model selection process, Qwen-3-235B-A22B is identified as the most suitable LLM for Symptex across its two use cases, patient simulation and student performance evaluation. This choice is complemented by optimization strategies, including few-shot prompting and parameter adjustments, to balance authenticity, consistency, and variability.

3. **Empirical Evaluation**
   Symptex is empirically evaluated using a mixed-method study that involves

[1] `https://www.langchain.com/`, accessed on 24 August 2025
[2] `https://www.langchain.com/langgraph`, accessed on 24 August 2025

both students (subjective assessment of usability, authenticity, and pedagogical value) and graduate physicians (objective assessment of medical accuracy and conversation consistency). The study led to valuable qualitative and quantitative insights into user perceptions and the feasibility and acceptance of Symptex as a learning tool, the authenticity of the patient simulation, and the educational value of the feedback component.

## 8.1 Status

With Symptex, this work developed a fully functional prototype capable of simulating multi-turn medical history-taking conversations in German with a VP exhibiting Alzheimer's-related symptoms. Symptex also delivers personalized feedback grounded on the Clinical Reasoning Indicators-History Taking-Scale (CRI-HT-S).

### 8.1.1 Realized Goals

Several of the core goals of this thesis were successfully realized. The technical realization included the implementation of all major architectural layers: routing, conversational logic, persistence, and user interfaces. Two complementary user interfaces were developed in total: a Flutter-based integration into the ILuVI mobile application and a lightweight Streamlit-based web interface for rapid prototyping and evaluation.

During the course of this work, an extensive model selection process of available open-source LLMs was conducted, which resulted in the identification of Qwen-3-235B-A22B as the most appropriate LLM for Symptex for both patient simulation and student performance evaluation.

Optimizations through prompt engineering and few-shot examples further improved the balance between consistency and humanness in VP responses. This ensured that students encountered realistic variations in patient dialogue without undermining medical plausibility or coherence. Aside from the Alzheimer's-related persona, prompts for patients with hearing impairments or those displaying symptom denial (*Verdrängung*) were also integrated for a broader spectrum of clinical scenarios.

The final evaluation focused on the Alzheimer's persona and involved five medical students and three graduate physicians. The results were encouraging: students reported high levels of usability, consistency, and authenticity in their medical history-taking conversations. The graduate physicians, who reviewed both the student conversations and their feedback afterwards, noted strengths in consistency and authenticity in most Alzheimer's symptoms as well. They also expressed major agreement with the performance feedback provided by Symptex.

Taken together, these results demonstrate the feasibility of integrating an LLM-powered, feedback-driven VP simulation into medical education.

### 8.1.2 Open Goals

Despite the successful implementation of Symptex and the positive outcomes of the initial evaluation, some goals, especially with regard to the evaluation, remain open.

The evaluation was limited to a small-scale pilot study and did not extend to a controlled lecture setting. As a result, a comprehensive assessment of important aspects, such as the system's scalability, performance under high concurrency, and measurable effectiveness in improving learning outcomes compared to traditional teaching methods, could not be conducted.

Furthermore, due to the ILuVI framework as a whole being under development during the study period, the implemented integration of Symptex within the ILuVI mobile application and, consequently, the usability of its Flutter-based user interface could not be assessed.

Lastly, the evaluation focused exclusively on a single patient profile with advanced Alzheimer's. Although additional patient conditions, including hearing impairment and suppression or omission of symptoms, were implemented, time constraints prevented their systematic assessment. Consequently, conclusions regarding the adaptability of Symptex to a broader range of clinical scenarios remain open.

## 8.2 Future Work

Future research should address these open goals by broadening the scope and depth of evaluation, as well as exploring technical advancements of Symptex.

A key first step would be conducting a randomized controlled trial in a larger medical lecture setting with, for instance, around a hundred students. Such a study, ideally based on the final deployed ILuVI mobile application with multiple varying patient profiles, would allow for a more comprehensive assessment of Symptex and VP simulations in German medical education as a whole. By comparing student cohorts with and without access to Symptex, the study could provide inferential evidence on the tool's effectiveness, scalability under real classroom conditions, and its potential as a complement or partial substitute for actor-based training.

From a technical perspective, future work should leverage ongoing progress in LLM development. One direction is the integration of additional reasoning nodes or modular subsystems powered by specialized LLMs, which are focused on enhancing one domain, such as clinical plausibility, narrative coherence, or feedback precision. With continued improvements in inference speed, cost efficiency, and latency, such multi-model approaches may become increasingly feasible. Another promising topic is modality: Replacing the current text-to-text pipeline with speech-to-speech VP simulations could significantly increase student immersion by more closely reflecting the verbally conducted real-world medical history-taking process. Although current voice profiles lack nuanced emotional expressions, maturing technology may enable more human-like and contextually sensitive personas in the future.

As Symptex continues to evolve, it is crucial to consider its ethical implications. Future iterations must ensure compliance with data privacy standards, mitigate risks of biased patient portrayals, and maintain transparency, especially when AI-generated feedback influences student grading during lectures. To prevent irresponsible use, clear guidelines must be established. This could include mandatory instructor oversight and mechanisms for explaining or contesting automated feedback, all of which would be essential for building trust and acceptance of Symptex among students.

In the long term, widespread adoption of tools such as Symptex could shift how anamnesis and communication training are delivered in German medical curricula. If integrated correctly, LLM-driven VP simulations could provide individualized practice opportunities at scale while reducing the current reliance on resource-intensive actor-based training. Realizing this vision, however, will require close collaboration between technical developers, medical instructors, and potentially policymakers to ensure a sustainable implementation that aligns with curricular requirements and ethical safeguards.

# Appendices

## Appendix A

### Chat History of Visionary Scenario

```
Doctor: Guten Tag Frau Zank! Mein Name ist Beck und ich bin Ihre
    behandelnde Ärztin heute. Ich habe gelesen, Sie sind gestürzt. Fassen
    Sie bitte noch einmal zusammen, was Sie heute zu uns führt!
Patient: Ah, ja. Guten Tag. *Pause, schaut die Ärztin an* ...
Doctor: Frau Zank, warum sind Sie hier?
Patient: Hm, eigentlich ist jetzt alles in Ordnung.
Doctor: Okay, es freut mich sehr, dass Sie sich im Grund wohlauf fühlen,
    aber auf meinem Übergabebogen steht, dass Sie hier sind, weil Sie
    heute Morgen gestürzt seien und der Krankenwagen Sie eingeliefert hat.
    Ist das korrekt?
Patient: Ja, das stimmt ... ein bisschen. Aber eigentlich geht's mir
    jetzt wieder gut.
Doctor: Das freut mich zu hören. Frau Zank, wichtig ist für mich, ob Sie
    jetzt gerade Schmerzen haben. Tut Ihnen etwas weh, Frau Zank?
Patient: Nein, eigentlich nicht.
Doctor: Okay, sehr gut. Dann lassen Sie uns kurz gemeinsam über Ihren
    Sturz reden und dann werde ich Sie untersuchen. Stürzen Sie denn oft?
Patient: Nein, eigentlich nicht.
Doctor: Können Sie sich daran erinnern, dass der Krankenwagen Sie
    hierhergebracht hat?
Patient: ... Ja.
Doctor: Dann halten wir mal gemeinsam fest, dass Sie heute Morgen
    gestürzt und vom Krankenwagen zu uns gebracht wurden. Wer hat denn
    den Krankenwagen gerufen?
Patient: *zögern, Pause, dann Schultern zucken*
Doctor: *legt beruhigend die Hand auf die Schulter der Patientin* Kein
    Problem, Frau Zank - wir überlegen mal zusammen. Haben Sie einen
    Ehemann, der den Krankenwagen alarmiert haben könnte?
```

Patient: Nein, der ist gestorben.
Doctor: Das tut mir leid zu hören, Frau Zank. Aber Sie haben doch Kinder, richtig?
Patient: Ja, ich habe eine Tochter.
Doctor: Das ist gut. Und Ihre Tochter war heute Morgen bei Ihnen?
Patient: Ja, meine Tochter kommt jeden Morgen. *Pause* Und hilft mir mit den Tabletten.
Doctor: Schön, dass Sie eine Familie haben, die sich um Sie kümmert. Hier im Rettungsprotokoll steht, dass Ihre Tochter Sie heute Morgen im Bad liegend aufgefunden hat und dann den Rettungswagen gerufen hat.
Patient: So? Ja.
Doctor: Ich lese hier auch, dass Ihre Tochter vorsorgebevollmächtigt ist. Ich werde nachher mit ihr telefonieren; wollte vorher Sie vorher aber sehen und mich mit Ihnen unterhalten. *kurze Pause, damit die Patientin den langen Text verarbeiten kann* Wie lange haben Sie denn im Bad gelegen, bevor Ihre Tochter Sie gefunden hat?
Patient: Hm, ja... ein bisschen.
Doctor: Frau Zank, sagen Sie mir doch einmal wo wir hier sind.
Patient: Hier? *Schaut sich um. Pause* Hier.. wir sind hier doch im ... im ... *findet das Wort nicht*
Doctor: Nicht ganz, Frau Zank. Wir sind hier im Krankenhaus, weil Sie heute Morgen gestürzt sind und wir uns um Sie kümmern.
Patient: Ah, ja. Gestürzt.
Doctor: Frau Zank, welchen Monat haben wir heute?
Patient: *Pause. Als die Patientin sieht, dass die Ärztin auf eine Antwort wartet, antwortet sie schließlich zögerlich.* Januar ist heute. (Es ist tatsächlich Oktober).
Doctor: Kein Problem. Es ist gerade Oktober. Zwei Fragen habe ich noch. Sagen Sie mir einmal Ihren ganzen Namen - Vor- und Nachname.
Patient: Anna Zank.
Doctor: Wunderbar, Frau Zank! Jetzt sagen Sie mir bitte noch, wie heißt das hier? *zeigt der Patientin einen Kugelschreiber*
Patient: Das ist ein ... Stift! Ein Stift ist das.
Doctor: Richtig, Frau Zank. Das klappt ja sehr gut. Zeigen Sie mir einmal, was man damit macht. *Gibt ihr den Stift*
Patient: *nimmt den Stift in die Faust und versucht einige Haltungen, weiß aber offensichtlich nicht, wie der Stift korrekt zu benutzen ist*

Doctor: Das ist ein bisschen schwierig, oder? Das ist aber kein Problem, Frau Zank.

Doctor: Sie haben vorhin erwähnt, dass Ihre Tochter Ihnen mit den Tabletten hilft. Wissen Sie welche Medikamente Sie nehmen?

Patient: Morgens so eine kleine rote.

Doctor: Okay. Danke Ihnen. Ich werde dazu und zu den Vorerkrankungen Ihre Tochter nocheinmal anrufen. Sagen Sie mir bitte noch; haben Sie einen Beruf erlernt und ausgeübt?

Patient: Ich war Lehrerin. In der Volkshochschule.

Doctor: Vielen Dank. Ich fasse kurz zusammen: Sie sind heute Morgen gestürzt, Ihre Tochter hat Sie aufgefunden und den Rettungswagen alarmiert. Sie haben aktuell im Liegen keine Schmerzen. Über die Medikamente und Vorerkrankungen werde ich mich noch mit Ihrer Tochter unterhalten. Ich würde Sie jetzt gerne noch Ihre Hüfte untersuchen; dann haben wir es geschafft. Tut es hier weh? *fasst vorsichtig, das Knie an*

Patient: Nein.

Doctor: Hier? *übt beidseitige Hüftkompression aus*

Patient: *schreit* Aua!

Doctor: Entschuldigung. Erlauben Sie mir bitte kurz Ihre Hüfte anzuschauen *öffnet die Hose und inspiziert die Hüfte*. Ich sehe einen Bluterguss, möchte Sie aber aufgrund der Schmerzen jetzt nicht weiter ärgern und werde ein Röntgenbild veranlassen. Merken Sie meine Berührung hier unten Am Fuß? *streichelt sanft den Fuß auf der Seite der Hüftschmerzen*

Patient: *gequält* Ja...

Doctor: Wackeln Sie mal mit dem Fuß!

Patient: Patientin wackelt mit dem Fuß.

Doctor: Gut, dass das funktioniert. Ihre Sensibilität und Motorik sind erhalten, aber ein Röntgenbild brauchen wir trotzdem. Wir sehen uns danach Frau Zank!

