FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Smart-Doc: AI-Powered Medical Education through Diagnostic Interactions

Bruno de Sena Pereira

Mestrado em Engenharia de Software

Supervisor: Prof. António Coelho

Second Supervisor: Prof. Nuno Rodrigues

November 14, 2025

Smart-Doc: AI-Powered Medical Education through Diagnostic Interactions

Bruno de Sena Pereira

Mestrado em Engenharia de Software

November 14, 2025

Resumo

Resumo

O erro de diagnóstico continua a ser uma causa relevante de dano evitável para os pacientes, frequentemente associado a enviesamentos cognitivos como a ancoragem, o viés de confirmação e o encerramento prematuro. A formação médica tradicional oferece poucas oportunidades, escaláveis e seguras, para treinar o reconhecimento e a correção destes desvios de raciocínio.

Esta dissertação investiga se interações com um paciente virtual (VP), sensíveis ao enviesamento e guiadas por intenção clínica, podem (i) suscitar e detetar enviesamentos durante o raciocínio diagnóstico e (ii) fomentar reflexão metacognitiva associada a maior exatidão diagnóstica. A hipótese central é que diálogos com divulgação progressiva, orientados pela intenção clínica, revelam marcadores mensuráveis de enviesamento e que a reflexão estruturada se associa a uma avaliação mais sistemática de hipóteses.

Para testar esta hipótese, foi desenvolvida a plataforma **Smart-Doc**, um paciente virtual suportado por IA. A solução combina diálogo não roteirizado com divulgação progressiva de informação baseada no reconhecimento de intenção clínica e integra deteção em tempo real de marcadores comportamentais de enviesamento através de análises híbridas (regras + semântica). O sistema suporta ainda reflexão metacognitiva pós-hoc, estruturada por prompts de feedback direcionados.

Realizou-se um estudo-piloto com seis médicos internos que resolveram um caso desenhado para induzir enviesamentos. Foram analisadas as trajetórias diagnósticas, a usabilidade e a carga de trabalho. O *Smart-Doc* reproduziu padrões de raciocínio propensos a erro: 50% dos participantes chegaram a diagnósticos incorretos, sobretudo por ancoragem em evidência precoce e reconciliação medicamentosa incompleta; os diagnósticos corretos associaram-se a avaliação sistemática de hipóteses e procura ativa de contra-evidência. A usabilidade foi aceitável (SUS: 66,5). A carga de trabalho (NASA-TLX) foi moderada e explicada sobretudo pela exigência temporal compatível com a latência do diálogo ($\sim$6 s por turno).

A arquitetura modular e case-agnostic da plataforma permite a autoria de novos casos clínicos sem reengenharia, apoiando a escalabilidade curricular. As principais limitações incluem caso único, amostra reduzida e dependência de pontuação automatizada; os prompts em tempo real foram desativados no piloto.

O trabalho futuro inclui redução de latência, introdução de prompts metacognitivos em tempo real com comparações controladas, expansão da biblioteca de casos e incorporação de adjudicação por peritos.

Em conjunto, os resultados sustentam a exequibilidade de interações VP sensíveis ao enviesamento para treino diagnóstico, posicionando o enviesamento cognitivo não apenas como fonte de erro, mas como uma oportunidade para aprendizagem estruturada e reflexão sistemática.

Abstract

Diagnostic error remains a leading cause of preventable patient harm, frequently driven by cognitive biases such as anchoring, confirmation bias, and premature closure. Traditional medical education offers limited, scalable opportunities to practice recognizing and correcting such reasoning flaws.

This dissertation investigates whether bias-aware, intent-driven virtual patient (VP) interactions can (i) elicit and detect cognitive bias during diagnostic reasoning and (ii) foster meta-cognitive reflection associated with improved diagnostic accuracy. The central hypothesis is that progressively disclosed VP dialogues, guided by clinical intent, will reveal measurable bias markers and that structured reflection will align with more systematic hypothesis evaluation.

To test this hypothesis, an AI-powered VP platform, Smart-Doc, was engineered as the experimental apparatus. Smart-Doc combines unscripted dialogue with progressive information disclosure based on clinical intent recognition and integrates real-time detection of behavioral bias markers via hybrid rule-based and semantic analyses. The system also supports structured, post-hoc meta-cognitive reflection through targeted feedback prompts.

A pilot study involving six clinical interns, who used a bias-engineered case, examined diagnostic trajectories, usability, and workload. Smart-Doc reproduced bias-prone reasoning patterns: 50% of participants reached incorrect diagnoses, predominantly due to anchoring on early evidence and incomplete medication reconciliation, whereas correct diagnoses were associated with systematic hypothesis testing and counter-evidence reasoning. Usability was acceptable (System Usability Scale: 66.5). Workload, measured with NASA-TLX, was moderate and driven primarily by temporal demand consistent with dialogue latency (6 s per turn).

The modular and case-agnostic architecture of the platform enables the authoring of new clinical cases without re-engineering, supporting the curricular scalability. Limitations include a single case, a small sample size, and reliance on automated scoring; real-time prompts were disabled during the pilot. Future work will reduce latency, introduce live meta-cognitive prompts with controlled comparisons, expand the case library, and incorporate expert adjudication.

Overall, the findings support the feasibility of bias-aware VP interactions for diagnostic training, positioning cognitive bias not merely as a source of error but as an opportunity for systematic, structured learning and reflection.

Acknowledgements

Antes de mais, gostava de agradecer à minha família, que, mesmo neste ano difícil, esteve sempre presente e disponível, com todo o apoio e amor incondicional. À minha **mãe**, por ser um verdadeiro exemplo de mulher guerreira e trabalhadora. Ao meu **pai**, por me proporcionar tudo sem nunca questionar, pelos seus valores inabaláveis e por sempre ter acreditado em mim. Ao meu **irmão**, que, apesar da distância, está comigo todos os dias, por ser uma pessoa tão querida e genuína.

À minha tia **Nelinha** e à avó **Laura**, que me aturaram diariamente na adolescência — santificadas sejam — e, em especial, ao meu avô **Sérgio**, que, embora tenha partido no início deste ano, permanece imortalizado nas minhas memórias mais íntimas.

Ao avô **Veríssimo** e à avó **Rosete**, por tudo o que construíram nesta vida para que os netos tivessem o que eles não tiveram.

Ao meu primo **João**, com quem cresci, como um irmão; à tia **Sandra**, à prima **Lili** e ao mais recente membro desta família, o **Dinis**, a quem desejo um futuro cheio de sucesso e felicidade.

Quero agradecer também ao meu orientador, o Professor **Nuno Rodrigues**, por toda a disponibilidade e mentoria, pela paciência e por ter acreditado em mim ao longo de todo este processo.

Aos meus companheiros de *fairway*, a *Sliice Nation* — **Sátão**, **Furos**, **Ribeiro** e **Tiro** — por me iniciarem no golfe, que acabou por se tornar o meu refúgio semanal, o meu momento de meditação. Obrigado por me acompanharem no orvalho matinal, com o vosso apoio incondicional, seja no *fairway*, no *green*, no *rough* ou no *bunker*.

Ao **Xico**, por todos estes anos de companheirismo, como colega de casa, sempre disponível para tudo.

Ao **Coelho**, por ser um grande amigo e por estar sempre por perto.

Aos amigos do *Água Faz Mal*, pela leveza e pela diversão que dão à vida.

Por último, mas não menos importante, ao meu girassol, a pediatra mais maravilhosa e competente — a **Xaninha** — que tem sido o meu pilar, que acredita em mim, me inspira e me faz querer ser melhor todos os dias. Obrigado pela tua paciência, pela tua compreensão e por todo o teu amor. Mereces o mundo.

Bruno de Sena Pereira

"As you walk down the fairway of life, you must smell the roses, for you only get to play one round."

Ben Hogan

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
API	Application Programming Interface
AR	Analytical Reasoning (System 2)
ASR	Automatic Speech Recognition
BNP	Brain Natriuretic Peptide
CDM	Clinical Decision-Making
CFS	Cognitive Forcing Strategies
CoT	Chain-of-Thought
COVID-19	Coronavirus Disease 2019
CSS	Cascading Style Sheets
CT	Computed Tomography
CUQ	Chatbot Usability Questionnaire
CXR	Chest X-Ray
GPT	Generative Pre-trained Transformer
Hb	Hemoglobin
HF	Heart Failure
HPI	History of Present Illness
ICC	Intraclass Correlation Coefficient
ILD	Interstitial Lung Disease
IQR	Interquartile Range
ITS	Intelligent Tutoring System
JSON	JavaScript Object Notation
LLM	Large Language Model

NAR	Non-Analytical Reasoning
NASA-TLX	National Aeronautics and Space Administration Task Load Index
NLP	Natural Language Processing
NLU	Natural Language Understanding
PE	Pulmonary Embolism
PMH	Past Medical History
QUIS	Questionnaire for User Interaction Satisfaction
RA	Rheumatoid Arthritis
RAG	Retrieval-Augmented Generation
RCT	Randomized Controlled Trial
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
RQ	Research Question
SQL	Structured Query Language
SUS	System Usability Scale
TB	Tuberculosis
TNF	Tumor Necrosis Factor
UI	User Interface
UX	User Experience
VP	Virtual Patient
VR	Virtual Reality
VSP	Virtual Standardized Patient
WBC	White Blood Cell

Chapter 1

Introduction

Diagnostic error remains a persistent threat to patient safety, with cognitive biases a major contributor to avoidable harm [21]. These biases—systematic deviations from rational judgment—can derail even experienced clinicians, leading to premature closure, selective information use, or unwarranted confidence in a favored hypothesis. As Croskerry [14] argues, traditional training often fails to help students recognize and mitigate these pitfalls under real-world pressure.

Transitioning from student to clinician requires robust clinical reasoning that integrates knowledge, pattern recognition, and reflective awareness [2]. Conventional approaches emphasize outcomes but rarely make the *process* of reasoning visible, let alone train learners to detect and counter bias in their own thinking. Addressing this gap demands pedagogy that elicits reasoning patterns, exposes bias, and scaffolds reflection—reliably and at scale.

This dissertation introduces **Smart-Doc**, an AI-powered virtual patient platform that meets this need. Leveraging large language models (LLMs), Smart-Doc provides naturalistic clinical interviews and embeds mechanisms to detect cognitive biases in real time. Beyond replicating encounters, it delivers structured meta-cognitive feedback that prompts deliberate reflection. In short, Smart-Doc moves from teaching *what* to diagnose toward teaching *how* to think—addressing root causes of diagnostic error in a controlled, scalable environment [3, 43].

1.1 Motivation

Two forces motivate this work. First, an educational gap: reasoning errors, not knowledge deficits, account for a substantial share of diagnostic mistakes, yet explicit support for bias awareness is rare [3]. Second, a technological opportunity: modern LLMs and principles from Intelligent Tutoring Systems enable simulations that combine realism with adaptive, reflective support. Smart-Doc sits at this intersection: it elicits naturalistic reasoning, surfaces bias-prone moments, and scaffolds reflective practice.

1.2 Research Questions

Building on the premise that diagnostic error is frequently linked to cognitive biases that shape how evidence is gathered and interpreted, the dissertation examines whether structured, bias-aware interaction with a virtual patient (VP) can make those biases visible and support better diagnostic reasoning. The guiding research question is:

> *How can AI-powered virtual patients be designed to help medical students recognize and mitigate cognitive biases in diagnostic reasoning?*

To operationalize this inquiry, three complementary research questions are articulated below.

1. **RQ1 — Elicitation and detection.** To what extent can an LLM-mediated diagnostic interview elicit and detect cognitive biases (specifically anchoring, confirmation bias, framing and premature closure) through observable interaction patterns?

2. **RQ2 — Metacognitive prompting.** How effective are meta-cognitive prompts—delivered either in real time or post hoc—in stimulating structured reflection (articulation of hypotheses, search for counter-evidence) and attenuating bias during case work-up?

3. **RQ3 — Scalability principles.** Which technical and pedagogical principles enable the scalable deployment of bias-aware VP simulations across curricula without re-engineering for each new case?

1.3 Objectives

Addressing these questions required the design of an experimental apparatus and an evaluation pipeline that links observable interaction patterns to learning outcomes. The work therefore advanced five objectives that together instantiate, instrument, and study the intervention.

Objective 1 — Simulation. Design and implement an AI-powered VP capable of realistic, unscripted clinical interviews, with cases intentionally engineered to surface bias-prone decision paths.

Objective 2 — Bias detection. Specify and compute behavioral markers of anchoring, confirmation bias, and premature closure using a hybrid approach that combines rule-based heuristics with LLM-assisted semantic analysis.

Objective 3 — Metacognitive tutoring. Deliver context-aware prompts that invite learners to articulate hypotheses, seek counter-evidence, and reconsider premature commitments when bias markers are detected.

Objective 4 — Evaluation framework. Develop analytics for diagnostic accuracy, information-gathering behavior, and manifestations of cognitive bias, providing both formative (feedback) and summative (scores) views.

Objective 5 — Effectiveness study. Conduct a pilot with clinical interns to characterize diagnostic trajectories, assess usability, and estimate workload, thereby gathering evidence relevant to accepting or rejecting the stated hypothesis.

Together, these objectives align the methodological apparatus (the VP platform and analytics) with the research questions, enabling a principled test of the central hypothesis introduced in this chapter.

1.4 Dissertation Structure

- **Chapter 2** presents the theoretical background: clinical reasoning, dual-process theory, cognitive bias taxonomy, debiasing strategies, and simulation principles that inform Smart-Doc.
- **Chapter 3** reviews AI-powered virtual patients, mapping feasibility, effectiveness, and design gaps that Smart-Doc addresses.
- **Chapter 4** details Smart-Doc's development, linking the two previous chapters—theory and evidence, to architecture, case design, and bias detection.
- **Chapter 5** evaluates Smart-Doc with clinical interns, reporting methods, results, and reflections on usability, realism, and educational impact.
- **Chapter 6** discusses implications, limitations, and directions for future research and curriculum integration.

With the problem, questions, and objectives established, the next step is to ground the study in theory. Chapter 2 introduces the cognitive foundations of clinical reasoning, the taxonomy of diagnostic bias, and debiasing strategies, then motivates AI-powered virtual patients as a pedagogical vehicle for Smart-Doc's design.

Chapter 2

Background

This chapter establishes the theoretical foundation for this dissertation. It introduces the cognitive mechanisms that underpin diagnostic reasoning, examines the pervasive role of cognitive biases, and situates simulation — particularly AI-powered virtual patients — as a promising medium for advancing medical education. By grounding the technical development of the Smart-Doc system in established cognitive science and educational principles, this chapter provides the conceptual scaffolding for the subsequent literature review (Chapter 3) and system design (Chapter 4).

2.1 The Cognitive Foundations of Medical Diagnosis

Clinical diagnosis is a demanding cognitive activity conducted under uncertainty and time pressure. Reasoning does not follow a linear pipeline. It proceeds through cycles of hypothesis formation, evidence gathering, and model revision, with constant negotiation between what is already known and what must still be learned from the patient. Understanding these mechanisms, along with their vulnerabilities, is crucial for designing effective educational interventions that enhance accuracy and safety.

2.1.1 The Diagnostic Process: An Iterative Quest for Coherency

The diagnostic encounter begins with sparse cues that prompt early hypotheses. These hypotheses guide a search for discriminating information and are repeatedly revised as new data emerge. In practice, clinicians alternate between abductive steps that generate explanations, inductive steps that generalize from findings, and deductive steps that test predictions of a working diagnosis [16, 18, 47]. The objective is not a perfect proof but a coherent account that best accommodates the totality of evidence available at the bedside.

Within this cycle, information acquisition is strategic rather than exhaustive. Clinicians prioritize questions and examinations with a high expected diagnostic yield, often beginning with the history of present illness and medication reconciliation before conducting targeted examinations and investigations [23, 49]. Progress depends on continuous updating of a problem representation that highlights the age, time course, salient risk factors, and key positive and negative findings

[56]. When alternative hypotheses are insufficiently considered, closure can occur prematurely. Educational designs should therefore cultivate habits of seeking counter evidence and revisiting earlier assumptions whenever discordant data appear [11].

Anamnesis as the Primary Diagnostic Instrument

The clinical interview supplies the majority of discriminative information for many presentations. A structured approach typically involves canvassing the history of present illness, past medical and surgical history, medications with explicit reconciliation, allergies, family and social history, and a systems review. Each component serves a different inferential role: medications and comorbidities refine pretest probabilities, social and occupational histories surface potential exposures, and the temporal pattern of symptoms refines the script that best explains the case [8, 49, 56]. Because narrative details shape early hypotheses, the interview is also the earliest point at which bias can take root or be corrected. Instruction that teaches learners to articulate a problem representation, to ask at least one question that could refute the leading hypothesis, and to reconcile medications systematically can reduce the risk of premature commitment [11, 18].

2.1.2 Dual-Process Theory in Clinical Reasoning

The cognitive architecture of diagnosis is often described with a dual-process account. One process is fast and intuitive, grounded in pattern recognition that draws on learned illness scripts and encapsulated knowledge [56]. The second process is slower and analytic, involving deliberate hypothesis testing, systematic search for disconfirming evidence, and explicit comparison of competing diagnoses [16, 47]. Expertise reflects flexible movement between these modes. Intuition supports efficiency in familiar conditions, whereas analysis is mobilized when presentations are atypical, when the stakes are high, or when uncertainty signals that the intuitive response may be unreliable [11].

From a pedagogical perspective, the goal is not to replace intuition but to calibrate it, recognizing situations that require a deliberate shift to analytic scrutiny, and to scaffold that shift with tools that make reasoning steps visible. Techniques include teaching problem representation, deliberate reflection on alternatives, and prompts that require articulation of counter-evidence before case closure [18, 44]. These methods align with the view that meta-cognitive oversight of reasoning is a learnable skill that can be trained in simulation and transferred to clinical practice.

Definitions.

System 1 (non analytical reasoning, NAR). Fast, intuitive, and automatic. Relies on pattern recognition and illness scripts. Reliable for common and familiar problems, but a frequent source of cognitive bias when heuristics are applied outside their range [11, 56].

System 2 (analytical reasoning, AR). Slow, deliberate, and resource intensive. Involves hypothesis testing, searching for disconfirming evidence, and explicit comparison of alternatives. Monitors and corrects intuitive responses when indicated [16, 47].

Diagnostic expertise lies in the fluid interplay between these systems: effective clinicians leverage System 1 for efficiency while engaging System 2 when uncertainty, complexity, or atypical presentations demand deeper scrutiny.

2.2 Cognitive Biases: The Pathologies of Clinical Reasoning

Although dual-process theory describes how reasoning ideally balances efficiency and accuracy, in practice, the system is fallible. Cognitive biases are systematic deviations from rational judgment that emerge from heuristic shortcuts. They are a leading contributor to diagnostic error and to preventable patient harm [21, 30, 59]. Biases rarely occur in isolation; they often interact with context, workload, and task framing, which makes clear definitions important for educational design.

2.2.1 A Taxonomy of Diagnostic Error: Identifying Common Cognitive Biases

Numerous biases have been cataloged in clinical reasoning research. The most prevalent include:

Table 2.1: Common Cognitive Biases in Clinical Diagnosis

Bias	Description and Clinical Implications
Anchoring	Fixating on initial impressions or salient features of the case, with insufficient adjustment when new, contradictory evidence emerges [14, 46].
Confirmation bias	Selectively seeking or interpreting information that confirms an existing hypothesis, while overlooking evidence that might refute it [3, 46].
Premature closure	Accepting a diagnosis before it has been adequately verified, halting further diagnostic exploration and increasing the risk of missed or incorrect diagnoses [21, 46].
Framing effect	Diagnostic reasoning shaped by how information is presented, steering attention toward certain explanations and away from others [14, 46].
Availability bias	Overestimating the likelihood of diagnoses that are recent, memorable, or vivid in experience [14, 21].
Overconfidence	Excessive faith in one's diagnostic accuracy, which reduces openness to alternatives and corrective feedback [4].
Search satisfaction	Stopping the diagnostic search once a plausible explanation has been found, resulting in missed comorbid or secondary conditions [14, 21].
Base-rate neglect	Underweighting disease prevalence and prior probabilities, which skews posterior judgments despite appropriate new data [47, 59].
Representativeness	Judging likelihood by similarity to a prototypical case, which can discount atypical presentations or comorbid states [18, 59].

Table 2.1 consolidates frequently cited cognitive biases in clinical reasoning and highlights their implications for diagnostic accuracy. While many biases have been described, the present work focuses on **anchoring**, **confirmation bias**, and **premature closure**, because they are common in clinical practice and amenable to instruction within simulation.

Measurement and operationalization. For the purposes of this dissertation, biases are treated as patterns in interaction and data use rather than as latent traits. Observable markers include selective information seeking aligned with a single hypothesis, early diagnostic commitment before key discriminators are queried, and disregard of disconfirming evidence. These markers are operationalized into detection rules and semantic probes that are applied to dialogue traces, with details provided in Chapter 4.

2.2.2 Debiasing Strategies: Interventions to Improve Clinical Judgment

Given the susceptibility of diagnostic reasoning to systematic error, a variety of debiasing approaches have been proposed. These approaches engage slower analytic processing to monitor, challenge, or override intuitive judgments, and can be grouped into complementary families: cognitive forcing strategies, structured reflection, diagnostic checklists and timeouts, simulation with targeted feedback, and computerized decision support.

Cognitive forcing strategies. Cognitive forcing strategies are metacognitive prompts that introduce a deliberate pause in intuitive reasoning, encouraging consideration of alternatives, search for disconfirming evidence, or reframing of the problem [12]. They can be taught at universal, generic, and specific levels, using prompts such as “What else could this be?” and “What finding would make my leading hypothesis unlikely?”

Structured reflection. Deliberate reflection is a structured procedure in which learners articulate competing diagnoses, list supporting and disconfirming features, and reassess the fit between data and diagnosis. It engages analytic processing and has been shown to improve recall of discriminating features and support more balanced reasoning, particularly when learners are taught how and when to apply it [35, 43, 45].

Diagnostic checklists and timeouts. Checklists provide structured prompts to broaden differential diagnosis, identify red flags, and prevent premature closure. Reviews report that checklists broaden information search and support verification, although effects on accuracy vary by context [17, 22]. Team-based diagnostic timeouts extend this idea by creating a deliberate pause to query assumptions and potential biases before committing to a diagnosis.

Simulation with just-in-time metacognitive feedback. Simulation environments trigger authentic information seeking and make reasoning processes observable. When combined with progressive disclosure and targeted prompts, simulation can surface bias-prone moments and provide

feedback that encourages reconsideration of alternatives [13, 22]. This aligns with experiential learning cycles in which action is followed by feedback and reflection, and it motivates the bias-aware design choices adopted for Smart-Doc.

Computerized decision support. Differential diagnosis generators and context-aware reminders can broaden consideration sets and support verification, especially for uncommon conditions. Their effectiveness depends on integration into workflow and on timely presentation of discriminating features [22, 47].

Summary and implications for design. Across these strategies, the common mechanism is deliberate engagement of analytic reasoning at critical junctures. Educational systems should therefore detect bias-prone patterns such as early closure or selective information seeking, interrupt with concise context-aware prompts, structure reflection on alternatives, and support verification through targeted tools. These principles inform the Smart-Doc architecture in Chapter 4, where they are translated into concrete design features.

2.3 Large Language Models for Educational Simulation

Modern natural language processing enables conversational agents that can participate in clinical teaching while making reasoning steps observable. Large language models (LLMs) provide the foundation for such agents, and their behavior can be shaped through prompting, alignment with human preferences, retrieval of external knowledge, and intent recognition for dialogue orchestration. The concepts surveyed here motivate the design choices implemented in Chapter 4.

2.3.1 Foundations of large language models

LLMs are neural sequence models that learn to predict the next token given prior context. Contemporary systems rely on the transformer architecture, which uses self-attention to model long-range dependencies efficiently [60]. After pretraining on very large text corpora, models acquire broad linguistic and factual competence, and performance scales with model size, data, and compute [5, 31]. Instruction tuning and preference learning further adapt these models to follow task descriptions and conversational norms [9, 48]. For educational simulation, two properties are central: the ability to sustain coherent multi-turn dialogue and the capacity to produce structured outputs that downstream analytics can interpret.

2.3.2 Techniques for controllable, bias-aware dialogue

Prompting for structure and reasoning. Prompts specify roles, context, and output formats. In a teaching setting, prompts can require intermediate reasoning steps or canonical output schemas that improve inspectability for feedback and automated scoring. Chain of thought prompting exposes stepwise reasoning [66]. Self-consistency samples multiple reasoning paths and aggregates

conclusions [65]. Lightweight instructions such as “think step by step” can improve transparency without large latency costs [33]. Structured outputs, for example, a JSON rubric with named fields, reduce ambiguity for analytics but increase verbosity, which must be balanced during live tutoring.