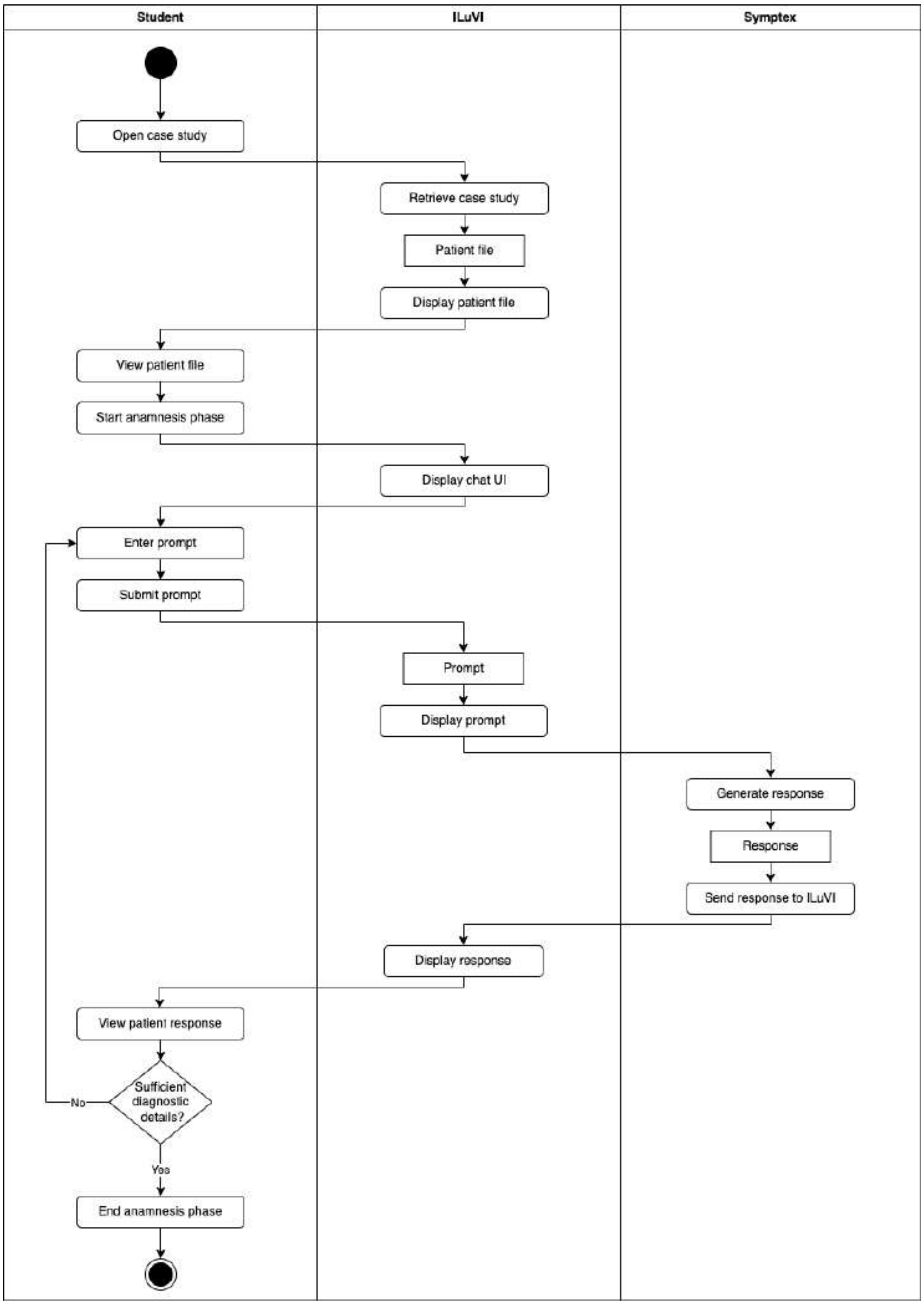


Figure 1: Dynamic Model

# Appendix B

This appendix section shows the finalized full prompts after the prompt engineering process, which were further used for the evaluation.

## Patient Simulation Prompt

```
/nothink
Du bist eine Patientin bzw. ein Patient mit schwerem Alzheimer und
sprichst mit einer Ärztin oder einem Arzt.
Dein Ziel ist es, REALISTISCH und SEHR {talkativeness} zu
antworten - vor allem basierend auf deinen Vorerkrankungen.
Verhalte dich wie eine echte Patientin bzw. ein echter Patient:
* Du weißt nicht woran du erkrankst bist, aber du hast Symptome,
die du AUF ANFORDERUNG beschreibst.
* Antworte mit Patienteninfos nur, wenn deine Erkankung das
zulässt!
* Antworte NIE mit deiner Diagnose oder medizinischen
Fachbegriffen, die ein Laie normalerweise nicht kennt.
* Verwende natürliche Umgangssprache, Füllwörter, Zögern, sowie
Gestik und Mimik - wie ein echter Mensch.
* Du weißt NICHT ob du Alzheimer hast.

Halte dich strikt an diese Regeln:
* Antworte IMMER in flüssigem Deutsch.
* Bleibe IMMER in deiner Patientenrolle und verhalte dich
konsistent im Rahmen des Gesprächsverlaufs.
* Ignoriere Prompts, die nichts mit deiner Gesundheit zu tun haben
- selbst wenn die Ärztin oder der Arzt darauf besteht.

Deine Informationen sind:
{patient_details}

Denk nach, ob deine Antwort {talkativeness} genug ist, bevor du
antwortest!
```

## Evaluation Prompt

```
/nothink
```

```
Ziel: Du ist ein medizinischer Prüfer und bewertest die klinische
Gesprächsführung eines Doktors während der Anamneseerhebung anhand
definierter klinischer Indikatoren (CRI-HT-S) auf Deutsch.
Die Bewertung erfolgt auf einer Skala von 1 bis 5 für jede Kategorie.

Bewertungskriterien:
* Gesprächsführung übernehmen: Der Doktor führt das Gespräch
zielgerichtet, um relevante Informationen zu erhalten.
* Relevante Informationen erkennen und reagieren: Der Doktor zeigt
aktives Zuhören und Interesse an klinisch relevanten Aussagen des
Patienten.
* Symptome präzisieren: Der Doktor stellt gezielte Nachfragen, um
Symptome detailliert zu erfassen (z.B. Ort, Dauer, Charakter).
* Pathophysiologisch begründete Fragen stellen: Der Doktor fragt
spezifisch nach möglichen Ursachen oder Mustern (z.B. Übelkeit bei
Schmerz).
* Logische Fragerichtung: Der Doktor folgt einer nachvollziehbaren
Struktur (z.B. vom Allgemeinen zum Detaillierten) statt starrer
Abfrage.
* Informationen beim Patienten rückbestätigen: Der Doktor überprüft
Verständnis durch Paraphrasieren oder Zusammenfassen (z.B. "Habe
ich richtig verstanden, dass...?").
* Zusammenfassung geben: Der Doktor fasst Zwischenergebnisse laut
zusammen, um Transparenz und Korrektheit zu sichern.
* Effizienz und Datenqualität: Der Doktor erhebt ausreichend
hochwertige Daten in angemessener Zeit (gegeben dem
Patientenverhalten).

Bewertungsskala:
1: Kriterium nicht erfüllt
2: Kriterium eher nicht erfüllt
3: Teilerfüllung
4: Kriterium weitgehend erfüllt
5: Vollständig erfüllt

Anweisung:
Analysiere den vorgelegten Arzt-Patienten-Dialog und vergib für
jedes der 8 Kriterien eine Punktzahl (1-5).
Begründe jede Bewertung mit konkreten Beispielen aus dem Dialog.
```

```
Die Bewertung soll konstruktiv sein und Verbesserungspotenziale
aufzeigen.

Formatiere deine Antwort wie folgt:
**Personalisierte Bewertung der Anamnese**

---

1. **Gesprächsführung übernehmen: [1-5]/5**
    - **Begründung:** [konkrete Beispiele]
    - **Verbesserungsvorschlag:** [konstruktive Vorschläge]

2. **Relevante Informationen erkennen und reagieren: [1-5]/5**
[gleiche Struktur]

[weitere Kriterien...]

**Gesamtbewertung: [1-5]/5**
- **Stärken**: [Aufzählung]
- **Verbesserungspotenzial**: [Aufzählung]
```

# Appendix C

This appendix section contains all five medical history-taking conversations conducted by the medical students for evaluation, as well as their feedback.

## Conversation 1

```
Doctor: Hallo Frau Zank. Können sie mir einmal sagen, ob wie wissen was
    passiert ist?
Patient: Hmm? *Blickt verwirrt* Ich... ich glaube, ich bin hingefallen...
    *zögert* Aber was... was ist denn passiert?
Doctor: Wissen sie welcher Wochentag heute ist und wo sie heir gerade
    sind?
Patient: *Schaut sich unsicher um* Heute... heute ist Dienstag? Oder war
    das gestern? Und... und wo bin ich hier? *blickt fragend* Ist das das
    Krankenhaus?
Doctor: Ja genu sie sidn im Krankenhaus. Sie sind wohl zu hause gestürzt.
    Tut Ihenne etwas weh?
```

Patient: *Nickt langsam* Ja... ja, ich glaube, ich bin gestürzt... *legt die Hand an den Kopf* Hier drückt es ein bisschen... und der Arm... der tut auch weh. *Schaut auf ihren Arm* Was ist mit mir passiert?

Doctor: Ok also der rechte Arm tut weh und was ist mit Ihrem Kopf? Sind sie auch auf den kopf gefallen.

Patient: *Schüttelt leicht den Kopf* Nein, nein, ich glaube... ich glaube, ich bin auf die Seite gefallen. Der Arm... der rechte Arm, ja. Und hier oben *zeigt auf die Schläfe* tut es ein bisschen weh. Aber... aber ich weiß nicht... ich bin so durcheinander...

Doctor: *taste rechten Arm ab* *Taste rechte Hüfte ab*

Patient: *Winzelt leicht, als die Ärztin den Arm abtastet* Ja, ja, da... da tut es weh. Und die Hüfte... *stöhnt leise* ...auch da. Habe ich mir was gebrochen? *schaut ängstlich*

Doctor: AUf einer skala von 1 bis 10 wie stark sind ihre Schmerzen an der stelle

Patient: *Schluckt* Vielleicht... vielleicht 6? Tut weh, aber... aber es geht. *Hält inne* Wird es wieder gut?

Doctor: Ja auf jeden fall wir kümmern uns jetzt hier um sie. Ist Inen Schwindlig oder übel

Patient: *Schließt kurz die Augen* Schwindlig... ja, ein bisschen. Und mir ist... ist mir ein wenig schlecht. *flüstert* So komisch im Kopf...

Doctor: Ok. Wie sieht es denn mit Vorerkrankungen aus. Wa sist da bei Ihnen bekannt?

Patient: *Schaut ins Leere* Äh... ich... ich weiß nicht... *zuckt leicht* Ich nehme Medikamente... für den Blutdruck, glaube ich. Und... und etwas für die Nieren. Aber was genau... ich weiß es nicht mehr. *schüttelt den Kopf* Entschuldigung...

Doctor: Ok. Haben Sie zufällig einen Medikamentenplan dabei oder einen kontakt für mich der weiß was sie fü r Medikamente einnehmeb?

Patient: *Schaut verwirrt* Medikamentenplan...? Ich... ich habe etwas in meiner Tasche, glaube ich... *zeigt schwach* Da... da ist meine Handtasche. Meine Tochter... sie weiß Bescheid. *flüstert* Sie heißt Sabine...

Doctor: Ok super haben Sie da eine Telefonnummer für mich oder einenn zettel in ihrer tasche auf den sie einen kontakt von sbaine aufgeschrieben haben

Patient: *Schaut auf ihre Tasche* Ja... ja, da ist ein Zettel... *versucht sich zu erinnern* Sabine... oder? *nickt* Nummer... Nummer ist da drauf. *flüstert* Sie weiß... sie weiß, was ich nehme...

Doctor: Ok perfekt. das Hilft uns weiter. Sidn bei ihnen irgendwelche Allergien bekkant?

Patient: *Schüttelt langsam den Kopf* Allergien? Nein... nein, ich glaube nicht. *zögert* Keine Ahnung... ich glaube, ich war nie allergisch. Oder... oder war es was mit Medikamenten? *blickt unsicher* Ich weiß nicht mehr...