Alignment with human feedback. Instruction following behavior benefits from learning human preferences. Reinforcement learning from human feedback aligns outputs with what evaluators judge to be helpful and safe [48]. Direct preference optimization provides a simpler alternative that optimizes on ranked pairs without an explicit reward model [52]. In education, alignment supports adherence to pedagogical constraints, for example, remaining in character as a patient and refusing to short-circuit inquiry by giving definitive diagnoses.

Context management through retrieval. Parametric knowledge can be incomplete or outdated. Retrieval augmented generation combines generation with access to external sources, which grounds responses and reduces unsupported assertions [32, 37]. In medical education, retrieval is valuable for guideline citations and post-session reflection. During patient interviews, however, unrestricted retrieval may reveal answers that bypass the intended process of inquiry. For bias-aware simulation, retrieval is therefore scoped to feedback phases rather than real-time dialogue, which preserves the need to ask discriminating questions.

Intent recognition and orchestration. Task-oriented dialogue systems map free text to structured intents and slots. Transformer-based classifiers support this mapping and allow orchestration of multi-phase scenarios such as history taking, focused examination, and reflection prompts [40, 68]. Because safety is critical in health-related applications, systems should detect unknown or out-of-scope inputs and decline to act when confidence is low [38].

Tool use and grammar constrained decoding. Beyond intent recognition, function calling allows the model to request actions, for example, retrieving a lab value or advancing the case state, while emitting a schema that the platform validates [55, 67]. Grammar-constrained decoding restricts outputs to a prescribed format, which improves controllability and reduces ambiguity during analytics. These mechanisms align with progressive disclosure and phase gating used later in the platform.

2.3.3 Evaluation and risk management in educational use

Educational use introduces requirements that differ from general chat and therefore needs explicit evaluation. Quality should be assessed at three levels. First, *interaction quality* captures coherence, adherence to role, and responsiveness under a defined latency budget. Second, *educational quality* refers to whether the agent elicits discriminating questions, supports deliberate reflection, and aligns with target learning outcomes, such as improvement in diagnostic accuracy or a more balanced consideration of alternative hypotheses. Third, *safety and reliability* covers the frequency and severity of hallucinated content, fabricated citations, and inappropriate advice, together with calibration of uncertainty and appropriate refusals [28, 48].

Risk management follows from these criteria. Grounding and post hoc verification reduce the risk of misleading statements in feedback phases [37]. Schema validation and grammar-constrained decoding limit unintended content in analytics. Unknown intent detection reduces

the likelihood of unsafe actions when the model is uncertain [38]. Practical constraints include cost and latency in multi-turn simulations and privacy when real cases are discussed. These considerations shape the platform's balance of transparency, control, and responsiveness described in Chapter 4.

2.4 AI Virtual Patients as a Crucible for Clinical Reasoning

To translate the theoretical insights of dual-process reasoning and bias awareness into practice, medical education has increasingly turned to simulation. Among the available approaches, **Virtual Patients (VPs)**—interactive, computer-based scenarios—have become an important tool for practicing diagnostic reasoning in safe, standardized, and repeatable environments [7, 50]. Their pedagogical value lies in making reasoning both visible and improbable: students must externalize their thinking, and the system can provide structured feedback [54].

2.4.1 The Evolving Role of Simulation in Medical Education

Traditional VPs are designed to mimic clinical encounters, offering opportunities to practice history-taking, decision-making, and reasoning without risk to real patients [50]. More recently, **AI-powered Virtual Standardized Patients (VSPs)** extend this paradigm by enabling natural language dialogue and adaptive responses, enhancing the realism of the encounter [7, 36]. Conceptually, this aligns with *experiential learning theory* (Kolb), which emphasizes learning through cycles of action, feedback, and reflection [34]. By situating learners in an authentic yet controlled interaction, AI-VSPs provide a fertile context for bias-prone reasoning to surface and be addressed [51].

2.4.2 Designing the Virtual Encounter: The Principle of Progressive Disclosure

A critical design principle for effective simulation is **progressive disclosure**: clinical information is revealed incrementally in response to learner queries, mirroring real-world diagnostic practice [27]. This structure ensures that learners actively seek information rather than receiving it passively, creating opportunities for biases such as anchoring or premature closure to appear [14, 21]. When combined with meta-cognitive scaffolding, progressive disclosure transforms the simulation from a scripted case into a dynamic learning environment, where both strengths and vulnerabilities in reasoning can be observed [43].

From Concept to Evidence. The conceptual framework highlights the potential of AI-powered VPs, but their actual impact depends on empirical validation. Chapter 3 therefore reviews the existing evidence base, identifying strengths, limitations, and gaps in current research.

Chapter 3

Literature Review

Building on the conceptual foundations outlined in Chapter 2, this chapter critically examines the empirical evidence on AI-powered Virtual Patients (AI-VPs) in medical education. While Chapter 2 established the theoretical rationale for using simulation, progressive disclosure, and bias-aware design to strengthen clinical reasoning, the actual effectiveness of these approaches must be determined through research.

To this end, a scoping review was conducted to map the current state of knowledge on AI-VPs across four dimensions: (i) feasibility and validity of implementation, (ii) effectiveness for skill development, (iii) system design and integration, and (iv) impact on clinical reasoning and decision-making.

This review aims to answer two guiding questions:

- How can AI-powered virtual patients transform medical education?
- What key considerations are essential for their integration and evaluation to maximize educational impact?

By synthesizing recent empirical studies, the review identifies both the strengths and limitations of existing approaches and highlights the gaps that Smart-Doc seeks to address in its design and evaluation.

To achieve this overarching goal, this review maps existing literature on AI-powered virtual patients in medical education, specifically focusing on:

- **Feasibility and Validity:** Examining evidence supporting practical implementation and accuracy of AI-VPs in simulating patient encounters and assessing student performance.
- **Effectiveness for Skill Development:** Synthesizing findings on AI-VP effectiveness in enhancing specific clinical skills, including communication, history-taking, and clinical reasoning.
- **Development and Design Considerations:** Exploring key design principles, technological underpinnings, and iterative development approaches employed in AI-VP system creation.

- **Impact on Clinical Reasoning and Decision-Making:** Investigating AI-VP influence on cognitive skills underpinning clinical practice, particularly clinical reasoning and decision-making abilities.

3.1 Methods

To ensure adherence to the ***PRISMA-ScR***, this scoping review employed established guidelines and a corresponding checklist.

3.1.1 Eligibility Criteria

This review focused on peer-reviewed, open-access empirical studies published in English between 2019 and 2025, investigating the application of AI in medical education for history-taking skills. This timeframe was selected to capture recent AI and e-learning trends post-COVID-19.

Eligible studies included research that:

- Involved medical and/or nursing students. This criterion ensured the review focused on populations relevant to general medical and nursing education, excluding specialized fields.
- Specifically examined AI-driven interventions designed to enhance clinical skills.
- Featured interventions that:
  - Employed Artificial Intelligence (AI).
  - Virtual Patients (VPs) were used as the delivery method.
  - Focused on clinical practice. This narrowed the scope to studies directly related to the acquisition and development of practical clinical skills.

The review excluded book chapters, reports, conference papers, grey literature, pay-walled articles, pre-prints, unpublished materials, theoretical papers, reviews, editorials, and opinion pieces. Studies focusing on specific medical specialties, such as surgery or dental education were also excluded, as the focus of the review is the clinical interview process.

3.1.2 Information Sources

A comprehensive literature search was conducted across Cochrane Library, PubMed (including PubMed Central - PMC), Scopus, ScienceDirect, and IEEE Xplore electronic bibliographic databases.

3.1.3 Search Strategy

A systematic search was performed across multiple databases for relevant studies.

A core search strategy, initially developed for PubMed and subsequently adapted for other databases to maintain consistency, was employed. The PubMed search string, restricted by publication date (2019-2025) and language (English), was:

> ("Anamnestic interview" OR "History Taking" OR "Anamnesis") AND "Artificial Intelligence"

The search was implemented in PubMed (All Fields), PMC (Abstract), Scopus (Title, Abstract, Keywords), ScienceDirect (Title, Abstract, Keywords), IEEE Xplore (All Metadata), and the Cochrane Library (Title, Abstract, Keywords).

3.1.4 Selection of Sources of Evidence

A two-stage screening process was employed:

1. **Title and Abstract Screening:** Search results were imported into Zotero, duplicates removed, and uploaded to Rayyan, a web-based systematic review screening tool. Titles and abstracts were reviewed by a single researcher against predefined eligibility criteria.
2. **Full-Text Screening:** Potentially relevant articles underwent full-text assessment, also conducted by the same reviewer.

While dual screening is often employed to mitigate bias, single-reviewer screening was utilized here due to the context of this review being conducted for a master's dissertation

3.1.5 Data Charting Process

Data extraction, refined iteratively after pilot testing, was performed using a predefined form on a subset of included studies. Extracted data included:

- **Article Information:** Title, publication year, authors, keywords, publication type, country of origin.
- **Study Characteristics:** Objective, design, data collection / analysis methods, study enablers / challenges.
- **Participant Characteristics:** Student type, educational background, learning setting, age range, sample size, gender distribution (when available).
- **AI Intervention Details:** AI application type, AI techniques, tools, models, services employed, delivery modality, integration platform, curriculum integration, intervention duration and supporting and hindering factors.
- **Outcomes and Impact:** Measurement methods, specific history-taking skill outcomes, reported benefits, challenges.

Data charting was loosely structured around the PICO framework, acknowledging typical scoping review inconsistencies in comparative elements across studies.

3.2 Results

Database searches identified 508 publications. After removing 14.17% (72/508) duplicates, the titles and abstracts of 79.92% (406/508) publications were screened. Based on the eligibility criteria, full texts of 5.91% (30/508) publications were reviewed. Ultimately, 3.35% (17/508) studies were included in this scoping review (Figure1).

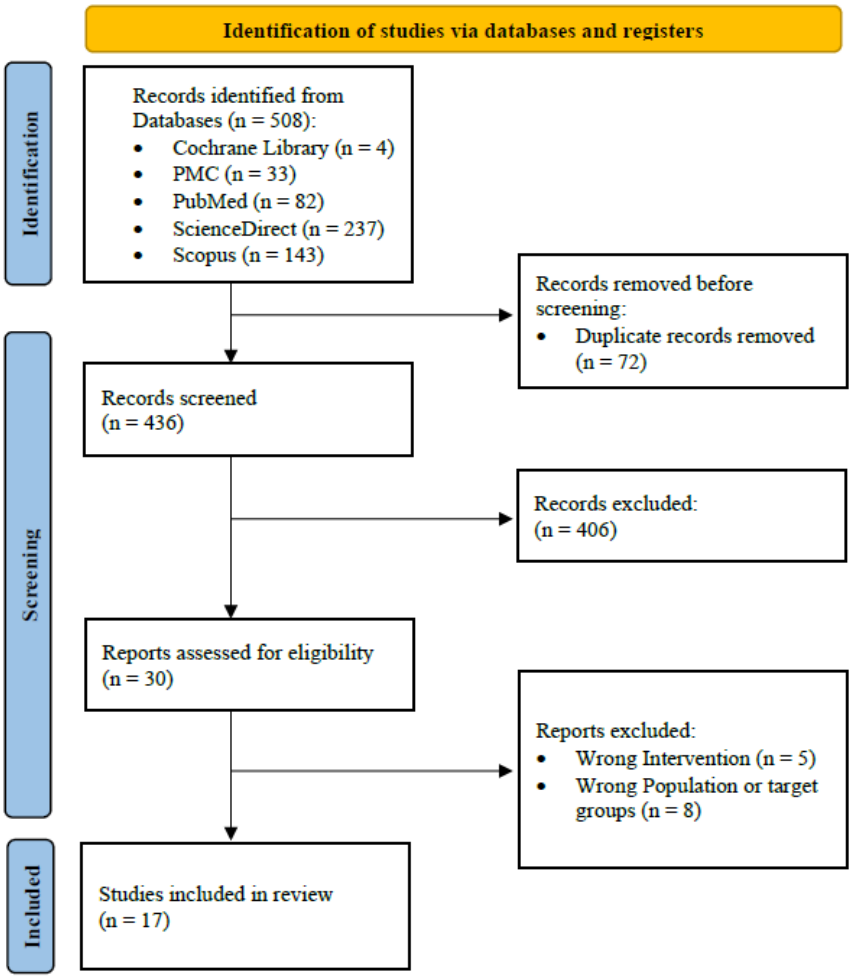

Figure 3.1: Prisma Flow Diagram

3.2.1 Publication Categorization

This scoping review categorized studies on AI-powered virtual patients (VPs) in medical education into aforementioned four key theme.

A Venn Diagram was developed to effectively visualizes the multidimensional nature of the reviewed literature by using a color gradient to represent scoring across four themes *(Figure. 2)*.

This allows for quick identification of research focus overlaps, and facilitates comparative analysis of thematic relevance across individual papers and clusters of studies.

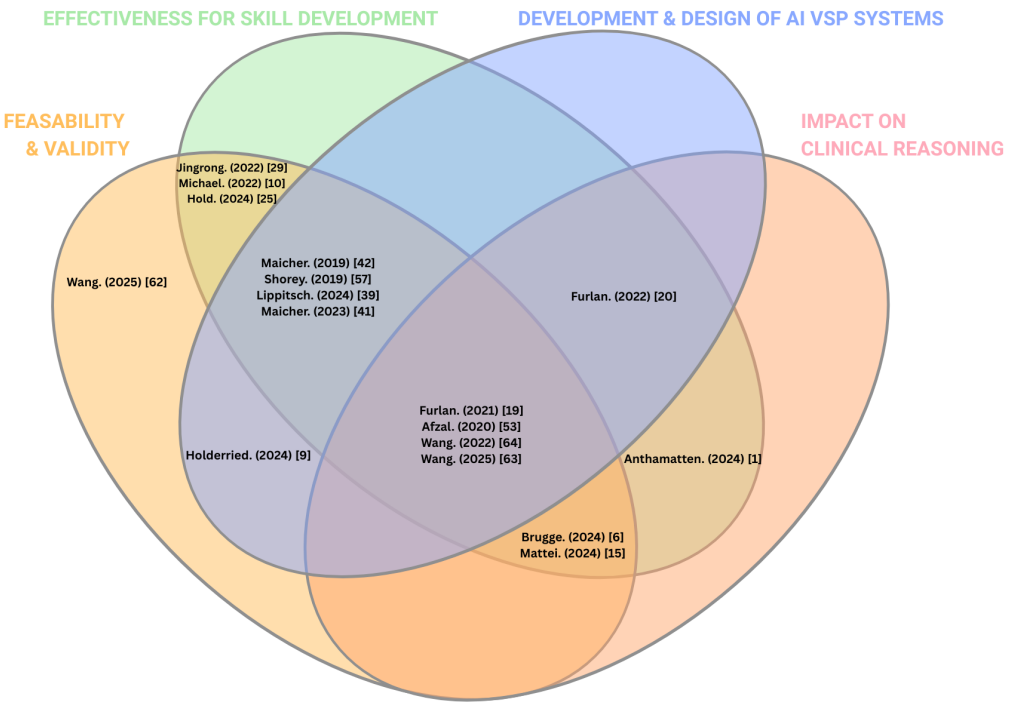

Figure 3.2: Venn Diagram Mapping

3.2.2 Study Design

To systematically classify the study designs, this review adopts the categorization proposed by *Hanna von Gerich et al. (2022)*[61], which groups research articles based on their primary aim. This framework includes developing new AI technologies, improving the accuracy or efficiency of existing AI technologies, testing different algorithms or AI systems, and assessing, evaluating, or validating existing AI technologies.

Hanna von Gerich et al. (2022) [61] utilized this categorization to analyze the progression of AI technology in nursing, revealing a predominance of studies in the early development phases. Applying this framework in the context of AI-powered virtual patients allows for a structured analysis of the research landscape, highlighting the primary aims of included studies and enabling comparisons within the field. Table II summarizes key study characteristics, including design and primary research aim, ensuring a structured and comparative analysis of the included studies.

Table 3.1: Study Design

Characteristic	**Author, (year of publication)**	**Publications (n=17), n(%)**
Study Design		
Experimental	*Wang. (2025)* [62], *Bruge. (2024)* [6], *Lippitsch. (2024)* [39], *Michael. (2022)* [10]	4 (23.5%)
Quasi-Experimental	*Shorey. (2019)* [57], *Anthamatten. (2024)* [1], *Wang. (2022)* [64]	3 (17.6%)
Descriptive	*Holderried. (2024)* [25], *Holderried. (2024)* [26], *Furlan. (2021)* [19], *Afzal. (2020)* [53], *Mattei. (2024)* [15], *Wang. (2025)* [63], *Maicher. (2023)* [41], *Jingrong. (2022)* [29], *Furlan. (2022)* [20], *Maicher. (2019)* [42]	10 (58.8%)
Aim of the research		
Development of new AI technologies:	*Furlan. (2021)* [19], *Shorey. (2019)* [57], *Afzal. (2020)* [53], *Maicher. (2023)* [41], *Wang. (2025)* [63], *Lippitisch. (2024)* [39], *Wang. (2022)* [64], *Furlan. (2022)* [20].	8 (47.1%)
Assessing, evaluating, or validating existing AI technologies:	*Holderried. (2024)* [25], *Holderried. (2024)* [26], *Wang. (2025)* [62], *Mattei. (2024)* [15]. *Jingrong. (2022)* [29], *Anthamatten. (2024)* [1], *Brugge. (2024)* [6], *Michael. (2022)* [10], *Maicher. (2019)* [42]	9 (53.9%)

3.2.3 Main findings

The subsequent section presents a concise overview of the key findings from the included publications, organized in table format for detailed discussion. Specifically, Table II encapsulates each study's characteristics, emphasizing feasibility, effectiveness, design considerations, and the impact of AI-driven Virtual Patients (VPs) on clinical training and decision-making. Complementing this, Table III identifies factors that either facilitate or challenge technology adoption, offering a balanced perspective on the potential benefits of VPs. Collectively, these tables deliver a comparative analysis that aids in understanding the current status, practical implications, and future research directions of AI-powered VPs.

Table 3.2: Condensed Summary of Studies on AI-Powered Virtual Patients

Author (Year)	Intervention	Population	Key Conclusions
		Feasibility and Validity	
Holderried (2024) [25]	GPT-3.5 chatbot	Med/Midwifery	Realistic; useful
Wang (2025) [62]	ChatGPT SP (prompts)	Medical	Effective; cost-efficient
Mattei (2024) [15]	AI-VSP cases	Healthcare	Better diagnostics; supplement
Maicher (2019) [42]	Web VSP	Medical	Reliable; feedback support
		Skill Development	
Shorey (2019) [57]	VP comm. training	Nursing	↑Self-efficacy; skills
Jingrong (2022) [29]	WeChat VSP	Nursing	Objective; influenced by skills
Michael (2022) [10]	Chatbot + tutor	Medical	Comparable to bedside
Holderried (2024) [26]	GPT-4 chatbot	Medical	>99% accurate; reliable
Lippitsch (2024) [39]	ViPATalk avatars	Medical	Equivalent to role-play
		Development and Design	
Furlan (2021) [19]	Hepius NLP/ITS	Medical	↑Test scores; reasoning
Wang (2025) [63]	XueYiKu app	Students, doctors	Active learning; feedback
Maicher (2023) [41]	AI VSP (ASR, RNN)	Medical	90% accuracy; human-like
Wang (2022) [64]	AIteach	Medical	↑Clinical thinking
Afzal (2020) [53]	Dengue tutor bot	Medical	High usability; engagement
		Clinical Reasoning	
Furlan (2022) [20]	VP NLP/ITS cases	Nurse practitioners	Consistent metrics; weak exam link
Brugge (2024) [6]	ChatGPT feedback	Medical	Better decision-making
Anthamatten (2024) [1]	VR + AI sims	Medical	Effective; screen > VR

Table 3.3: Supporting and Hindering Factors for AI Adoption in Medical Education

Factor Category	Supporting Factors	Hindering Factors
Educational Potential	• Innovative, accessible learning [15, 25] • Overcomes traditional training limitations [25] • Cost-effective [6, 25] • High-quality feedback [26] • Realistic simulations [26] • Enhances learning through practice [26]	• Concerns about accuracy [6, 19, 42, 53] • Limitations in subjective judgment and feedback mechanisms [39, 62] • Reliance on AI instead of learning [26] • Unexpected AI behaviors [26]
Technology & Implementation	• Technological advances in AI and NLP [19, 41, 64] • Improved system accuracy with training data [41] • Consistent answers [10] • Distance learning capabilities [10, 19, 64]	• Technical issues [1, 15, 57] • High implementation costs [15] • Reliance on third-party services [39] • Limited case variety [63] • Lack of system integration [63] • Difficulty in communication with AI [1]
Realism & Human Interaction	• Simulates standardized patients effectively [62] • Bridges theoretical and clinical practice [15]	• Lack of realism in avatars [15] • Inability to fully replace human interaction [15] • Absence of non-verbal cues [6]
Cost & Accessibility	• Cost reduction [42] • Increased accessibility [6] • Resource sharing [63]	• Resource-intensive programs [41]
Other	• Need for safe practice environment [15] • Potential for AI improvement [15] • Positive student feedback [41]	• Negative student attitudes towards AI feedback [26] • Potential disengagement [53] • Preference for screen-based format over immersive VR [1]

Discussion

This section begins with a bullet-point summary of the key findings across the four identified themes, followed by a detailed discussion of these findings.

Key Findings

- **Feasibility and Validity:** AI-powered virtual patients (VSPs) are feasible and valid for medical education, showing high accuracy and realism [15, 25, 42, 62]. However, continuous refinement is needed to optimize accuracy and address context-dependent variability in realism.
- **Effectiveness for Skill Development:** AI-VPs effectively enhance foundational skills, particularly history-taking, offering scalable, feedback-rich learning comparable to traditional methods [10, 26, 29, 39, 57]. Further research is needed to assess their impact on advanced communication skills, empathy, and long-term skill retention.
- **Development and Design:** AI-VP systems leverage advanced technologies like NLP and ITS for realism, scalability, and pedagogical soundness [19, 41, 63, 64]. The use of real medical records and hybrid AI designs further enhances authenticity and sophistication.
- **Impact on Clinical Reasoning and Decision-Making:** AI-VPs show promise in enhancing clinical reasoning and decision-making (CDM) skills, with AI feedback significantly improving CDM scores [1, 6, 20]. However, further research is needed to fully quantify their long-term impact on cognitive skills.
- **Balance between AI Innovation and Evaluation:** The field demonstrates a balance between rapid AI innovation and rigorous evaluation, with a growing body of experimental research validating AI-VPs' educational effectiveness [6, 26].
- **Implications for Clinical Training:** AI-VPs have transformative potential for medical education, offering scalable and effective training for foundational skills, personalized feedback, and data-driven insights [10, 20]. However, strategic integration and ongoing evaluation are crucial to address limitations and maximize their pedagogical impact.

Feasibility and Validity of AI-Powered Virtual Patients

AI-powered virtual patients (VSPs) demonstrate strong feasibility and validity, although optimizing accuracy and realism remains a challenge. *Holderried et al. (2024)* established GPT-3.5's feasibility for patient simulation, with 94.4% answer alignment to illness scripts and 97.9% plausibility ratings, indicating educational acceptability and good user experience (CUQ *(tool measures UX and usability of chatbots)* score: 77/100) [25]. However, rare implausible answers (2.1%) highlighted the need for refined model constraints to improve output accuracy and reduce socially desirable responses [25].

Wang et al. (2025) reinforced AI-VP validity, showing prompt engineering significantly enhanced ChatGPT's clinical accuracy. Iterative prompt design decreased scoring discrepancies from 29.83% to 6.06%, achieving 89.3% question categorization accuracy, emphasizing iterative prompt design for valid assessments [62]. Similarly, *Maicher et al. (2019)* validated AI assessment of history-taking skills, showing computer-generated scores (87% accuracy) closely mirrored human rater scores (90%), with high inter-rater reliability (Intraclass correlation coefficient (ICC) 0.75 for 81% of categories) confirming system consistency, although complex domains require further refinement [42].

De Mattei et al. (2024) further evidenced AI-VP feasibility, with 50–87% of students across disciplines rating them as realistic. Supporting their utility, 72–90% reported improved diagnostic confidence. However, variable realism ratings (e.g., 16% among physician assistant students) indicated context-dependent feasibility [15].