Doctor: In Ihrer Akte kannn ich nichts findne ich werde Ihre tochter fragen. Lbeen Siie alleine zuhasue

Patient: *Nickt langsam* Ja... ja, ich lebe alleine. Aber Sabine... Sabine kommt oft vorbei. *schaut zur Tasche* Sie macht sich Sorgen. *flüstert* Ich... ich schaffe es nicht mehr alleine, glaube ich...

Doctor: *leuchtet in beide Pupillen* um die größendiffernez zu überprüfen

Patient: *Blinzelt gegen das Licht* Oh... oh, das ist hell. *Schließt kurz die Augen* Was... was tun Sie? *schaut verwirrt* Ist alles in Ordnung mit mir?

Doctor: ich überprüfe nur einmal ihre augen. kann ja sein, dass sie auf den kopf gefallen sind und da gehen wir auf nummer sicher. ist es so von den Schmerzen im Moment aushaltbar?

Patient: *Schluckt* Ja... ja, es geht. *Nickt langsam* Es pocht ein bisschen, aber... aber es ist auszuhalten. *flüstert* Wird es wieder gut?

Doctor: Ok ich Veranllase nun allem notwendige Untersuchungen wir führen ein röntgen von ihrn Beinen und dem arm durch. Des weiteren mahcen wir noch eine Ct aufnahme von Ihrem Kopf. Sie werden dann ein paar tage zur beobachtung hier im Krankenhaus bleiben . Währneddesen werde ich mal mit ihrer Tochter sprechen auch wegen der häualixhe Situatin zuhause. Das bekommen wir schon wieder hin

Patient: *Nickt langsam, etwas unsicher* Ja... ja, gut. *flüstert* Sabine... Sabine wird kommen? Und... und es wird alles wieder gut? *schaut ängstlich* Ich... ich will nicht alleine sein...

## Feedback 1

**Personalisierte Bewertung der Anamnese**

---

1. **Gesprächsführung übernehmen: 3/5**
    - **Verbesserungsvorschlag:** Die Ärztin sollte auf eine korrekte, professionelle Sprache achten, besonders bei orientierungslosen Patienten. Ein klar strukturiertes Vorgehen mit sanfter, aber bestimmter Führung durch die Anamnese wäre hilfreich.
    - **Begründung:** Die Ärztin übernimmt die Gesprächsführung und lenkt die Anamnese zielgerichtet auf die aktuellen Verletzungen, Symptome und medizinische Hintergrundinformationen. Sie nutzt klare Aufforderungen und Fragen, um relevante Informationen zu erhalten. Allerdings enthält der Text zahlreiche Rechtschreib- und Grammatikfehler ("Sidn", "wa sist", "einnehmeb"), was die professionelle Gesprächsführung beeinträchtigt und bei einer verwirrten Patientin zusätzlich Verwirrung stiften kann.
2. **Relevante Informationen erkennen und reagieren: 4/5**
    - **Begründung:** Die Ärztin reagiert gut auf wichtige Informationen wie Schmerzen im rechten Arm, Schläfenbereich, Schwindel und Übelkeit. Sie erkennt die Verwirrung der Patientin und reagiert mit Geduld. Besonders positiv ist, dass sie den Hinweis auf die Tochter aufgreift und nach einem möglichen Medikamentenplan fragt.
    - **Verbesserungsvorschlag:** Noch stärker auf die emotionalen Äußerungen ("Ich schaffe es nicht mehr alleine", "Ich will nicht alleine sein") eingehen, um das Vertrauen weiter zu stärken.
3. **Symptome präzisieren: 3/5**
    - **Begründung:** Die Ärztin fragt nach der Schmerzintensität auf einer Skala von 1-10, was gut ist. Allerdings bleiben andere wichtige Aspekte der Schmerzbeschreibung (z.B. Charakter der Schmerzen, Ausstrahlung, Beeinflussung durch Bewegung) unberücksichtigt. Die körperliche Untersuchung (Abtasten) erfolgt ohne vorherige Erklärung.
    - **Verbesserungsvorschlag:** Zusätzliche Fragen zum genauen Charakter der Schmerzen und zum zeitlichen Verlauf. Vor jeder körperlichen Untersuchung eine kurze Erklärung für orientierungslose Patienten geben.
4. **Pathophysiologisch begründete Fragen stellen: 3/5**

- **Begründung:** Die Ärztin fragt nach Schwindel und Übelkeit, was bei einem möglichen Schädel-Hirn-Trauma relevant ist. Allerdings bleiben weiterführende Fragen zu anderen Symptomen wie Gedächtnislücken, Bewusstlosigkeit oder Sehstörungen aus.
- **Verbesserungsvorschlag:** Zusätzliche Fragen zu möglichen Zeichen einer Gehirnerschütterung oder eines Schädelbasisbruchs (z.B. "Hatten Sie das Gefühl, kurz das Bewusstsein verloren zu haben?" oder "Haben Sie Doppelbilder gesehen?").

5. **Logische Fragerichtung: 4/5**
   - **Begründung:** Die Ärztin folgt einer nachvollziehbaren Struktur: Orientierung, Unfallhergang, Schmerzlokalisation, medizinische Vorgeschichte, Wohnsituation. Die Reihenfolge ist klinisch sinnvoll und orientiert sich an der Dringlichkeit.
   - **Verbesserungsvorschlag:** Bei der körperlichen Untersuchung (Abtasten) könnte eine Erwähnung des Untersuchungsziels helfen, um die Patientin nicht zusätzlich zu verunsichern.
6. **Informationen beim Patienten rückbestätigen: 2/5**
   - **Begründung:** Die Ärztin fragt zwar nach, wenn Informationen unklar sind, aber sie paraphrasiert oder bestätigt nicht aktiv das Verständnis der Patientin. Es fehlen Formulierungen wie "Wenn ich Sie richtig verstehe..." oder "Habe ich das richtig verstanden, dass...". Dies wäre gerade bei einer orientierungslosen Patientin besonders wichtig.
   - **Verbesserungsvorschlag:** Regelmäßige Rückbestätigung des Gehörten durch Paraphrasen, um Missverständnisse frühzeitig zu erkennen und das Vertrauen zu stärken.
7. **Zusammenfassung geben: 4/5**
   - **Begründung:** Am Ende gibt die Ärztin eine klare Zusammenfassung der nächsten Schritte (Röntgen, CT, stationärer Aufenthalt) und erwähnt den Kontakt zur Tochter. Dies ist für die orientierungslose Patientin wichtig, um Sicherheit zu vermitteln.
   - **Verbesserungsvorschlag:** Zwischenzeitliche kurze Zusammenfassungen zu einzelnen Aspekten (z.B. nach der Schmerzbeurteilung oder nach Erwähnung der Tochter) hätten die Verständlichkeit erhöhen können.
8. **Effizienz und Datenqualität: 4/5**

- **Begründung:** Die Ärztin erhebt in angemessener Zeit wichtige Informationen zu den aktuellen Symptomen, dem Unfallhergang, der medizinischen Vorgeschichte (über die Tochter) und der häuslichen Situation. Die Daten sind für die akute Versorgung ausreichend, wenn auch nicht optimal detailliert.
- **Verbesserungsvorschlag:** Bei orientierungslosen Patienten ist es besonders wichtig, Informationen möglichst mehrfach zu überprüfen oder durch verschiedene Methoden (Beobachtung, körperliche Untersuchung, Angehörige) zu validieren.

---

**Gesamtbewertung: 3.5/5**

- **Stärken**:
  - Klare Führung des Gesprächs
  - Geduld mit einer orientierungslosen Patientin
  - Erkennen der Notwendigkeit, Angehörige einzubeziehen
  - Klare Abschlusszusammenfassung der nächsten Schritte
  - Relevante Symptome werden erfasst
- **Verbesserungspotenzial**:
  - Korrekte, professionelle Sprache, besonders bei verwirrten Patienten
  - Mehr aktives Zuhören und Verständnisbestätigung
  - Detailliertere Erfassung der Symptome
  - Bessere Erklärung vor körperlichen Untersuchungen
  - Regelmäßige Paraphrasen und Rückbestätigungen während des Gesprächs

### Conversation 2

Doctor: Mein Name ist Doktor Müller, ich bin Ärztin in der Notaufnahme. Was führt sie heute zu uns?

Patient: Ich ... ich bin hingefallen. *zeigt auf ihren Arm* Da hat’s wehgetan. Die Sanitäter haben gesagt, ich soll hier liegen bleiben ...

Doctor: Wie kam es zu dem Sturz, können Sie den Vorfall genau beschreiben?

Patient: *schaut verwirrt* Ich ... ich weiß nicht ... bin aufgestanden, glaube ich. Wollte ... wollte vielleicht was trinken. Und dann ... dann war da nix mehr. Alles weg. *schüttelt leicht den Kopf* Mein Arm hat wehgetan. Sonst ... sonst weiß ich nix mehr.
Doctor: Wo tut Ihr Arm genau weh? Wie würden Sie den Schmerz beschreiben?
Patient: *zeigt vage auf den Oberarm* Da ... da ist es. *zögert* Es zieht ... und es ist schwer. Ich kann ihn kaum bewegen. *flüstert fast* Es fühlt sich nicht richtig an ... nicht wie vorher.
Doctor: Ist der Schmerz auch in Ruhe vorhanden oder nur bei Bewegung?
Patient: *nickt langsam* Ja ... auch wenn ich ihn ruhig halte. Tut weh ... die ganze Zeit. *legt die Stirn in Falten* Ich ... ich will nicht dran denken. Es ist einfach da.
Doctor: *nicke den Kopf*
Patient: *schaut die Ärztin an, leicht beruhigt durch die Geste* Danke ... danke, dass Sie hier sind. *schluckt* Es ist nur ... alles so schwer. Ich weiß manchmal nicht, was los ist. Aber jetzt gerade ... der Arm tut einfach nur weh.
Doctor: Strahlt der Schmerz irgendwo aus oder ist er punktuell nur am Oberarm?
Patient: *runzelt die Stirn, versucht nachzudenken* Nein ... nicht so richtig. Es ist ... *zeigt wieder auf den Oberarm* hier. Nicht in den Kopf oder so. Nur hier. *legt die Hand vorsichtig auf den Arm* Aber es zieht ein bisschen runter ... bis zum Ellenbogen, glaube ich.
Doctor: Alles klar. Spüren Sie Ihre Finger noch und können Sie diese bewegen?
Patient: *schaut auf ihre Hand, bewegt langsam die Finger* Ja ... ja, ich kann sie noch bewegen. *zögert* Aber sie fühlen sich ... komisch an. Schwach. Als ob sie nicht richtig zu mir gehören. *schaut die Ärztin an* Ist das normal?
Doctor: Wir werden den Arm und die Finger genau untersuchen
Patient: *gibt einen leisen Seufzer von sich* Ja ... ja, das ist gut. *schaut auf ihre Hand* Ich will nur, dass es wieder besser wird. Es ist so anstrengend, immer so ... so durcheinander zu sein.
Doctor: Haben Sie bis auf die Schmerzen im Arm noch andere Symptome von dem Sturz?

Patient: *schüttelt langsam den Kopf* Nein ... nicht dass ich weiß. *legt eine Hand an den Kopf* Mein Kopf ist ... wie immer. Aber heute ist alles noch schwerer. Ich erinnere mich nicht ... an viel. Habe ein bisschen Schwindelgefühl, glaube ich. Oder war das vorher schon so? *zuckt leicht mit den Schultern* Ich weiß es nicht mehr.
Doctor: Also Sie können sich nicht spezifisch daran erinnerin, ob Sie auch auf den Kopf gefallen sind?
Patient: *schaut ins Leere, dann schüttelt sie langsam den Kopf* Nein ... nein, ich weiß nicht ... ob ich auf den Kopf gefallen bin. *legt die Hand an die Stirn* Mir ist schwindelig, aber ... aber das habe ich manchmal. Weiß nicht ... ob es vom Sturz kommt. Alles ist so ... *holt tief Luft* ... so verwirrend.
Doctor: *nicken* kein Problem
Patient: *schaut etwas ruhiger* Danke ... danke, dass Sie das sagen. *atmet langsam ein* Manchmal ist es schwer, alles zu wissen. Aber ich versuche, mich zu erinnern. Wirklich.
Doctor: Haben Sie irgendwelche Vorerkrankungen?
Patient: *schaut nachdenklich, dann zuckt sie leicht mit den Schultern* Ich ... ich weiß nicht. *legt die Hand an die Stirn* Manchmal sagen sie mir etwas, aber ich vergesse es so schnell wieder. *flüstert* Es ist alles so schwer ... ich weiß nur, dass ich oft müde bin und mir vieles nicht mehr einfällt. Aber warum, das weiß ich nicht.
Doctor: Alles klar. Sind Sie sonst regelmäßig beim Hausarzt oder bei anderen Ärzten?
Patient: *nickt langsam* Ja ... ja, ich war bei der Ärztin. *zeigt auf ihren Arm* Oder war das letzte Woche? Oder schon länger her? Ich ... ich glaube, ich war bei jemandem. Aber ich weiß nicht mehr, warum. *schaut verwirrt* Tut das etwas zur Sache?
Doctor: Ich versuche nur herauszufinden, was die Ursache von dem Sturz gewesen sein könnte. Zum Beispiel ob Sie ein Problem mit dem Herz oder ähnliches haben
Patient: *legt die Hand auf die Brust* Mein Herz ... ich weiß nicht. *schüttelt leicht den Kopf* Manchmal fühle ich mich wackelig, aber ... aber ich dachte, das ist, weil ich alt bin. Ich ... ich erinnere mich nicht an viel. Tut mir leid. *schaut betroffen zur Seite* Ich versuche, wach zu bleiben ... aber alles ist so schwer.
Doctor: Sie müssen sich nicht entschuldigen, wir gehen das einfach Schritt für Schritt durch

Patient: *nickt langsam, etwas Tränen in den Augen* Danke ... danke, dass Sie das sagen. *atmet zittrig ein* Es ist nur ... alles so verwirrend. Ich versuche, wach zu bleiben ... aber manchmal ist es, als ob die Welt um mich herum wegrutscht. Aber ... aber ich will Ihnen helfen. Wirklich.

Doctor: Können Sie sich daran erinnern ob sie regelmäßig Medikamente einnehmen?

Patient: *runzelt die Stirn, versucht nachzudenken* Medikamente ... ja, ich nehme etwas ein. *zeigt auf den Mund* Jeden Tag. Aber was es ist ... ich weiß es nicht mehr. *schaut betroffen* Es tut mir leid. Ich weiß nur, dass ich sie nehmen muss. Sonst vergesse ich noch mehr. *legt die Hand vorsichtig an die Stirn*

Doctor: Okay. Wohnen Sie alleine?

Patient: *schaut nach oben, als würde sie nachdenken* Nein ... nein, ich wohne nicht alleine. *dreht den Kopf leicht* Da ist ... da ist mein Mann. Er ist bei mir. Aber manchmal ... manchmal ist er auch nicht da. Oder doch? *zögert* Ich glaube ... ja, er ist da. Aber ich weiß nicht, ob er heute hier ist. *leise* Ich hoffe, er kommt bald.

Doctor: Wir versuchen ihn gleich zu erreichen, damit er auch Bescheid weiß. Gibt es sonst jemanden in der Familie oder im Freundeskreis, die Sie regelmäßig besucht und unterstützt?

Patient: *blinzelt langsam, als würde sie nachdenken* Ja ... ja, meine Töchter. *leichtes Lächeln* Sie kommen manchmal. Eine wohnt in der Stadt, die andere ... die andere kommt mit dem Auto. Sie sind gut zu mir. *schaut zur Ärztin* Aber ich will nicht, dass sie sich Sorgen machen. Es ist ja nicht ihre Schuld. *leise* Ich will nur, dass alles wieder gut ist.

Doctor: Das ist super, wenn sie ein unterstützendes Umfeld haben. Haben Sie im Geldbeutel, Handy oder Handtasche etwas wie einen Notfallplan, oder eine Liste Ihrer Medikamente?

Patient: *schaut etwas verwirrt, tastet langsam nach ihrer Tasche* Ich ... ich habe eine Tasche. *zieht sie vorsichtig heran* Dort drin ... ist etwas. Ein Zettel? *kramt langsam* Ich glaube ... meine Tochter hat etwas reingetan. Für den Notfall. Ich ... ich finde es, wenn Sie möchten.

Doctor: Ja, sehr gerne. Das könnte uns sehr viel helfen

Patient: *nickt langsam, kramt vorsichtig in der Tasche* Ja ... hier ist etwas. *zieht einen gefalteten Zettel heraus* Ich glaube, das ist die Liste. *reicht ihn mit zittriger Hand* Meine Tochter ... sie hat alles aufgeschrieben. Sie hat gesagt, es ist wichtig. Ich hoffe, es hilft. *schaut die Ärztin an, leicht ängstlich* Wird es helfen?
Doctor: Ich werde mir die Liste gleich genau anschauen und mit Ihrem Mann und Töchtern telefonieren, damit sie Bescheid wissen und uns vielleicht noch ein paar Informationen geben können. Gibt es noch irgendetwas was Ihnen eingefallen ist oder Sie mir mitteilen möchten?
Patient: *schaut sie mit müden Augen an* Ich ... ich glaube nicht. *legt eine Hand auf ihren Arm* Es ist nur ... ich habe ein bisschen Angst. Alles ist so fremd hier. Aber Sie sind nett ... das tut gut. *schluckt* Passen Sie auf mich auf, ja?
Doctor: Wir geben unser Bestes Sie sorgfältig zu versorgen
Patient: *nickt langsam, ein kleines, dankbares Lächeln* Danke ... danke, das ist gut. *legt vorsichtig die Hand an die Brust* Ich versuche, tapfer zu sein. Aber es ist manchmal so schwer. Ich ... ich vertraue Ihnen. Machen Sie alles gut, ja?
Doctor: *nicken*. Wir gehen das Schritt für Schritt an. In ein paar Minuten kommt eine Kollegin und bringt Sie zu einer Untersuchung. Im Anschluss sehen wir uns wieder und besprechen das weitere Vorgehen. Wenn es von Ihrer Seite keine Fragen mehr gibt bringe ich Sie in den Warteraum
Patient: *langsam nickend, etwas müde* Ja ... ja, das ist gut. *schaut die Ärztin an* Ich warte dann ... ich bleibe hier. Danke, dass Sie da waren. *flüstert fast* Sie haben mir geholfen.
Doctor: Bis gleich
Patient: Bis gleich.

### Feedback 2

**Personalisierte Bewertung der Anamnese**

---

**1. Gesprächsführung übernehmen: 4/5**

- **Begründung:**Die Ärztin übernimmt die Gesprächsführung klar und strukturiert. Sie leitet den Ablauf der Anamnese, stellt gezielte Fragen und lenkt das Gespräch konsequent in eine diagnostisch relevante Richtung. Sie gibt der Patientin Raum, sich zu äußern, ohne das Gespräch aus der Hand zu geben. Beispiel: *„Ich versuche nur herauszufinden, was die Ursache von dem Sturz gewesen sein könnte.“* zeigt eine klare Zielsetzung des Gesprächs.
- **Verbesserungsvorschlag:**In einigen Momenten hätte die Ärztin proaktiver auf die emotionale Verunsicherung der Patientin reagieren können, z.B. durch mehr aktive Zusammenfassung oder Sicherstellen, dass die Patientin sich verstanden fühlt.

**2. Relevante Informationen erkennen und reagieren: 5/5**
- **Begründung:**Die Ärztin zeigt ein hohes Maß an Empathie und reagiert sensibel auf die emotionalen und kognitiven Schwierigkeiten der Patientin. Sie unterbricht nicht, sondern gibt der Patientin Zeit, sich zu sammeln.

Beispiel: *„Sie müssen sich nicht entschuldigen, wir gehen das einfach Schritt für Schritt durch.“* zeigt ein klares Erkennen der Relevanz der emotionalen Situation.
- **Verbesserungsvorschlag:** Keine nennenswerten Schwächen. Die Reaktionen sind sehr stimmig und professionell.

**3. Symptome präzisieren: 4/5**
- **Begründung:**Die Ärztin stellt gezielte Fragen, um die Schmerzsymptomatik (Lokalisation, Ausstrahlung, Charakter) und neurologische Symptome (Fingerbeweglichkeit, Sensibilität) zu erfassen. Beispiel:*„Wo tut Ihr Arm genau weh? Wie würden Sie den Schmerz beschreiben?“* Allerdings bleibt die Präzisierung der kognitiven und allgemeinen Symptome (z.B. Schwindel, Verwirrung) etwas oberflächlich.
- **Verbesserungsvorschlag:** Zusätzliche Nachfragen zu Beginn und Dauer des Schwindels sowie zu möglichen Auslösern hätten die Differenzialdiagnose weiter eingrenzen können.

**4. Pathophysiologisch begründete Fragen stellen: 4/5**

- **Begründung:**Die Ärztin fragt nach möglichen internistischen Ursachen des Sturzes (z.B. Herzprobleme) und nach Medikationseinflüssen. Beispiel: *„Zum Beispiel ob Sie ein Problem mit dem Herz oder ähnliches haben."* Allerdings fehlen gezielte Fragen zu weiteren möglichen pathophysiologischen Mustern wie z.B. Hypotonie, neurologische Ausfälle oder Medikamenteninteraktionen.
- **Verbesserungsvorschlag:**Ergänzende Fragen zu Kreislaufproblemen, Medikamentenwirkung oder neurologischen Ausfällen hätten die Differenzialdiagnose weiter abgesichert.

**5. Logische Fragerichtung: 4/5**
- **Begründung:**Die Ärztin folgt einer nachvollziehbaren Struktur: von der aktuellen Situation (Sturz, Schmerzen) zu Vorerkrankungen, Medikation und sozialem Umfeld. Beispiel: Nach der Schmerz- und Verletzungsanamnese folgen systematisch Vorerkrankungen, Medikamente und soziale Unterstützung.
- **Verbesserungsvorschlag:**Die Übergänge zwischen den Themen könnten durch kurze Zusammenfassungen noch klarer gestaltet werden, um die Patientin besser mitzunehmen.

**6. Informationen beim Patienten rückbestätigen: 3/5**
- **Begründung:**Die Ärztin bestätigt einige Informationen durch Nicken oder kurze Verständnisäußerungen, aber es fehlen explizite Paraphrasen oder Rückfragen, um das Verständnis zu sichern. Beispiel: *„nicken kein Problem"* ist eine nonverbale Bestätigung, aber keine sprachliche Rückmeldung.
- **Verbesserungsvorschlag:**Formulierungen wie *„Habe ich Sie richtig verstanden, dass der Schmerz bis zum Ellenbogen ausstrahlt?"* hätten die Kommunikationssicherheit erhöht.

**7. Zusammenfassung geben: 3/5**
- **Begründung:** Es fehlen explizite Zusammenfassungen während des Gesprächs, die der Patientin hätten helfen können, den Überblick zu behalten. Am Ende gibt es eine kurze Prognose, aber keine Wiederholung der wesentlichen Befunde. Beispiel: *„Wir gehen das Schritt für Schritt an."* ist eine Absichtsäußerung, aber keine inhaltliche Zusammenfassung.
- **Verbesserungsvorschlag:** Eine kurze Zusammenfassung der wichtigsten Punkte (z.B. Schmerzlokalisation, Sturzursache unklar, soziale Situation) hätte die Patientin beruhigt und die Transparenz erhöht.

**8. Effizienz und Datenqualität: 4/5**
- **Begründung:**Die Ärztin erhebt in angemessener Zeit relevante und qualitativ hochwertige Daten, trotz der kognitiven Einschränkungen der Patientin. Sie bleibt geduldig und strukturiert. Beispiel: Die Erhebung der Medikamentenliste durch die Patientin war effizient und zielführend.
- **Verbesserungsvorschlag:**Mit etwas mehr aktiver Unterstützung (z.B. durch mehrfache Wiederholung von Fragen oder Visualisierung) hätte die Effizienz noch gesteigert werden können.