Effectiveness of AI-Powered VPs for Skill Development

AI-powered VPs show promise in effectively enhancing specific skills, particularly foundational skills, although nuances exist. *Holderried et al. (2024)* directly addressed AI-VP effectiveness, showing GPT-4 provided structured feedback on history-taking, aligning with human raters and supporting skill improvement. *Lippitsch et al. (2024)* further indicated ViPATalk's effectiveness, demonstrating equivalence to role-play for history-taking completeness while offering automated feedback and self-study benefits. This suggests AI-VPs offer a scalable alternative for foundational skill practice [26, 39].

Shorey et al. (2019) highlighted AI-VPs' potential to enhance broader communication skills, noting improved student self-efficacy and engagement, though primarily based on perceived improvements. *Du et al. (2022)* indirectly supported effectiveness by validating AI-VPs' ability to provide objective history-taking skill assessments, crucial for targeted feedback. Finally, *Co et al. (2022)*'s case-control study directly compared AI-VP training to traditional bedside teaching, finding comparable student performance for history-taking, suggesting AI-VPs can be as effective as real-patient interactions for foundational skills training [10, 29, 57].

However, while AI-VPs are effective for foundational skills and feedback delivery, studies indicate limitations. *Lippitsch et al. (2024)* found equivalence, not superiority, to role-play [39]. *Shorey et al. (2019)* focused on perceived self-efficacy [57], and *Du et al. (2022)* primarily on assessment validity [29]. Further research is needed to assess effectiveness for advanced communication skills, empathy, and long-term skill retention. Nevertheless, evidence indicates AI-VPs are effective tools for foundational skill development in medical education, offering scalable, accessible, feedback-rich learning.

Development and Design of AI-Powered Virtual Patient Systems

AI-VP system development consistently leverages advanced technologies for realism, scalability, and pedagogical soundness. *Furlan et al. (2021)* emphasized NLP integration in Hepius for

realistic interaction and ITS for personalized feedback. *Maicher et al. (2023)* detailed a hybrid AI system combining rule-based and neural network NLU for improved conversational fidelity. *Wang et al. (2022 & 2025)* highlighted NLP's crucial role in processing real medical records for authentic case content in AIteach and XueYiKu [19, 41, 63, 64].

Beyond NLP, Intelligent Tutoring Systems (ITS) are increasingly integrated for pedagogical enhancement. *Furlan et al. (2021)* incorporated an ITS module in *Hepius* for step-by-step feedback. *Wang et al. (2025)* implemented multidimensional automatic evaluation in XueYiKu, using radar charts for visualized feedback. The use of real medical records, as seen in AIteach and XueYiKu [63, 64], and hybrid AI designs [41] further underscores the trend towards sophisticated, data-driven, and pedagogically informed AI-VP systems. Mobile accessibility and scalable architectures are also key design considerations [53, 63].

Impact of AI-Powered Virtual Patients on Clinical Reasoning and Decision-Making

AI-powered VPs demonstrate a promising, nuanced impact on cognitive skills, particularly clinical reasoning and decision-making. *Brügge et al. (2024)* provides direct evidence of enhanced CDM with AI feedback. Their RCT showed significant CDM score improvements in students receiving AI feedback, especially in "creating context" and "securing information" subdomains [6]. *Furlan et al. (2022)* further developed learning analytics metrics to assess specific CDM components within AI-VP simulations, offering granular performance insights [20].

Anthamatten & Holt (2024)'s qualitative findings align, with NP students reporting VR-AI simulations as most impactful on decision-making abilities and responsiveness to clinical changes. Students also reported increased diagnostic confidence and perceived AI-VPs as useful for history-taking and differential diagnosis development [1].

However, limitations exist. *Furlan et al. (2022)* found limited correlation with external hematology exam scores, suggesting VPs and traditional exams assess different competencies [20]. *Anthamatten & Holt (2024)* primarily relied on student perceptions, requiring further objective validation. Despite these nuances, evidence suggests AI-VPs can positively impact cognitive skills, particularly CDM, though further research is needed to fully quantify long-term impact. [1]

Balance between AI innovation and evaluation

Reflecting on methodological approaches employed across reviewed studies, a notable trend emerges concerning balance between AI innovation and rigorous evaluation. A significant portion of literature, particularly within 'Development and Design' theme, is descriptive, detailing innovative architectures, functionalities, technological advancements of AI-VP systems, showcasing rapid AI innovation in medical education. However, alongside descriptive accounts, a growing body of experimental research, as seen in 'Effectiveness' and 'Impact on Clinical Reasoning' themes, is dedicated to rigorously evaluating AI-VP pedagogical value and impact on student learning and cognitive skill development. This increasing experimental validation emphasis showcases a maturing field, moving beyond demonstrating technological feasibility towards establishing evidence

for AI-powered virtual patients' educational effectiveness and optimal implementation in medical training.

Interpretation in Context

The findings of this scoping review resonate with and expand upon existing reviews in AI-powered medical education. Similar to *Stamer et al. (2023)'s* [58] scoping review focused on communication skills training, it confirms the promising potential of AI and machine learning to enhance medical education, particularly by offering cost-effective, readily accessible training modalities. Both reviews align in identifying AI-VP feasibility and acceptability as key strengths, echoing *Hindelang et al. (2024)'s* [24] finding that chatbots can increase patient engagement and streamline data collection. This consensus across reviews strengthens the argument for the growing maturity and acceptance of AI-VP technology in medical education.

However, the current study also adds nuanced perspectives and extends beyond prior review scope. While *Stamer et al. (2023)'s* scoping review focused on communication skills training, *Hindelang et al. (2024)* and this study acknowledge technological limitations and challenges related to AI-VP interaction authenticity and natural language flow. Unlike those reviews, it delves deeper into the specific technological advancements (NLP, ITS, hybrid systems, learning analytics) employed to address these limitations, particularly within the "Development and Design" theme. Furthermore, unlike the narrower focus of *Stamer et al. (2023)* on communication skills and *Hindelang et al. (2024)* on history-taking chatbots, its broader scope, encompassing the impact on clinical reasoning and decision-making, provides a more holistic, integrated understanding of AI-VP pedagogical potential across a wider spectrum of cognitive skills essential for clinical practice. Notably, the consistent emphasis in this review on AI-generated feedback's crucial role in enhancing effectiveness, as highlighted in *Holderried et al. (2024)* [26] and *Brügge et al. (2024)* [6], represents a key area of focus not explicitly emphasized in other scoping reviews. This highlights AI-generated feedback as a potentially critical mechanism for maximizing AI-VP pedagogical impact, as the reviewed studies demonstrate that detailed feedback provides a higher level of learning.

Implications for Clinical Training

Building upon these findings and considering identified trends, key implications for clinical training and medical education emerge. The demonstrated feasibility, validity, and effectiveness of AI-powered VPs suggest transformative potential for curriculum design and technology adoption.

Firstly, findings strongly advocate for judicious AI-VP simulation integration into medical curricula, particularly in early clinical training stages, offering a scalable, accessible means for repeated practice in foundational skills. AI-VP training can be comparable to real-patient bedside teaching for history-taking skills, suggesting a viable, ethical alternative for novice learners [10]. For example, AI-VPs can be strategically integrated into early clinical training to provide students

with repeated practice in history-taking, allowing them to develop proficiency before encountering real patients.

Secondly, emphasis on AI-generated feedback and learning analytics points towards a future of more personalized, data-driven medical education. AI-VPs' ability to provide structured, automated, potentially individualized feedback offers a mechanism to enhance learning efficiency and address individual student needs. Furthermore, learning analytics metrics enable continuous program improvement and curriculum refinement based on data-driven insights [20]. This data can be used to identify areas where students struggle and to tailor instruction to meet their specific needs.

An interesting aspect of the clinical reasoning process is the presence of cognitive errors and biases. Furlan et al. (2021) highlight that the majority of diagnostic errors are caused by cognitive errors, rather than knowledge deficiency [19]. Subsequently, Furlan et al. (2022) found that many students skipped crucial steps of the clinical interview, falling into the "early closure" mistake [20]. This suggests that students may be susceptible to cognitive biases during the clinical reasoning process, which is an important consideration for future research and AI-VP development. AI-VPs can be designed to simulate scenarios that are likely to trigger cognitive biases, providing students with the opportunity to recognize and correct these errors in a safe environment.

However, reviewed literature also underscores the importance of thoughtful, strategic implementation. As highlighted by Stamer et al. (2023) [58] and Hindelang et al. (2024) [24], challenges related to authenticity, natural language flow, technological limitations, and ongoing refinement needs remain. Future curriculum design and technology adoption should prioritize:

- **Enhancing Realism and Authenticity:** Focus on improving AI-VP interaction realism and authenticity, particularly in nuanced communication aspects beyond factual accuracy.
- **Strategic Integration of AI-VPs:** Integrate AI-VPs as a supplement, not a replacement, for traditional clinical experiences. For instance, AI-VPs can be used for initial skill development and practice, while real-world clinical experiences remain essential for developing advanced clinical judgment, empathy, and complex decision-making skills that require human-to-human interaction.
- **Stronger Evaluation Methodologies:** Focus on robust evaluation methodologies, including RCTs and objective skill measures, to further validate AI-VP interventions' long-term educational impact. For example, future research could employ longitudinal studies to assess the long-term impact of AI-VP training on clinical skills and patient outcomes, or use standardized patient assessments to objectively measure the improvement in students' clinical skills after AI-VP training.
- **Other Considerations:** Carefully consider ethical implications, data privacy, and user-centered design for responsible, equitable AI-VP implementation.

Strengths and Limitations

This scoping review benefits from a solid search strategy across major databases and a detailed thematic analysis, providing a strong synthesis of current literature. However, it also has methodological limitations. Grey literature exclusion may have overlooked relevant findings. Initial single reviewer screening introduces potential bias, and the focus on peer-reviewed, open-access publications carries a risk of publication bias towards positive results. Future research should consider incorporating grey literature, implementing dual screening, and acknowledging potential biases when interpreting findings.

Conclusion

This scoping review examined the current landscape of AI-powered virtual patients in medical education, synthesizing evidence across feasibility and validity, effectiveness for skill development, system design and impact on clinical reasoning.

The findings demonstrate the promising potential of AI-VPs as a valuable and increasingly sophisticated tool for medical training. AI-VPs offer a feasible, scalable, and engaging modality for delivering standardized, repeatable practice opportunities, particularly for foundational communication and history-taking skills. Moreover, AI-driven feedback and learning analytics integration within these systems offers pathways towards personalized, data-driven medical education with the potential to enhance clinical reasoning and decision-making abilities.

While challenges such as achieving complete realism, overcoming technological limitations and the need for further rigorous validation remain, the AI-VP development trajectory is clearly towards more authentic, user-friendly and pedagogically effective tools. To fully leverage this potential, medical educators should strategically integrate AI-VPs into curricula, focusing on supplementing traditional methods rather than replacing them. Researchers should prioritize investigating AI-VP impact on advanced clinical skills and long-term outcomes, specifically through longitudinal studies assessing skill retention beyond immediate training. Evaluating AI-VPs' sustained impact on clinical competence in this manner would provide stronger evidence for their educational value and inform best practices for curriculum integration. Furthermore, developers should continue to focus on improving AI-VP realism and addressing current technological limitations. Furthermore, developers should focus on improving AI-VP realism and addressing current technological limitations. These efforts will maximize the unique strengths of AI-VPs, holding significant promise for transforming medical education and cultivating a more competent and future-ready healthcare workforce. It is crucial to invest in this rapidly evolving field and ensure AI-VPs' responsible and effective implementation in medical education and beyond.
Taken together, the evidence shows both the promise and the current limitations of AI-powered virtual patients. While feasibility and usability studies highlight their potential, questions remain regarding their role in teaching complex reasoning skills, supporting metacognition, and addressing cognitive biases in authentic contexts.

To respond to these gaps, the next chapter presents the design and development of the **SmartDoc** platform. Chapter 4 demonstrates how the theoretical principles outlined in Chapter 2 and the empirical findings reviewed here informed the system's architecture, case design, and educational features.

Chapter 4

Developing Smart-Doc: A Bias-Aware Clinical Simulation Platform

This chapter presents the methods developed to address the research questions stated in Chapter 1.2. To operationalize these questions, a decision was taken to develop **Smart-Doc**, a conversational clinical simulation platform. The rationale is threefold. First, the research requires an environment where diagnostic reasoning is observable at the level of turns, intents, and evidence use, so that bias-related behaviors can be detected and measured. Second, the study needs a controllable setting that preserves the inquiry process. Information should be disclosed progressively in response to learner actions, rather than supplied upfront, in order to reveal how hypotheses are formed and revised. Third, the platform must support systematic feedback and analytics that map interaction patterns to learning outcomes while remaining feasible for curricular deployment.