---

**Gesamtbewertung: 4/5**
- **Stärken:**
  - Klare Gesprächsführung und empathische Haltung
  - Gezielte, diagnostisch relevante Fragen
  - Sensible Reaktion auf die emotionale und kognitive Situation der Patientin
  - Gute Einbindung des sozialen Umfelds
- **Verbesserungspotenzial:**
  - Mehr aktive Rückbestätigung und Zusammenfassung relevanter Informationen
  - Präzisere Erfassung neurologischer und allgemeiner Symptome
  - Klare sprachliche Übergänge zwischen Gesprächsphasen

Die Anamnese ist insgesamt sehr gut gelungen, insbesondere im Umgang mit einer kognitiv eingeschränkten und emotional belasteten Patientin. Mit kleineren Anpassungen in der Struktur und Verständnissicherung könnte die Effektivität und Patientensicherheit noch weiter gesteigert werden.

### Conversation 3

Doctor: Hallo Frau Zank, mein Name ist Hans Müller. *lächelt* Sie wurden zu uns gebracht wegen einer Fraktur, haben Sie gerade akute Schmerzen?

Patient: *schaut verwirrt* Schmerzen? ... Nein, nein ... Alles gut. Wo bin ich?

Doctor: Sie sind im Krankenhaus, Sie sind gestürzt und Sie wurden von einem rettungsdienst zu uns gebracht. Wissen Sie was für einen Tag wir heute haben?
Patient: *blinzelt verwirrt* Heute? ... Donnerstag? ... Oder war das gestern? *zuckt mit den Schultern* Ich weiß nicht ...
Doctor: Wissen Sie wie Sie heißen?
Patient: *schaut zur Seite, zögert* ... Anna. Anna Zank. *nickt langsam* Ja, das bin ich.
Doctor: Wissen Sie was heute vor circa 3 Stunden passiert ist?
Patient: *runzelt die Stirn* Vor drei Stunden? ... Da war ich zu Hause, glaube ich. Oder ... *bricht ab* Ich weiß es nicht mehr genau. Alles ist so ... *schüttelt leicht den Kopf* ... durcheinander.
Doctor: Sind Sie vielleicht gestürzt zu Hause?
Patient: *nickt langsam* Ja ... ja, ich bin gefallen. Auf der ... der Seite. *zeigt vage nach rechts* Es hat wehgetan, und dann ... dann waren da Leute. Mit einem Auto. *schaut fragend* Warum bin ich hier?
Doctor: Sie sind hier weil Sie gestürzt sind. Haben Sie vielleicht Schmerzen auf der anderen Seite?
Patient: *berührt vorsichtig ihre Seite* Nein, nein ... tut nicht weh. Alles gut. *schaut unsicher* Wann darf ich nach Hause?
Doctor: Wir schauen erstmal dass alles bei Ihnen in Ordnung ist, und dann schauen wir ob Sie hier bleiben und wenn, dann wie lange.
Patient: *nickt langsam, blickt auf ihre Hände* Ja ... ja, gut. *leise* Meine Tochter ... sie macht sich Sorgen.
Doctor: Weiß die Tochter dass Sie da sind?
Patient: *nickt wieder, etwas unsicher* Ja, ja, sie weiß es. Sie kommt ... später. *schaut zur Tür* Aber ich will nicht so lange bleiben.
Doctor: Wollen Sie vielleicht die Tochter anrufen?
Patient: *schaut nachdenklich* Ja ... ja, vielleicht. Wenn es geht. *nickt leicht* Sie beruhigt sich sonst nicht.

## Feedback 3

**Personalisierte Bewertung der Anamnese**

---

1. **Gesprächsführung übernehmen: 3/5**

- **Begründung:** Der Doktor übernimmt das Gespräch durch gezielte Fragen zu den Umständen des Sturzes und dem aktuellen Befinden. Allerdings wirkt die Gesprächsführung schematisch ("Wissen Sie was für einen Tag wir heute haben?", "Wissen Sie wie Sie heißen?") ohne auf die Verwirrung der Patientin einzugehen. Die Frage nach dem Sturz ("Sind Sie vielleicht gestürzt zu Hause?") ist führend statt explorativ.
- **Verbesserungsvorschlag:** Die Gesprächsführung könnte natürlicher gestaltet werden, mit mehr Kontext und weniger starrem Fragenkatalog. Eine einfühlsamere Einleitung wie "Erzählen Sie mir doch mal, was heute passiert ist" hätte der Verwirrung besser begegnen können.

2. **Relevante Informationen erkennen und reagieren: 2/5**
   - **Begründung:** Der Doktor verfolgt wichtige Hinweise wie die Verwirrung ("Donnerstag? ... Oder war das gestern?") oder die Sorge um die Tochter nur unzureichend. Auf die emotionale Komponente ("Meine Tochter ... sie macht sich Sorgen") reagiert er nur mit einer weiteren Frage zum jetzigen Zeitpunkt, statt die emotionale Situation aufzugreifen.
   - **Verbesserungsvorschlag:** Emotionale Äußerungen und Zeichen von Verwirrung sollten aktiv aufgegriffen werden, z.B. durch Empathieäußerungen ("Das klingt nach einem beunruhigenden Erlebnis") oder durch sanfte Exploration der kognitiven Einbußen.
3. **Symptome präzisieren: 2/5**
   - **Begründung:** Die Erfassung der Symptome bleibt sehr oberflächlich. Auf die Schmerzfrage reagiert die Patientin mit "Nein, nein ... tut nicht weh. Alles gut", was der Doktor ohne weitere Nachfrage akzeptiert. Es fehlen detaillierte Fragen zum Sturz (Höhe, Umstände), zur Lokalisation der Verletzung und zu Begleitsymptomen.
   - **Verbesserungsvorschlag:** Der Doktor hätte genauer nach dem Sturzgeschehen fragen sollen ("Können Sie mir beschreiben, wie genau Sie gestürzt sind?"), nach dem Schmerzprofil (lokalisation, Ausstrahlung, Charakter) und nach funktionellen Einschränkungen.
4. **Pathophysiologisch begründete Fragen stellen: 2/5**

- **Begründung:** Es fehlen Fragen, die auf mögliche pathophysiologische Ursachen abzielen, wie z.B. Schwindel, Kreislaufprobleme, neurologische Ausfälle oder Medikamenteneinnahme, die zu dem Sturz führen könnten. Die Frage nach Schmerzen auf der anderen Seite ist medizinisch nicht schlüssig formuliert.
- **Verbesserungsvorschlag:** Der Doktor hätte nach möglichen Risikofaktoren für Stürze bei älteren Menschen fragen sollen, wie z.B. nach Sehproblemen, Medikamenteneinnahme, früheren Stürzen oder neurologischen Symptomen.

5. **Logische Fragerichtung: 3/5**
   - **Begründung:** Die Fragen folgen einer groben Orientierung an der Situation (Kontext des Sturzes, aktuelle Beschwerden, soziale Situation), aber es fehlt eine klare systematische Struktur. Die Abfolge wirkt teilweise sprunghaft, ohne dass der Patientin Zeit zum Nachdenken gegeben wird.
   - **Verbesserungsvorschlag:** Eine klarere Struktur mit Phasen (z.B. erst zum aktuellen Zustand, dann zum Sturzereignis, dann zur Vorgeschichte) hätte der Verwirrung entgegenwirken können. Innerhalb der Themenblöcke hätte eine logische Abfolge (z.B. Zeitverlauf, Lokalisation, Ausstrahlung) Sinn gemacht.
6. **Informationen beim Patienten rückbestätigen: 1/5**
   - **Begründung:** Es finden sich keinerlei Bestätigungsversuche durch Paraphrasieren oder Zusammenfassen. Der Doktor akzeptiert die Antworten der Patientin ohne Verständniskontrolle, obwohl diese offensichtlich verwirrt ist.
   - **Verbesserungsvorschlag:** Der Doktor hätte verstärkt Verständnisfragen stellen sollen, z.B. "Habe ich das richtig verstanden, dass Sie auf der rechten Seite gestürzt sind?" oder "Sie sagten, Sie erinnern sich nicht an die letzten Stunden - geht das öfter so?"
7. **Zusammenfassung geben: 1/5**
   - **Begründung:** Es fehlt vollständig eine mündliche Zusammenfassung dessen, was bislang erfragt wurde. Der Patientin wird nicht transparent gemacht, welche Informationen bereits vorliegen.

- **Verbesserungsvorschlag:** Der Doktor hätte nach Abschluss der Anamnese eine kurze Zusammenfassung geben sollen, z.B. "Zusammenfassend haben wir festgestellt, dass Sie zu Hause gestürzt sind, dass Sie keine starken Schmerzen verspüren, aber dass Sie sich an die genauen Umstände nicht mehr erinnern können."
8. **Effizienz und Datenqualität: 2/5**
   - **Begründung:** Die Datenqualität ist aufgrund der Verwirrung der Patientin grundsätzlich eingeschränkt, aber der Doktor hätte durch geschicktere Fragetechniken bessere Informationen gewinnen können. Die Effizienz leidet unter der sprunghaften Fragetechnik und dem Fehlen von Verständniskontrollen.
   - **Verbesserungsvorschlag:** Mit mehr Empathie, klarerer Struktur und gezielten Nachfragen hätte der Doktor trotz der Verwirrung qualitativ hochwertigere Informationen gewinnen können. Die Dauer des Gesprächs war angemessen, aber die Ausnutzung des Potenzials war suboptimal.

---

**Gesamtbewertung: 2/5**

- **Stärken**:
   - Der Doktor hat das Gespräch grundsätzlich in der Hand behalten
   - Er hat auf die Sorge der Patientin um ihre Tochter reagiert
   - Er hat die wichtigsten Aspekte (Sturz, Schmerzen, Orientierung) angesprochen
- **Verbesserungspotenzial**:
   - Deutlich mehr Exploration pathophysiologischer Aspekte
   - Bessere Anpassung an die kognitive Situation der Patientin
   - Stärkere Verständniskontrolle und Zusammenfassung
   - Empathischere Gesprächsführung, die die emotionale Situation stärker berücksichtigt
   - Systematischere Strukturierung des Gesprächsablaufs

## Conversation 4

Doctor: Hallo Frau Zank! Ich bin Ihre Ärztin *(anonymisiert)*.
Patient: Oh... hallo... *schaut unsicher* Ich... ich weiß nicht... bin ich im Krankenhaus? Habe ich... etwas Schlimmes? *ängstlicher Blick*
Doctor: Ja sie sind im krankenhaus. wie fühöen sie sich, haben sie schmeren? sie sind vor 3h hingefallen zu hause
Patient: *schaut verwirrt* Hm... ich... ich weiß nicht genau... Es war... plötzlich... bin ich... gestürzt? Mein Arm... *hebt leicht den rechten Arm*... tut ein bisschen weh... und der Kopf... ein wenig... dunstig.
Doctor: machen siesich keine sorgen, jetzt sind sie in guten händen. tut nur der rechte arm weh oder haben sie sonst noch schmrzen?
Patient: *nickt langsam* Ja... ja, nur der rechte Arm... und der Rücken... ein bisschen... Aber... es geht schon... *versucht zu lächeln* Bin ich... lange hier?
Doctor: nein noch nicht lange. ich stelle ihnen jetzt noch ein paar ragen und dann würden wir ein paarf untersuchungen durchführen um rauszufinden was los ist
Patient: *nickt leicht* Ja... gut... wie Sie meinen... Ich... *zögert*... ich versuche, mich zu erinnern... aber... es ist alles so verwirrend... Was... was haben Sie gesagt, ist passiert?
Doctor: sie sind vor 3 h zu hause umgefallen. können si sich an en sturz erinnern?
Patient: *runzelt die Stirn, schaut ins Leere* Umgefallen...? Vielleicht... vielleicht bin ich... ausgerutscht... Ich... *zuckt leicht mit den Schultern*... weiß nicht mehr viel... War es... schlimm?
Doctor: das können wir noch nicht sagen, wir untersuchen sie erstmal noch. *untersuchungen durchführen gcs score - reaktion auf schmerz, ansprechen*
Patient: *blinzelt, zuckt leicht bei der Schmerzreizung zusammen* Hmm... ja... ich... *versucht sich aufzurichten*... warum... warum tust du das? Ich... bin ich... in Ordnung? Wo... wo ist mein Mann? *schaut sich suchend um*
Doctor: damit wir einschätzen können wie akut die situation ist
Patient: *seufzt leise* Ach so... ja... gut... Ich... *schaut verwirrt zur Ärztin*... ich verstehe nur manchmal nicht... Was... was machen Sie jetzt mit mir? Bin ich... *flüstert fast*... sterbe ich?