Virtual patients are established teaching tools, yet most implementations emphasize factual content or procedural competence. Few systems target the cognitive processes that lead to diagnostic error, and those that do often rely on post hoc analysis without interactive support. Smart-Doc addresses this gap by combining structured dialogue control with large language models to sustain natural conversation while maintaining educational oversight. The design centers on intent recognition for dialogue orchestration, progressive disclosure to prevent premature closure, and a bias analysis pipeline that treats bias as an observable pattern in interaction.

Alternative approaches were considered and rejected. Fully neural, end-to-end dialogue without explicit control was deemed inadequate for educational purposes because it limits transparency, reduces the ability to enforce phase boundaries, and complicates measurement and evaluation. Conversely, entirely scripted dialogues lack realism and do not elicit authentic information seeking. Smart-Doc adopts a hybrid strategy that joins rule-based pedagogical logic with model-generated language. This provides realism for the learner and accountability for the instructor.

The remainder of the chapter is organized as follows. Section 4.1 introduces the conceptual and pedagogical design, including the progression of phases and the operationalization of bias markers in support of RQ 1 and RQ 2. Section 4.2 details the technical implementation, covering

algorithms, schemas, and deployment choices that enable scalable use and address RQ 3. Section 4.3 illustrates a complete diagnostic workflow using the study case. Section 4.4 describes the learner and administrator interfaces. The chapter concludes with a summary in Section 4.5 and a bridge to the empirical evaluation in Chapter 5.

4.1 System Design

Smart-Doc is designed to make diagnostic reasoning observable while preserving the authenticity of a clinical encounter. The design translates three pedagogical commitments into concrete mechanisms. First, **authenticity:** the interaction should resemble a real consultation, with natural language anamnesis, focused examination, and selective access to investigations. Second, **bias-awareness:** the system should detect patterns that signal anchoring, confirmation, or premature closure and respond with prompts that encourage reflection without breaking immersion. Third, **structured reflection:** learners should periodically articulate evidence, alternatives, and uncertainty so that reasoning steps become explicit. These commitments, introduced in Chapters 2 and 3, motivate a platform that couples conversational flexibility with deliberate control in service of RQ 1–3.

4.1.1 Architecture and core components

The platform follows a three-layer architecture that separates pedagogy from orchestration and presentation. The **domain logic** layer encodes rules for information disclosure and bias detection so that educational intent governs what the learner can discover and when. The **API layer** coordinates modules, maintains the case state, and enforces phase boundaries across the interview, examination, and investigations. The **presentation layer** provides the browser interface for dialogue, results, and feedback. At the center is the **Intent Driven Disclosure Manager**, which keeps the learning loop coherent by linking what the learner asks, what the case reveals, and when reflective prompts appear.

Concretely, the manager orchestrates three processes that run on every turn: *intent classification* maps the learner's message to a structured action; *progressive disclosure* reveals only the information justified by that action and the current case state; and *bias monitoring* looks for patterns that warrant a brief, context aware nudge toward alternative hypotheses.

Smart-Doc integrates the following modules into a closed pedagogical loop.

Intent Classifier. Learner queries are mapped to a taxonomy of 33 diagnostic intents spanning anamnesis, examination, and investigations (Table 4.1). Classification is contextual, so identical strings can be interpreted differently by phase. This supports intuitive conversation while maintaining structural coherence.

Discovery Processor. Clinical facts are represented as modular *information blocks* annotated with level, prerequisites, and pedagogical importance. The processor releases blocks only when

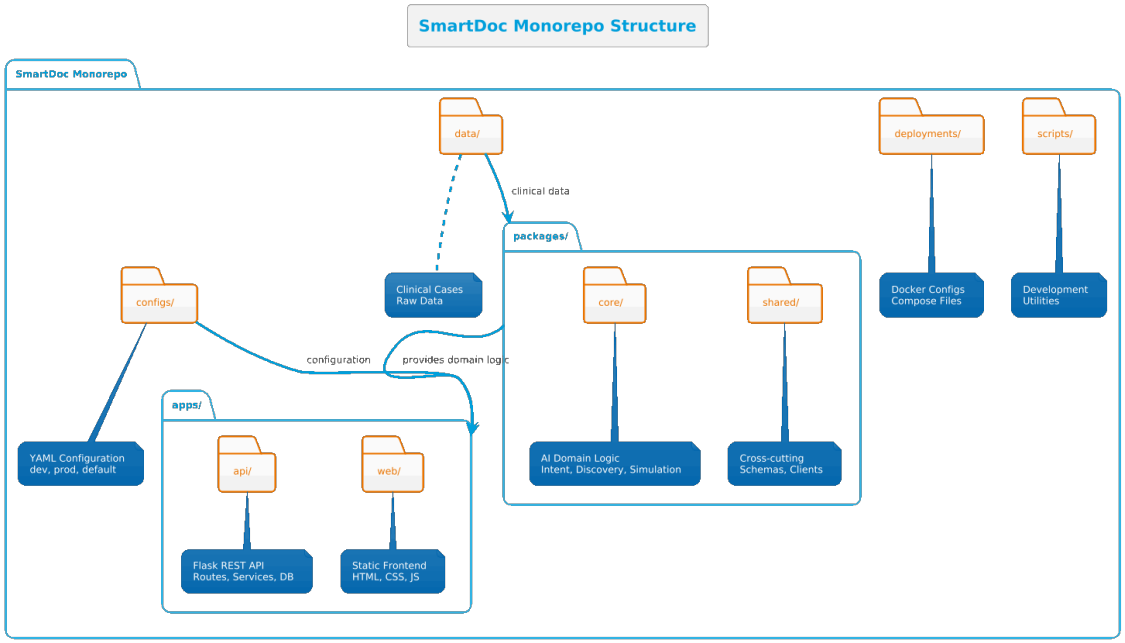

Figure 4.1: High-level architecture with three layers and modular components.

the learner's intent and case state justify it, which preserves inquiry and prevents answer reveals that would short-circuit reasoning.

Bias Analyzer. A hybrid of rule-based heuristics and model-assisted semantic checks monitors for selective information seeking aligned with a single hypothesis, early commitment before discriminators are queried, or disregard of counter-evidence. When such patterns emerge, the system issues a brief prompt that invites reconsideration, for example *"What else could explain these findings?"*.

Clinical Evaluator. A single model under strict temperature control scores information gathering, diagnostic accuracy, and indicators of bias awareness against structured rubrics designed for reproducibility. Scores and narrative feedback are presented at planned checkpoints.

Simulation Engine and Responders. The Simulation Engine coordinates the cycle of query, classification, disclosure, bias analysis, and response generation. Responder templates adapt tone and format to phase, conversational in anamnesis, objective in examination, and concise for laboratory or imaging results.

Session Management. All interactions are logged with timestamps to create a reasoning trace used for feedback, assessment, and research.

Table 4.1: Intent categories by diagnostic phase (selected examples).

Phase	Category	Example intents	Count
Anamnesis	Past Medical History	`pmh_general`, `pmh_family_history`, `pmh_surgical_history`	3
Anamnesis	History of Present Illness	`hpi_chief_complaint`, `hpi_onset_duration`, `hpi_fever`, `hpi_chills`	9
Anamnesis	Medications	`meds_current`, `meds_ra_specific`, `meds_full_reconciliation`	3
Anamnesis	Social History	`social_smoking`, `social_alcohol`, `social_occupation`	3
Exam	General / System	`exam_vital`, `exam_cardiovascular`, `exam_respiratory`	4
Investigations	Laboratory	`labs_general`, `labs_cbc`, `labs_bnp`	5
Investigations	Imaging	`imaging_chest_xray`, `imaging_ct_chest`, `imaging_echo`	6

4.1.2 Case Modeling and Bias Integration

Each Smart-Doc case embeds pedagogical logic directly in its structure, linking bias triggers, learning objectives, and reflection prompts.

Information blocks. Every block defines its level of inquiry, prerequisites, criticality, and trigger intents. This schema supports both shallow and deep questioning and mirrors expert diagnostic reasoning.

Bias triggers. Metadata identifies potential anchors, contradictory evidence, and corrective prompts. When a learner focuses excessively on one hypothesis despite conflicting data, a brief bias card invites reconsideration.

Educational scaffolding. Repeated unproductive queries activate dynamic hints that steer the interview without providing answers. This adaptive feedback maintains engagement while avoiding frustration.

Reflection prompts. At case closure, learners answer structured meta-cognitive questions on evidence, counter-evidence, and alternatives. These responses foster deliberate reflection and make judgment processes explicit.

Prototype case. The pilot case adapts Mull et al. [46], where an elderly woman with dyspnea was misdiagnosed with heart failure due to anchoring. Smart-Doc reproduces this trajectory

through controlled framing, contradictory findings, and the critical discovery of immunosuppressant therapy.

Taken together, these elements convert each case from a scripted storyline into a declarative, testable specification that binds content to pedagogy. The same schema supports rapid authoring of new scenarios, controlled variation for A/B comparisons, and precise logging of high-value events for analytics. Bias annotations align with RQ 1, reflection prompts operationalize RQ 2, and the case agnostic structure contributes to scalability in RQ 3. This integration ensures that learning signals are generated by inquiry rather than answer reveal, and it provides a reproducible basis for the evaluation reported in Chapter 5.

4.2 Technical Implementation

4.2.1 Runtime platform and deployment

Smart-Doc runs on a modular, containerized platform designed for reproducibility, privacy, and ease of adoption in research and teaching. The backend exposes a simple API, the model host runs locally for privacy, and the front-end is lightweight for easy deployment. Table 4.2 lists the core technologies.

Table 4.2: Core technologies and versions.

Component	Technology	Version
Backend framework	Flask	3.0+
Language	Python	3.13+
Dependency mgmt.	Poetry	1.8+
ORM and migrations	SQLAlchemy + Alembic	2.0+
Data validation	Pydantic	2.0+
LLM interface	Ollama	Latest
LLM model	Gemma 3:4b-it-q4_K_M	4B params
Frontend	HTML/CSS/JS	ES2020+
Containerisation	Docker + Compose	24.0+
Prod server	Gunicorn	21.0+

Two deployment modes are supported. A local mode prioritizes privacy and is suited to small cohorts or lab studies. A cloud mode supports larger classes and centralized monitoring. Structured logs record latency and fallback events, and per-turn traces enable audit and research without storing raw audio or video.

Observability. Structured logs capture per-turn latency, token counts, classification confidence, validation outcomes, and fallback reasons. Metrics are aggregated by session and module to monitor performance and safety. In the study deployment, median response time was approximately 6 seconds per turn, so a request timeout threshold is enforced with a single retry using a compacted prompt. Timeouts and retries are recorded for audit, and error messages are downgraded to clarifications to preserve conversational flow.

4.2.2 Interaction pipeline and control algorithms

Each turn follows the same loop: the learner submits a query, the system classifies intent with phase awareness, eligible information blocks are disclosed, a phase-appropriate responder generates a reply, the bias monitor evaluates recent behavior, and events are logged to the session state. Ambiguous queries trigger clarification rather than errors to preserve conversational flow.