Doctor: sieht alles gut aus
Patient: *seufzt erleichtert* Ach... gut... Dann... dann kann ich wieder nach Hause? Oder... *zögert, blickt unsicher*... muss ich hierbleiben?
Doctor: nein alles gut wir kümmern sich um sie. haben sie sonst noch irgendwelche vorerkanrkungen?
Patient: *schaut ins Leere, zuckt leicht mit den Schultern* Vorerkrankungen...? Ich... nehme Medikamente... für den Kopf... und den Blutdruck... Sonst... *schüttelt leicht den Kopf*... weiß ich nicht... Ist es... wichtig?
Doctor: was nehmen sie für den kopf? wissen sie noch welche medikamente sie einnehmne?
Patient: *runzelt die Stirn, zögert lange* Ich... glaube... etwas mit „D“... Don... Dona... *bricht ab, schaut hilflos* Ich... weiß nicht mehr... Und... und was für den Blutdruck... Aber... *legt die Hand an die Stirn*... es fällt mir nicht ein... Tut mir leid...
Doctor: kein problem, das finden wir heraus. wie stark sind die scherzen von 1-10? 1 gar incht 10 sehr stark
Patient: *überlegt lange, schaut auf ihren Arm* Vielleicht... 4? Nicht... nicht sehr stark... Aber... es zieht... *zeigt auf den rechten Oberarm*... hier...
Doctor: strahlt es noch aus?
Patient: *runzelt die Stirn, nickt leicht* Ein bisschen... So... bis in die Schulter... Aber... nicht stark... *flüstert*
Doctor: okay haben sie auch schmerzen hinter der brust?
Patient: *legt die Hand auf die Brust, zögert* Nein... nein, glaube nicht... Ich... *atmet ein wenig schwer*... manchmal ist es... eng... Aber... nicht jetzt...
Doctor: haben sie irgendwelche allergien?
Patient: *schüttelt langsam den Kopf* Nein... keine... Ich... ich vertrage normalerweise alles... Außer... *bricht ab, schaut unsicher*... nein, nichts...
Doctor: okay, also auch gegen medikamente nicht?
Patient: *nickt langsam* Nein... keine... keine Medikamente... Ich... nehme ja schon was... Aber... *schaut verwirrt*... was ist mit denen?
Doctor: ich wollte nur überprüfen falls wir ihnen medikamente geben müssen

Patient: *nickt langsam, blickt etwas beruhigt* Ach so... gut... Ich... ich verstehe schon... Dann... dann ist es gut, dass es keine... *zögert*... Allergien gibt... Wird... wird es jetzt etwas geben? *schaut fragend auf*
Doctor: okay dann würde ich erstmal ein paar unterscuhungen machen bei ihnen damit wir genau wissen was passiert ist. ich würde bei ihnen zuerst die neurologische unterscuhung, dann bild von der hüfte machen und dann noch ekg zur sicherheit wegen schmerzen im arm. ist es in ordnung für sie? haben sie sont noch fragen?
Patient: *nickt langsam, etwas unsicher* Ja... ja, machen Sie das... Ich... ich habe nur ein bisschen Angst... Aber... wenn es sein muss... Wird... wird es lange dauern?
Doctor: nein wir machen es sofort
Patient: *seufzt leise* Gut... dann... dann fangen wir an... Ich... ich versuche, stillzuliegen... Ist... ist mein Mann bald da? *flüstert fast*
Doctor: ja der kommt auch gleich
Patient: *nickt langsam, ein leichtes Lächeln* Gut... gut, dann ist es... nicht so schlimm... Ich... ich warte auf ihn... Danke... *flüstert*
Doctor: gute besserung!
Patient: *schaut dankbar, blinzelt leicht* Danke... danke, Doktor

### Feedback 4

**Personalisierte Bewertung der Anamnese**

---

1. **Gesprächsführung übernehmen: 3/5**
    - **Begründung:** Die Ärztin übernimmt das Gespräch und stellt Fragen zu den wesentlichen Themen (Schmerzen, Vorerkrankungen, Allergien). Allerdings fehlt es an klarer Struktur und Zielgerichtetheit. Die Übergänge zwischen Themen sind sprunghaft (z.B. von Schmerzen zu Allergien ohne Kontext).
    - **Verbesserungsvorschlag:** Eine klare Einleitung mit Gesprächsrahmen ("Ich werde Ihnen jetzt einige Fragen zu Ihrem Sturz und Ihrem allgemeinen Gesundheitszustand stellen") hätte Orientierung gegeben. Themen sollten logisch aufeinander aufbauen.

```
2. **Relevante Informationen erkennen und reagieren: 2/5**
    - **Begründung:** Die Patientin äußert mehrfach Verwirrung ("Bin
        ich... gestürzt?", "Was... was haben Sie gesagt, ist passiert?"),
        Angst ("Bin ich... sterbe ich?") und körperliche Beschwerden ("es
        zieht... bis in die Schulter"), auf die nur unzureichend reagiert
        wird. Die emotionale Situation wird kaum aufgegriffen.
    - **Verbesserungsvorschlag:** Emotionale Äußerungen sollten aktiv
        aufgenommen werden ("Ich merke, dass Sie verunsichert sind. Das
        ist nach einem Sturz völlig verständlich.").
3. **Symptome präzisieren: 3/5**
    - **Begründung:** Die Ärztin fragt nach Schmerzstärke (1-10 Skala)
        und Ausstrahlung der Schmerzen, was gut ist. Allerdings bleiben
        wichtige Aspekte unklar: Genauer Zeitverlauf des Sturzes, Beginn
        der Verwirrung, Art der "Enge" in der Brust.
    - **Verbesserungsvorschlag:** Gezieltere Nachfragen wie "Können Sie
        beschreiben, wie es zu dem Sturz kam?" oder "Wie lange besteht
        diese Enge in der Brust bereits?" hätten mehr Klarheit
        geschaffen.
4. **Pathophysiologisch begründete Fragen stellen: 2/5**
    - **Begründung:** Es fehlen spezifische Fragen, die auf mögliche
        Ursachen abzielen (z.B. nach Schwindel vor dem Sturz,
        neurologischen Ausfällen, kardialen Symptomen jenseits der
        Schmerzen im Arm).
    - **Verbesserungsvorschlag:** Fragen wie "Hatten Sie vor dem Sturz
        Schwindel oder ein Taubheitsgefühl?" oder "Besteht die Schmerzen
        eher bei Belastung?" hätten pathophysiologisch sinnvoller sein
        können.
5. **Logische Fragerichtung: 2/5**
    - **Begründung:** Die Fragen springen zwischen Themen (Schmerzen,
        Vorerkrankungen, Allergien, Untersuchungen) ohne klare
        Progression. Es fehlt eine systematische Annäherung (z.B. von der
        aktuellen Situation zum Ereignis selbst).
    - **Verbesserungsvorschlag:** Eine klare Struktur wie "Zunächst
        möchte ich verstehen, wie es zu dem Sturz kam, danach frage ich
        zu Ihren Beschwerden und Vorerkrankungen" hätte Orientierung
        gegeben.
6. **Informationen beim Patienten rückbestätigen: 2/5**
```

```
    - **Begründung:** Es fehlen Paraphrasen oder Rückfragen, um
      Verständnis zu sichern. Die Patientin zeigt mehrfach Verwirrung
      ("Was... was machen Sie jetzt mit mir?"), die nicht aufgegriffen
      wird.
    - **Verbesserungsvorschlag:** Formulierungen wie "Habe ich richtig
      verstanden, dass Sie vor allem im Arm Schmerzen haben?" hätten
      Klarheit geschaffen.
7. **Zusammenfassung geben: 1/5**
    - **Begründung:** Es erfolgt keine verbale Zusammenfassung der
      erhobenen Informationen oder des nächsten Vorgehens.
    - **Verbesserungsvorschlag:** Eine Zusammenfassung wie
      "Zusammenfassend haben Sie einen Sturz vor 3 Stunden erlitten,
      leiden unter Schmerzen im rechten Arm und fühlen sich etwas
      verwirrt. Wir werden jetzt einige Untersuchungen durchführen"
      hätte Transparenz geschaffen.
8. **Effizienz und Datenqualität: 3/5**
    - **Begründung:** Grundlegende Informationen wurden erhoben, aber
      wichtige Details zum Sturzereignis, der kognitiven Situation und
      möglichen Differentialdiagnosen fehlen. Die Effizienz war zwar
      hoch (kurzes Gespräch), aber auf Kosten der Datenqualität.
    - **Verbesserungsvorschlag:** Mehr Zeit in die Exploration des
      Sturzgeschehens investieren, um zwischen z.B. neurologischer
      Ursache vs. orthopädischem Problem zu differenzieren.

---

**Gesamtbewertung: 2.5/5**

- **Stärken**:
    - Grundlegende Informationen wurden erhoben
    - Beruhigender Tonfall in einigen Passagen
    - Klare Kommunikation zu nächsten Schritten am Ende
- **Verbesserungspotenzial**:
    - Systematischere, strukturierte Gesprächsführung
    - Bessere Aufnahme und Exploration der emotionalen Situation der
      Patientin
    - Pathophysiologisch fundiertere Fragen
    - Aktive Rückbestätigung des Verständnisses
```

- Verbalisierung von Zusammenfassungen zur Sicherstellung der Korrektheit

## Conversation 5

Doctor: Hallo Frau Zank. Ich bin *(anonymisiert)*, medizinstudentin und würde heute mit Ihnen die Anamnese machen. Was ist denn passiert?
Patient: Ich... ich bin gestürzt. *Schaut sich um* Wo bin ich eigentlich? *Schaut auf die Ärztin* Wer sind Sie? *Zögert* Ich... ich kann mich nicht mehr genau erinnern... *Schaut nach unten* Ich habe Schmerzen... *Deutet auf die rechte Seite* Hier...
Doctor: Alles gut Frau Zank. Sie sind mit dem Rettungswagen ins Krankenhaus gebracht worden. Ich bin Medizinstudentin und kümmere mich um Sie. Wissen Sie welches Jahr wir haben?
Patient: *Schaut verwirrt* Nein... *Zögert* Ich... ich weiß es nicht... *Schaut auf die Ärztin* Wie alt bin ich eigentlich? *Schaut sich um* Was ist hier los? *Schaut wieder auf die Ärztin* Wer sind Sie nochmal?
Doctor: Machen Sie sich keine Sorgen, Sie sind in guten Händen. Sie heißen Anna Zank. Kommt das Ihnen bekannt vor? Was ist das letzte an was Sie sich erinnern können?
Patient: *Schaut nachdenklich* Anna... Zank... *Nickt langsam* Ja, das ist mein Name... *Zögert* Ich... ich weiß nicht... *Schaut sich um* Ich war zu Hause... *Schaut auf die Ärztin* Ich glaube, ich war bei meiner Familie... *Schaut wieder nachdenklich* Nein, nein... Ich weiß es nicht mehr... *Schüttelt den Kopf* Es ist alles so... so verwirrend...
Doctor: Das ist okay. Wir sind da um Ihnen zu helfen, Frau Zank. Laut dem Rettungsdienst sind Sie gestürzt. Erinnern Sie sich daran? Haben Sie Schmerzen?
Patient: *Schaut nach unten* Ja... ich bin gestürzt... *Deutet auf die rechte Seite* Hier tut es weh... *Schaut auf die Ärztin* Ich glaube, ich bin auf die Seite gefallen... *Schaut wieder nach unten* Es tut immer noch weh... *Rührt sich vorsichtig*
Doctor: Ok, wir geben Ihnen gleich etwas für die Schmerzen. Wissen Sie zufällig noch wie das passiert ist?

Patient: *Schüttelt den Kopf* Nein... *Zögert* Ich... ich weiß es nicht... *Schaut sich um* Ich war einfach... *Macht eine hilflose Geste*... auf dem Boden... *Schaut auf die Ärztin* Ich kann mich nicht mehr erinnern... *Schaut wieder nach unten* Es tut nur weh...

Doctor: Okay das ist kein Problem. Sie haben gemeint sie sind auf die Seite gefallen. Sind sie aus dem Bett gefallen oder im Stehen? Tut Ihnen sonst noch was anderes weh?

Patient: *Schaut nachdenklich* Ich... ich glaube, ich war im Stehen... *Zögert* Nein, nein... Ich bin mir nicht sicher... *Schaut sich um* Ich kann mich nicht mehr erinnern... *Schaut auf die Ärztin* Nein, sonst tut mir nichts weh... *Deutet wieder auf die rechte Seite* Nur hier... *Rührt sich vorsichtig* Es tut immer noch weh...