Intent-driven progressive disclosure. The procedure below governs information revelation, scaffolding, and bias checks during each interaction.

```
Input: user_query, session_context, revealed_blocks
Output: information_blocks, educational_hints, bias_warnings

phase ← session_context.current_phase
available ← filter_intents_by_phase(phase)
cls ← LLM_classify(user_query, available)
if cls.confidence < 0.30 then
    return request_clarification(user_query, available)
end if
intent ← cls.intent_id
mapped ← get_intent_block_mappings(intent, case)
eligible ← filter_by_prerequisites(mapped, revealed_blocks)

// Just-in-time scaffolding
if intent == "meds_ra_specific_initial_query" then
    n ← count_previous_queries(intent, session)
    if n > 1 AND not revealed("critical_infliximab") then
        hint ← "Maybe you could check her previous
            hospital records?"
        return (eligible, hint, None)
    end if
end if

// Real-time bias check
focus ← calculate_focus(session_context.working_diagnosis)
if focus > 0.70 then
    contra ← detect_contradictions(revealed_blocks,
         session_context.working_diagnosis)
    if not empty(contra) then
        warn ← make_bias_warning("anchoring",
         "What else could explain these findings?")
        return (eligible, None, warn)
    end if
end if

return (eligible, None, None)
```

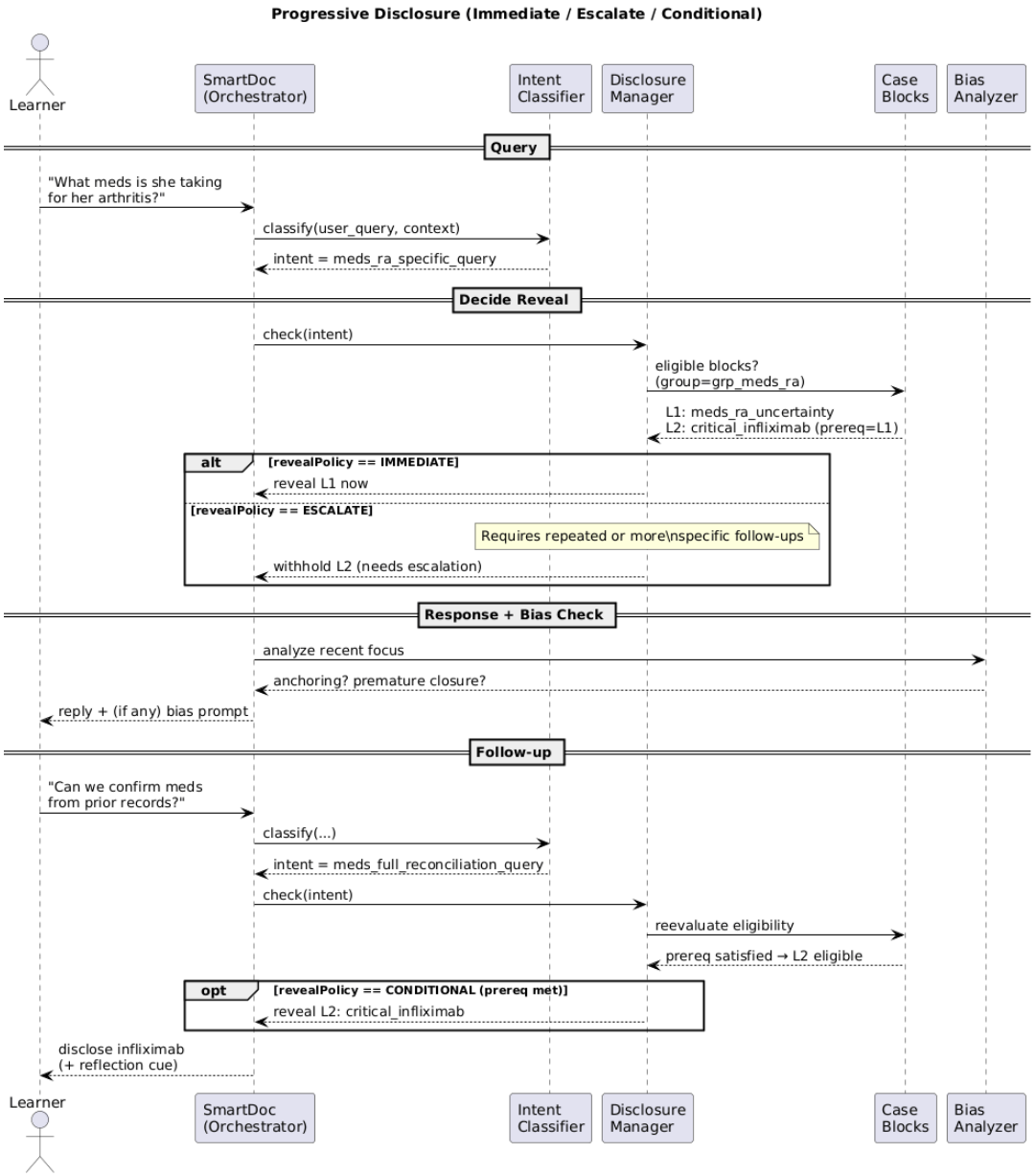

Figure 4.2: Progressive disclosure sequence: from query to information revelation with prerequisite checking.

Bias detection and warnings. Bias is treated as a pattern in interaction, not a trait. The monitor estimates focus on a single hypothesis, checks for contradictions that have been revealed, and looks for selective evidence seeking or early closure.

```
Input: session_history, current_hypothesis, revealed_information
Output: bias_warning or None

1:  recent ← last_n(session_history.queries, 5)
2:  f_ratio ← focus_ratio(recent, current_hypothesis)
```

```
3:  // Anchoring
4:  if f_ratio > 0.70 then
5:      contra ← detect_contradictions(revealed_information,
            current_hypothesis)
6:      if not empty(contra) then
7:          log_bias("anchoring", evidence=contra)
8:          return warn("anchoring",
              "You seem focused on one diagnosis.
              What else could explain this?",
              "moderate")
9:      end if
10: end if

11: // Confirmation bias
12: s ← count_supporting(recent, current_hypothesis)
13: r ← count_refuting(recent, current_hypothesis)
14: if s > 0 and r == 0 and len(recent) >= 4 then
15:     log_bias("confirmation", pattern="seeking only
            supporting evidence")
16:     return warn("confirmation",
                    "Consider evidence that might contradict
                         your hypothesis", "low")
17: end if

18: // Premature closure
19: needed ← critical_blocks(case)
20: have ← intersect(needed, revealed_information)
21: if len(have) < 0.6 * len(needed) and
    diagnosis_submitted(session_history) then
22:     log_bias("premature_closure", coverage=len(have))
23:     return warn("premature_closure", None, "high")
24: end if

25: return None
```

Classification refinements and accuracy. Iterative testing exposed two failure modes. Queries about past medical history were misclassified as medication requests because negative examples were sparse. Enriching intent definitions with explicit negatives improved that contrast from 57% to 95% and raised overall accuracy. Cross-phase acceptance errors, such as medication queries during examination, were removed through strict phase gating. The result was a stable 96% overall accuracy (Table 4.3).

Table 4.3: Intent classification accuracy improvements through iterative refinement.

Configuration	PMH vs. Meds	Cross-phase errors	Overall
Baseline (generic defs.)	57%	23%	78%
+ Enhanced definitions	95%	18%	87%
+ Strict context filtering	95%	0%	96%

4.2.3 Model configuration, validation, and data layer

Smart-Doc operates on a local Ollama host with the quantized `gemma3:4b-it-q4_K_M` model for classification, dialogue, and evaluation. Module-specific temperatures balance determinism and naturalness (Table 4.4). All structured outputs pass Pydantic validation. If validation fails, a retry with a compacted prompt is issued, and a conservative fallback response keeps the session moving.

Table 4.4: LLM temperature settings by module.

Module	Temp.	Rationale
Intent classification	0.2	High consistency with minimal variability
Clinical evaluation	0.3	Reproducible scoring across sessions
Responders (dialogue)	0.5	Natural patient dialogue with limited variability
Responders (labs)	0.3	Deterministic, professional reporting

Persistent data is stored in SQLite with an event-driven schema. Entities include *sessions*, *messages*, *discoveries*, *bias_events*, and *evaluations*. In-memory state supports real-time performance, while write-ahead logging preserves complete traces.

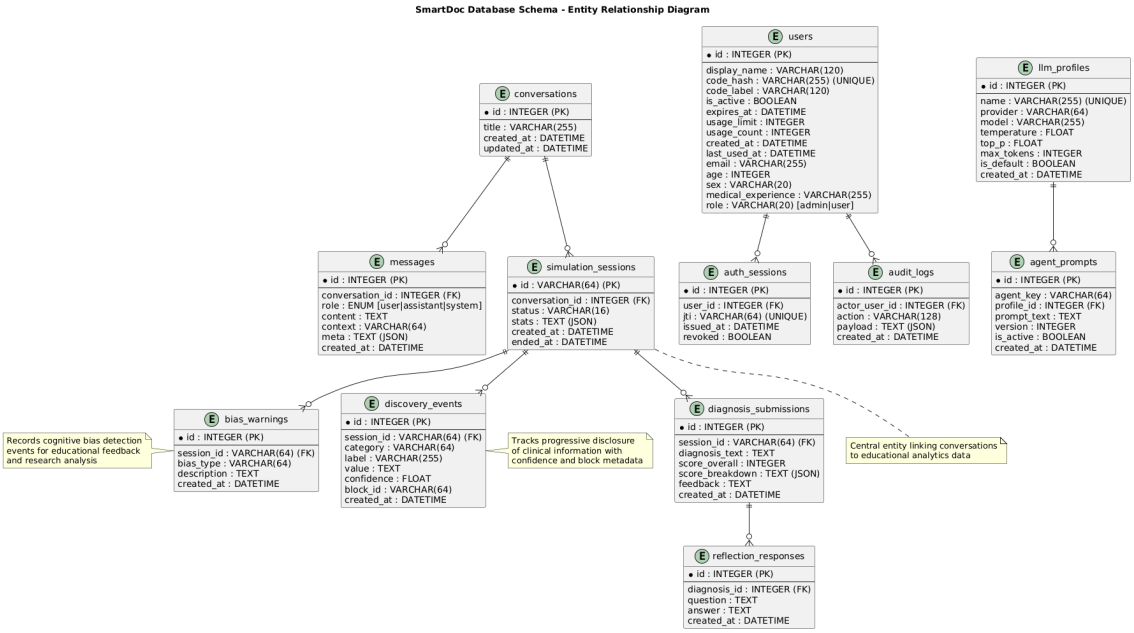

Figure 4.3: Conceptual schema capturing reasoning traces, bias events, and reflection data.

4.3 Example Workflow: The Miliary Tuberculosis Case

This section demonstrates Smart-Doc in practice through a full diagnostic workflow. The case is adapted from Mull et al. and corresponds to representative Session `SESS_0W451OZEJ`. Appendix C provides the complete dialogue transcript with intent tags and justifications, the submitted diagnosis, the reflective responses, and the automated evaluation summary.

4.3.1 Case Overview

Presentation. An 85-year-old Spanish-speaking woman with progressive dyspnoea. **Framing.** Prior label of "heart failure". **Correct diagnosis.** Miliary tuberculosis related to infliximab therapy. **Common misdiagnosis.** Acute heart failure exacerbation.

4.3.2 Turn-by-turn highlights

Anamnesis. The learner confirms comorbidities (diabetes, hypertension, rheumatoid arthritis) and begins medication reconciliation. After repeated queries about rheumatoid arthritis therapy, a just-in-time hint to check previous hospital records prompts retrieval of prior documentation and reveals infliximab, which reframes risk and narrows differentials.

Examination. Normal heart sounds and absence of edema reduce the likelihood of decompensated heart failure. Diffuse crackles remain non-specific and defer judgment to imaging.

Investigations. A chest radiograph read as pulmonary vascular congestion briefly supports the initial frame and triggers an anchoring warning. Computed tomography then shows a miliary

pattern, and echocardiography confirms a normal ejection fraction, which redirects the working diagnosis toward disseminated infection rather than cardiac failure.

Table 4.5: Information revealed by category and timing. See Appendix C for the complete dialogue and event log.

Category	Count	Critical findings	Timeline
Presenting symptoms	6	None	First 2 min
Current medications	3	Infliximab	about min 7
Physical examination	3	None	7 to 10 min
Imaging	3	Miliary nodules; normal echo	10 to 15 min
Diagnostic results	3	None	15 to 18 min
Total	**18**	**2 critical**	**18 min**

Micro-excerpts from the transcript.

High-confidence intent mapping.
Doctor: "First, what is her past medical history?"
Intent: `pmh_general` (confidence 0.95).

Imaging request mapped to intent.
Doctor: "Is there any chest X-ray?"
Intent: `imaging_chest_xray` (confidence 0.95).

Scaffolding hint that unlocked the pivot.
System hint: "Maybe you could check her previous hospital records?"

4.3.3 Outcome and reflection

The learner submits *miliary tuberculosis* as the final diagnosis. Automated sub-scores are Information 75, Accuracy 88, Bias 80, which yield an Overall 81 out of 100. In the reflection, the learner links infliximab to tuberculosis reactivation risk, acknowledges contradictions such as a normal echocardiogram despite an elevated BNP, and articulates reasonable alternatives with tests that would arbitrate between them. The evaluation card and the full narrative feedback appear in Appendix C.