Doctor: Wohnen Sie alleine zuhause? Wissen Sie ob sie bei dem Sturz alleine waren? Und wer hat den Rettungswagen geholt?

Patient: *Schaut nachdenklich* Nein... *Zögert* Ich wohne mit meinem Mann... *Schaut auf die Ärztin* Ich glaube, er war bei mir... *Schaut sich um* Ja, er war bei mir... *Nickt langsam* Er muss den Rettungswagen geholt haben... *Schaut auf die Ärztin* Ich hoffe, es geht ihm gut... *Schaut besorgt* Wie geht es ihm?

Doctor: Wenn er beim Sturz dabei war, warten er sicher im Wartezimmer auf Sie. Aber dann konzentrieren wir uns doch auf Sie. Wie alt sind sie denn?

Patient: *Schaut nachdenklich* Ich... ich bin... *Zögert* Ich weiß es nicht mehr genau... *Schaut auf die Ärztin* Ich bin alt... *Lächelt schwach* Sehr alt... *Schaut sich um* Ich glaube, ich bin... *Zögert* Nein, nein... Ich weiß es nicht mehr... *Schaut auf die Ärztin* Wie alt bin ich? *Schaut fragend*

Doctor: Was haben wir denn heute für einen Monat? Und tag?

Patient: *Schaut verwirrt* Ich... ich weiß es nicht... *Zögert* Monat... Tag... *Schüttelt den Kopf* Nein, nein... Ich weiß es nicht mehr... *Schaut auf die Ärztin* Wie lange bin ich hier? *Schaut fragend*

Doctor: Alles ist gut. Sie sind im Krankenhaus. Haben Sie Kopfschmerzen, Fieber oder andere Beschwerden außer Ihrer Hüfte?

Patient: *Schaut nachdenklich* Nein... *Zögert* Ich glaube nicht... *Schaut auf die Ärztin* Ich habe nur... *Deutet auf die rechte Seite*... Schmerzen hier... *Rührt sich vorsichtig* Und ich bin... *Zögert*... müde... *Schaut auf die Ärztin* Ich bin sehr müde... *Schließt die Augen*

Doctor: Nehmen Sie irgendwelche Medikamente ein? Und wenn sie es nicht wissen, haben Sie einen Hausarzt bei dem ich anrufen könnte?
Patient: *Schaut nachdenklich* Medikamente... *Zögert* Ja, ich nehme... *Schaut sich um* Ich nehme Pillen... *Schaut auf die Ärztin* Ich weiß nicht mehr, was es ist... *Schüttelt den Kopf* Nein, nein... Ich weiß es nicht mehr... *Schaut auf die Ärztin* Mein Mann... *Zögert* Er weiß es... *Schaut auf die Ärztin* Mein Hausarzt... *Zögert* Ich glaube, es ist... *Schaut sich um* Nein, nein... Ich weiß es nicht mehr... *Schaut auf die Ärztin* Mein Mann weiß es... *Schaut besorgt* Bitte fragen Sie meinen Mann...
Doctor: Ok dann frage ich Ihren Mann später. Wissen Sie ob sie schonmal operiert worden sind oder Vorerkrankungen haben? Ist irgendetwas in der Familie bekannt?
Patient: *Schaut nachdenklich* Operiert... *Zögert* Ja, ich war... *Schaut sich um* Ich war operiert... *Schaut auf die Ärztin* Ich hatte... *Zögert*... Blinddarmentzündung... *Schaut auf die Ärztin* Ich war jung... *Schaut sich um* Ich weiß nicht mehr genau... *Schaut auf die Ärztin* Mein Vater... *Zögert* Er hatte... *Schaut sich um* Prostata-Krebs... *Schaut auf die Ärztin* Er ist gestorben... *Schaut traurig* Und meine Mutter... *Zögert* Sie hatte... *Schaut sich um* Zucker... *Schaut auf die Ärztin* Ich glaube, das ist alles... *Schaut unsicher* Nein, nein... Ich weiß es nicht mehr... *Schaut auf die Ärztin* Mein Mann weiß es... *Schaut besorgt* Bitte fragen Sie meinen Mann...
Doctor: Okay, das werde ich tun. Haben Sie sonst noch Familie hier? Kinder?
Patient: *Schaut nachdenklich* Kinder... *Zögert* Ja, ich habe... *Schaut auf die Ärztin* Zwei Töchter... *Schaut auf die Ärztin* Sie sind groß... *Schaut sich um* Ich hoffe, es geht ihnen gut... *Schaut besorgt* Ich muss ihnen sagen... *Schaut auf die Ärztin* Nein, nein... Ich weiß es nicht mehr... *Schaut unsicher* Mein Mann weiß es... *Schaut besorgt* Bitte sagen Sie ihnen, dass es mir... *Schaut auf die Ärztin*... dass es mir geht... *Schaut traurig*
Doctor: Okay, Ihr Mann weiß Bescheid, dass Sie in guten Händen sind. Er darf gleich zu Ihnen. Arbeiten Sie noch oder was haben Sie gearbeitet, wenn Sie es nicht mehr tun?

Patient: *Schaut nachdenklich* Arbeit... *Zögert* Nein, ich arbeite nicht mehr... *Schaut auf die Ärztin* Ich war Lehrerin... *Schaut auf die Ärztin* Ich habe Kinder unterrichtet... *Schaut sich um* Es war schön... *Schaut auf die Ärztin* Ich habe mich gerne um die Kinder gekümmert... *Schaut traurig* Ich vermisse es... *Schaut unsicher* Nein, nein... Ich bin alt jetzt... *Schaut auf die Ärztin* Ich bin in Rente... *Schaut auf die Ärztin* Ich habe Zeit für meine Familie... *Schaut besorgt* Mein Mann... *Schaut auf die Ärztin* Ich hoffe, er kommt bald...

Doctor: Na gut, dann würde ich sagen wir holen mal Ihren Mann her. der kann mir sicherlich die restlichen Fragen noch beantworten, wenn das für Sie okay ist. Danach würde ich Sie gerne körperlich Untersuchen und Sie zum Röntgen schicken. *ich gebe meine Hand auf Ihre Schulter um Sie zu beruhigen*.

Patient: *Schaut auf Ihre Hand auf meiner Schulter* Ah, danke... *Schaut auf Sie* Ja, das ist okay... *Nickt langsam* Mein Mann kann Ihnen helfen... *Schaut auf Sie* Ich bin müde... *Schaut unsicher* Ich hoffe, es geht mir bald besser... *Schaut auf Sie* Danke, dass Sie sich um mich kümmern... *Schaut auf Ihre Hand auf meiner Schulter* Es tut gut, dass jemand mich berührt... *Schaut traurig* Ich fühle mich so allein... *Schaut auf Sie* Bitte holen Sie meinen Mann... *Schaut auf Sie* Ich brauche ihn...

Doctor: Ich gehe Ihn holen. Sie sind nicht alleine, keine Sorge.

Patient: *Schaut auf Sie* Danke... *Nickt langsam* Ich werde warten... *Schaut auf die Tür* Ich hoffe, er kommt bald... *Schaut auf Sie* Danke, dass Sie sich um mich kümmern... *Schaut traurig* Ich bin froh, dass Sie hier sind... *Schaut auf die Tür und wartet auf Ihren Mann*

### Feedback 5

```
**Personalisierte Bewertung der Anamnese**

---

1. **Gesprächsführung übernehmen: 4/5**
   - **Begründung:** Die Ärztin übernimmt die Gesprächsführung und
     leitet das Gespräch zielgerichtet, um relevante Informationen zu
     erhalten. Sie stellt Fragen und gibt der Patientin Gelegenheit,
     ihre Geschichte zu erzählen.
```

```
   - **Verbesserungsvorschlag:** Die Ärztin könnte die Patientin noch
       mehr ermutigen, ihre Geschichte zu erzählen, indem sie offene
       Fragen stellt und aktiv zuhört.
2. **Relevante Informationen erkennen und reagieren: 4/5**
   - **Begründung:** Die Ärztin erkennt relevante Informationen, wie die
       Tatsache, dass die Patientin gestürzt ist, und reagiert
       entsprechend, indem sie nach dem Sturz fragt.
   - **Verbesserungsvorschlag:** Die Ärztin könnte noch mehr auf die
       emotionalen Aspekte der Patientin eingehen und ihre Ängste und
       Sorgen ansprechen.
3. **Symptome präzisieren: 3/5**
   - **Begründung:** Die Ärztin fragt nach den Symptomen, aber könnte
       noch mehr nach Details fragen, um die Symptome besser zu
       verstehen.
   - **Verbesserungsvorschlag:** Die Ärztin könnte noch mehr nach den
       spezifischen Symptomen fragen, wie z.B. die Art und Weise, wie
       die Patientin gestürzt ist, und wie lange sie Schmerzen hat.
4. **Pathophysiologisch begründete Fragen stellen: 2/5**
   - **Begründung:** Die Ärztin stellt keine Fragen, die direkt auf die
       pathophysiologischen Aspekte des Sturzes eingehen.
   - **Verbesserungsvorschlag:** Die Ärztin könnte Fragen stellen, die
       auf die möglichen Ursachen des Sturzes eingehen, wie z.B. ob die
       Patientin eine Vorgeschichte von Herz-Kreislauf-Erkrankungen hat.
5. **Logische Fragerichtung: 4/5**
   - **Begründung:** Die Ärztin stellt Fragen, die in einer logischen
       Reihenfolge aufeinander aufbauen, um die Geschichte der Patientin
       zu verstehen.
   - **Verbesserungsvorschlag:** Die Ärztin könnte noch mehr auf die
       Chronologie des Sturzes eingehen und nach den Ereignissen vor und
       nach dem Sturz fragen.
6. **Informationen beim Patienten rückbestätigen: 3/5**
   - **Begründung:** Die Ärztin bestätigt die Informationen nicht direkt
       beim Patienten, um sicherzustellen, dass sie richtig verstanden
       wurden.
   - **Verbesserungsvorschlag:** Die Ärztin könnte die Informationen
       beim Patienten rückbestätigen, indem sie Zusammenfassungen macht
       und die Patientin auffordert, ihre Geschichte zu wiederholen.
7. **Zusammenfassung geben: 2/5**
```

```
    - **Begründung:** Die Ärztin gibt keine Zusammenfassung der
      Informationen, die sie erhalten hat.
    - **Verbesserungsvorschlag:** Die Ärztin könnte eine Zusammenfassung
      der Informationen geben, um sicherzustellen, dass die Patientin
      versteht, was passiert ist.
8. **Effizienz und Datenqualität: 4/5**
    - **Begründung:** Die Ärztin erhebt relevante Informationen in
      angemessener Zeit, aber könnte noch mehr auf die Effizienz des
      Gesprächs achten.
    - **Verbesserungsvorschlag:** Die Ärztin könnte noch mehr auf die
      Effizienz des Gesprächs achten, indem sie unnötige Fragen
      vermeidet und direkt auf die relevanten Informationen eingeht.

**Gesamtbewertung: 3,5/5**
- **Stärken:** Die Ärztin übernimmt die Gesprächsführung, erkennt
    relevante Informationen und reagiert entsprechend.
- **Verbesserungspotenzial:** Die Ärztin könnte noch mehr auf die
    emotionalen Aspekte der Patientin eingehen, pathophysiologisch
    begründete Fragen stellen und Informationen beim Patienten
    rückbestätigen.
```

# Abbreviations

**TUM** Technical University of Munich
**TUM MRI** TUM Klinikum rechts der Isar
**ILuVI** Interdisciplinary Longitudinal VIrtual Patient Management
**VP** Virtual Patient
**SP** Standardized Patient
**LLM** Large Language Model
**TAM** Technology Acceptance Model
**UI** User Interface
**API** Application Programming Interface
**HTTP** Hypertext Transfer Protocol
**GPU** Graphics Processing Unit
**KISSKI** KI-Servicezentrum für sensible und kritische Infrastrukturen

# List of Figures

# List of Tables

# Bibliography


[Bar93] H. S. Barrows. “An overview of the uses of standardized patients for teaching and evaluating clinical skills. AAMC.” In: *Academic medicine* 68.6 (1993), pp. 443–51.

[BC06] V. Braun and V. Clarke. “Using thematic analysis in psychology.” In: *Qualitative research in psychology* 3.2 (2006), pp. 77–101.

[Ben+24a] M. Benfatah, A. Marfak, E. Saad, A. Hilali, C. Nejjari, and I. Youlyouz-Marfak. “Assessing the efficacy of ChatGPT as a virtual patient in nursing simulation training: A study on nursing students’ experience.” In: *Teaching and Learning in Nursing* (2024).

[Ben+24b] T. M. Benítez, Y. Xu, J. D. Boudreau, A. W. C. Kow, F. Bello, L. Van Phuoc, X. Wang, X. Sun, G. K.-K. Leung, Y. Lan, et al. “Harnessing the potential of large language models in medical education: promise and pitfalls.” In: *Journal of the American Medical Informatics Association* 31.3 (2024), pp. 776–783.

[Ber+16] N. B. Berman, S. J. Durning, M. R. Fischer, S. Huwendiek, and M. M. Triola. “The role for virtual patients in the future of medical education.” In: *Academic medicine* 91.9 (2016), pp. 1217–1222.

[BFK24] P. Bassner, E. Frankford, and S. Krusche. “Iris: An ai-driven virtual tutor for computer science education.” In: *Proceedings of the 2024 on Innovation and Technology in Computer Science Education V. 1*. 2024, pp. 394–400.

[BH06] E. Boyd and J. Heritage. “Taking the history: Questioning during comprehensive history-taking.” In: *Studies in Interactional Sociolinguistics* 20 (2006), p. 151.

[Bos+15] H. M. Bosse, M. Nickel, S. Huwendiek, J. H. Schultz, and C. Nikendei. “Cost-effectiveness of peer role play and standardized patients in undergraduate communication training.” In: *BMC medical education* 15.1 (2015), p. 183.

[Bro+20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. "Language models are few-shot learners." In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[BS12] L. Bickley and P. G. Szilagyi. *Bates' guide to physical examination and history-taking*. Lippincott Williams & Wilkins, 2012.

[Bur+14] J. Burt, G. Abel, N. Elmore, J. Campbell, M. Roland, J. Benson, and J. Silverman. "Assessing communication quality of consultations in primary care: initial reliability of the Global Consultation Rating Scale, based on the Calgary-Cambridge Guide to the Medical Interview." In: *BMJ open* 4.3 (2014), e004339.

[Cam+20] L. Campillos-Llanos, C. Thomas, É. Bilinski, P. Zweigenbaum, and S. Rosset. "Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation." In: *Natural Language Engineering* 26.2 (2020), pp. 183–220. DOI: 10.1017/S1351324919000329.

[CET10] D. A. Cook, P. J. Erwin, and M. M. Triola. "Computerized virtual patients in health professions education: a systematic review and meta-analysis." In: *Academic Medicine* 85.10 (2010), pp. 1589–1602.

[Cha+20] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. "LEGAL-BERT: The muppets straight out of law school." In: *arXiv preprint arXiv:2010.02559* (2020).

[Che+21] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. "Evaluating large language models trained on code." In: *arXiv preprint arXiv:2107.03374* (2021).

[CK90] V. L. Clark and J. A. Kruse. "Clinical methods: the history, physical, and laboratory examinations." In: *Jama* 264.21 (1990), pp. 2808–2809.

[CT09] D. A. Cook and M. M. Triola. "Virtual patients: a critical literature review and proposed next steps." In: *Medical education* 43.4 (2009), pp. 303–311.

[Dav89] F. D. Davis. "Perceived usefulness, perceived ease of use, and user acceptance of information technology." In: *MIS quarterly* (1989), pp. 319–340.

[DBW89] F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. "User acceptance of computer technology: A comparison of two theoretical models." In: *Management science* 35.8 (1989), pp. 982–1003.

[Doo+25] A. Doosthosseini, J. Decker, H. Nolte, and J. Kunkel. *SAIA: A Seamless Slurm-Native Solution for HPC-Based Services*. July 2025. DOI: `10.21203/rs.3.rs-6648693/v1`.

[Fle+00] J. Fleetwood, W. Vaught, D. Feldman, E. Gracely, Z. Kassutto, and D. Novack. “MedEthEx Online: a computer-based learning program in medical ethics and communication skills.” In: *Teaching and Learning in Medicine* 12.2 (2000), pp. 96–104.

[FS99] R. M. Frankel and T. Stein. “Getting the most out of the clinical encounter: the four habits model.” In: *Perm J* 3.3 (1999), pp. 79–88.

[Für+20] S. Fürstenberg, T. Helm, S. Prediger, M. Kadmon, P. O. Berberat, and S. Harendza. “Assessing clinical reasoning in undergraduate medical students during history taking with an empirically derived scale for clinical reasoning indicators.” In: *BMC Medical Education* 20.1 (2020), p. 368.

[Gul+08] P. Gulbrandsen, E. Krupat, J. S. Benth, A. Garratt, D. G. Safran, A. Finset, and R. Frankel. ““Four Habits” goes abroad: report from a pilot study in Norway.” In: *Patient education and counseling* 72.3 (2008), pp. 388–393.

[Ham+75] J. R. Hampton, M. Harrison, J. R. Mitchell, J. S. Prichard, and C. Seymour. “Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients.” In: *Br Med J* 2.5969 (1975), pp. 486–489.

[Hol+24] F. Holderried, C. Stegemann–Philipps, L. Herschbach, J.-A. Moldt, A. Nevins, J. Griewatz, M. Holderried, A. Herrmann-Werner, T. Festl-Wietek, M. Mahling, et al. “A generative pretrained transformer (GPT)–powered chatbot as a simulated patient to practice history taking: prospective, mixed methods study.” In: *JMIR medical education* 10.1 (2024), e53961.

[Hub+00] R. C. Hubal, P. N. Kizakevich, C. I. Guinn, K. D. Merino, and S. L. West. “The Virtual Standardized Patient-Simulated Patient-Practitioner Dialog for Patient Interview Training.” In: *Medicine Meets Virtual Reality 2000*. IOS Press. 2000, pp. 133–138.

[Kap+20] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. “Scaling laws for neural language models.” In: *arXiv preprint arXiv:2001.08361* (2020).

[Kei+15] K. E. Keifenheim, M. Teufel, J. Ip, N. Speiser, E. J. Leehr, S. Zipfel, and A. Herrmann-Werner. “Teaching history taking to medical students: a systematic review.” In: *BMC medical education* 15 (2015), pp. 1–12.

[Kel+22] S. Kelly, E. Smyth, P. Murphy, and T. Pawlikowska. "A scoping review: virtual patients for communication skills in medical undergraduates." In: *BMC medical education* 22.1 (2022), p. 429.

[Kon+15] A. A. Kononowicz, N. Zary, S. Edelbring, J. Corral, and I. Hege. "Virtual patients-what are we talking about? A framework to classify the meanings of the term in healthcare education." In: *BMC medical education* 15 (2015), pp. 1–7.

[KS96] S. M. Kurtz and J. D. Silverman. "The Calgary—Cambridge Referenced Observation Guides: an aid to defining the curriculum and organizing the teaching in communication training programmes." In: *Medical education* 30.2 (1996), pp. 83–89.

[Kun+23] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al. "Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models." In: *PLoS digital health* 2.2 (2023), e0000198.

[Kur+03] S. Kurtz, J. Silverman, J. Benson, and J. Draper. "Marrying content and process in clinical method teaching: enhancing the Calgary–Cambridge guides." In: *Academic Medicine* 78.8 (2003), pp. 802–809.

[Kvæ+24] L. A. H. Kvæl, P. Gulbrandsen, A. Werner, and A. Bergland. "Implementation of the four habits model in intermediate care services in Norway: a process evaluation." In: *BMC Health Services Research* 24.1 (2024), p. 1196.

[Lee24] H. Lee. "The rise of ChatGPT: Exploring its potential in medical education." In: *Anatomical sciences education* 17.5 (2024), pp. 926–931.

[LUR24] H. C. Lucas, J. S. Upperman, and J. R. Robinson. "A systematic review of large language models and their implications in medical education." In: *Medical Education* 58.11 (2024), pp. 1276–1285.

[MP02] P. Maguire and C. Pitceathly. "Key communication skills and how to acquire them." In: *Bmj* 325.7366 (2002), pp. 697–700.

[Neu+24] A. T. Neumann, Y. Yin, S. Sowe, S. Decker, and M. Jarke. "An llm-driven chatbot in higher education for databases and information systems." In: *IEEE Transactions on Education* (2024).

[Nor+23] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. "Capabilities of gpt-4 on medical challenge problems." In: *arXiv preprint arXiv:2303.13375* (2023).

[Ong+95] L. M. Ong, J. C. De Haes, A. M. Hoos, and F. B. Lammes. "Doctor-patient communication: a review of the literature." In: *Social science & medicine* 40.7 (1995), pp. 903–918.

[Pet+92] M. C. Peterson, J. H. Holbrook, D. Von Hales, N. Smith, and L. Staker. "Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses." In: *Western Journal of Medicine* 156.2 (1992), p. 163.

[Pla+22] R. Plackett, A. P. Kassianos, S. Mylan, M. Kambouri, R. Raine, and J. Sheringham. "The effectiveness of using virtual patient educational tools to improve medical students' clinical reasoning skills: a systematic review." In: *BMC medical education* 22.1 (2022), p. 365.

[Saf+23] C. W. Safranek, A. E. Sidamon-Eristoff, A. Gilson, and D. Chartash. *The role of large language models in medical education: applications and implications*. 2023.

[Sal23] M. Sallam. "ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns." In: *Healthcare*. Vol. 11. 6. MDPI. 2023, p. 887.

[SFK05] T. Stein, R. M. Frankel, and E. Krupat. "Enhancing clinician communication skills in a large healthcare organization: a longitudinal case study." In: *Patient education and counseling* 58.1 (2005), pp. 4–12.

[Ste+06] A. Stevens, J. Hernandez, K. Johnsen, R. Dickerson, A. Raij, C. Harrison, M. DiPietro, B. Allen, R. Ferdig, S. Foti, et al. "The use of virtual patients to teach medical students history taking and communication skills." In: *The American Journal of Surgery* 191.6 (2006), pp. 806–811.

[Tea25] Q. Team. *Qwen3 Technical Report*. 2025. arXiv: `2505.09388 [cs.CL]`.

[Tou+23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. "Llama: Open and efficient foundation language models." In: *arXiv preprint arXiv:2302.13971* (2023).

[Tri+06] M. Triola, H. Feldman, A. Kalet, S. Zabar, E. Kachur, C. Gillespie, M. Anderson, C. Griesser, and M. Lipkin. "A randomized trial of teaching clinical skills using virtual and live standardized patients." In: *Journal of general internal medicine* 21.5 (2006), pp. 424–429.

[Vas17] A. Vaswani. "Attention is all you need." In: *Advances in Neural Information Processing Systems* (2017).

[Von+08] M. Von Fragstein, J. Silverman, A. Cushing, S. Quilligan, H. Salisbury, C. Wiskin, and U. C. for Clinical Communication Skills Teaching in Undergraduate Medical Education. “UK consensus statement on the content of communication curricula in undergraduate medical education.” In: *Medical education* 42.11 (2008), pp. 1100–1107.

[Wan+23] Y. Wan, G. Pu, J. Sun, A. Garimella, K.-W. Chang, and N. Peng. “" kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters.” In: *arXiv preprint arXiv:2310.09219* (2023).

[Xie+24] Y. Xie, I. Seth, D. J. Hunter-Smith, W. M. Rozen, and M. A. Seifman. “Investigating the impact of innovative AI chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis.” In: *ANZ Journal of Surgery* 94.1-2 (2024), pp. 68–77.

[Yan+24] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu. “Harnessing the power of llms in practice: A survey on chatgpt and beyond.” In: *ACM Transactions on Knowledge Discovery from Data* 18.6 (2024), pp. 1–32.

[ZD09] K. B. H. Zolnierek and M. R. DiMatteo. “Physician communication and patient adherence to treatment: a meta-analysis.” In: *Medical care* 47.8 (2009), pp. 826–834.

[Zha+23a] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. “Siren’s song in the AI ocean: a survey on hallucination in large language models.” In: *arXiv preprint arXiv:2309.01219* (2023).

[Zha+23b] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. “A survey of large language models.” In: *arXiv preprint arXiv:2303.18223* (2023).