Educational signals. Progressive disclosure required 17 targeted queries to reveal 18 information blocks, which indicates active inquiry rather than passive reading. The medication hint successfully guided reconciliation after two unsuccessful attempts. Bias detection was activated when preliminary imaging favored heart failure and was logged as an anchoring warning. Intent classification reached 100% for this session, and median response latency was close to six seconds per turn. Together, these signals show how the platform elicits bias-prone trajectories, supports timely reflection, and records high-value events for later analysis.

4.4 User Interfaces

The user interfaces translate Smart-Doc's pedagogical logic into a low-friction experience that makes reasoning steps visible, encourages reflection, and preserves the authenticity of a clinical encounter. Each screen is organized to foreground the learner's inquiry, to reveal information progressively, and to surface bias-related prompts without breaking immersion..

4.4.1 Simulation interface

The simulation runs in a single-page application with four context tabs that mirror the clinical workflow: *Patient Information*, *Clinical Interview*, *Physical Examination*, and *Labs & Imaging*. Dialogue is chat-based and distinguishes student and system turns by color and alignment.

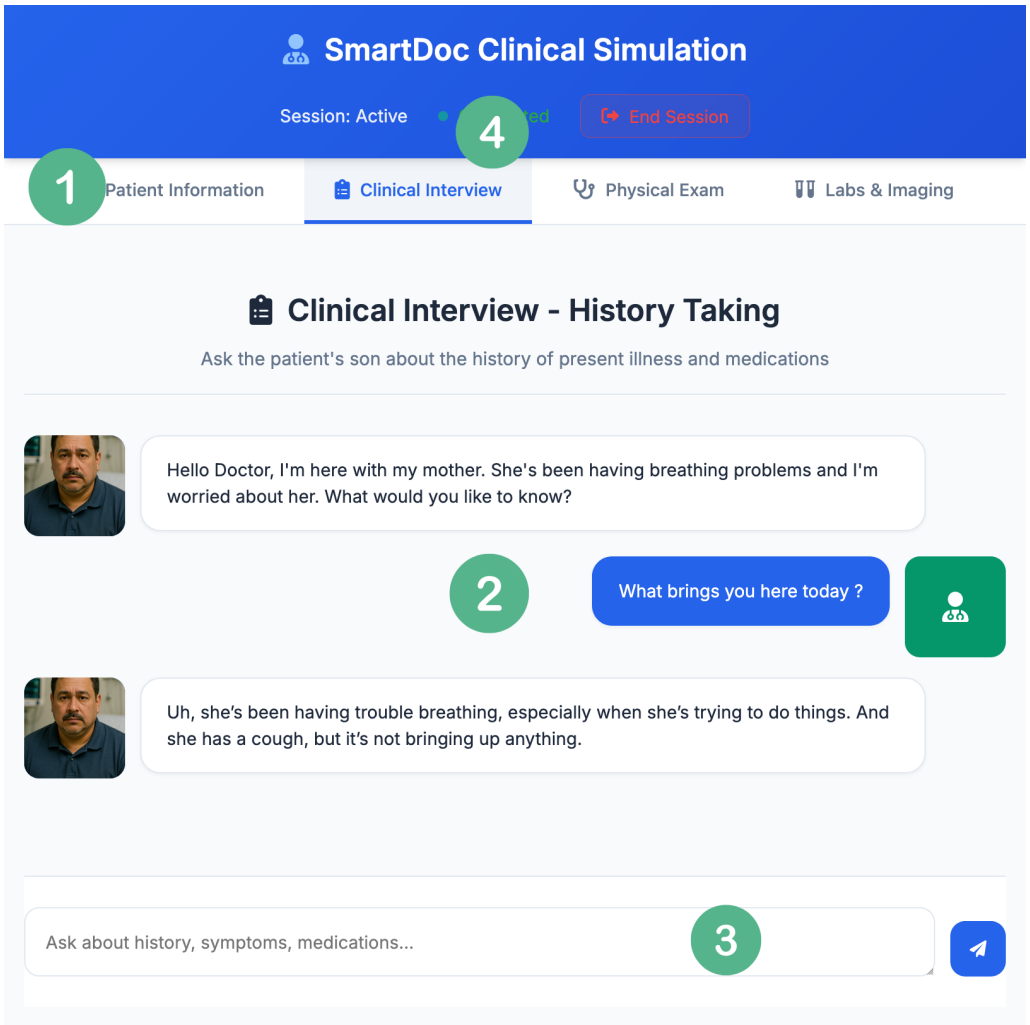

Figure 4.4: Simulation interface overview. Numbered Legend indicates: (1) context tabs aligned to workflow;(2) dialogue pane with student and system turns; (3) input area where prompts are entered; (4) phase indicator that shows whether the learner is in interview, examination, or investigations.

In the patient information tab, revealed information is written to the discovery panel as structured items with concise labels and values. Items are grouped by category and displayed in the order in which they were discovered, which helps learners reconstruct their reasoning and supports later reflection. A counter with the information discovered/yet to reveal information is also shown in this tab, with an indicator of how many bias warnings were triggered so far. This, together with the bias warning card can be disabled in admin page.

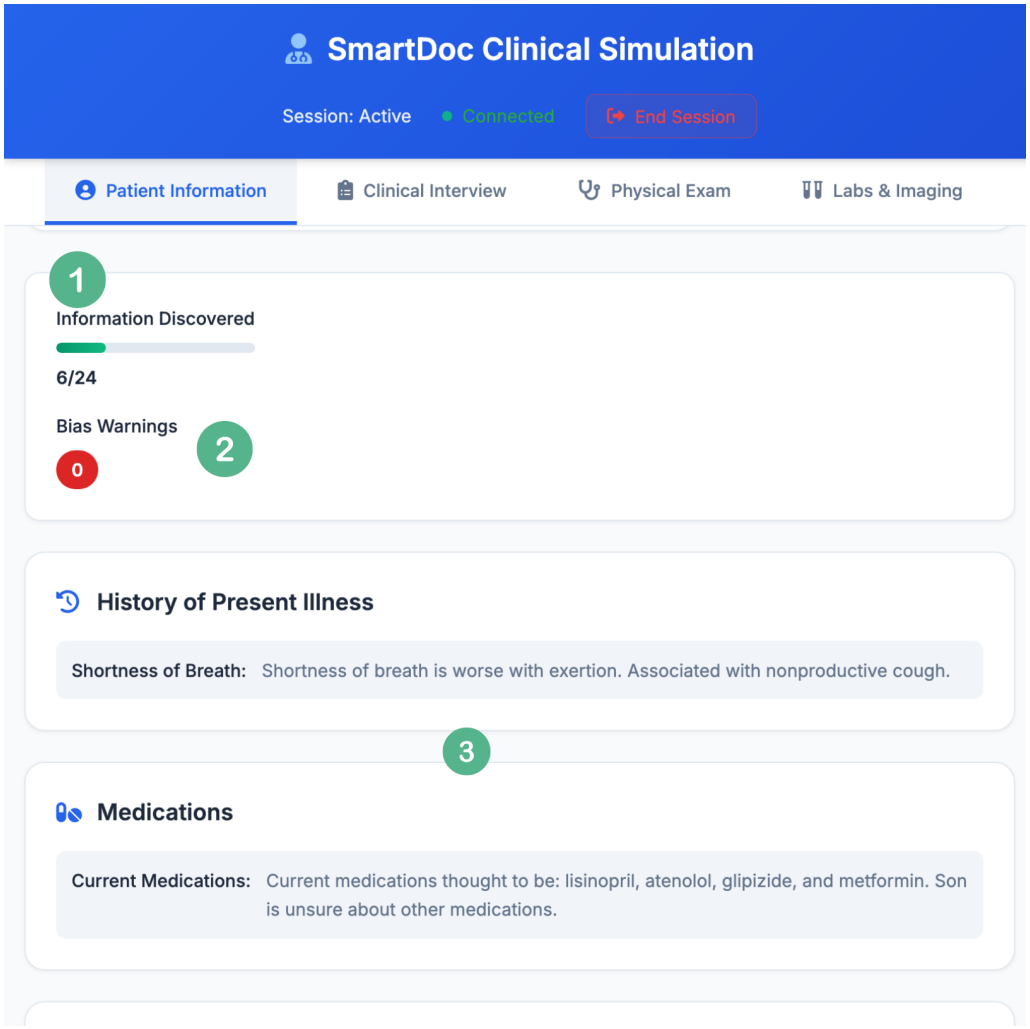

Figure 4.5: Revealed information organized by category. Legend: (1) Information Discovered tracker; (2) Rule-based cognitive-bias tracker; (3) Discovered information blocks, grouped into categories.;

When the system detects a bias-prone pattern, a persistent card appears right below the VSP response. Also, clarification messages are used when intent confidence is low, which preserves conversational flow and avoids hard errors.

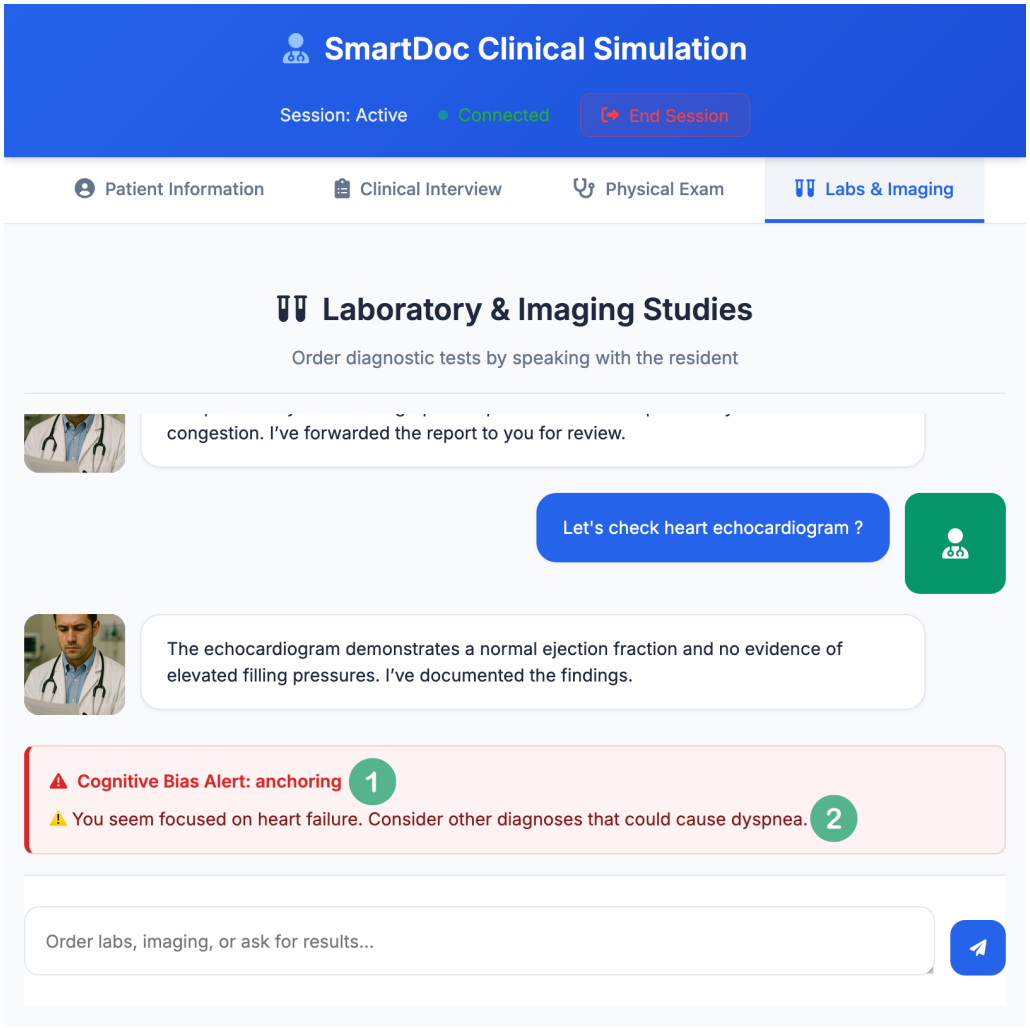

Figure 4.6: Bias warning card in the discovery panel. Legend: (1) bias label, for example, anchoring; (2) short rationale that cites the pattern observed and a meta-cognitive prompt that invites the learner to consider alternatives.

At case conclusion, the submission screen requires a structured diagnosis and a short reflection. The form asks the learner to list supporting and disconfirming features and to record the must-not-miss conditions. The layout encourages a brief, disciplined summary that can be scored automatically and also read quickly by an educator.

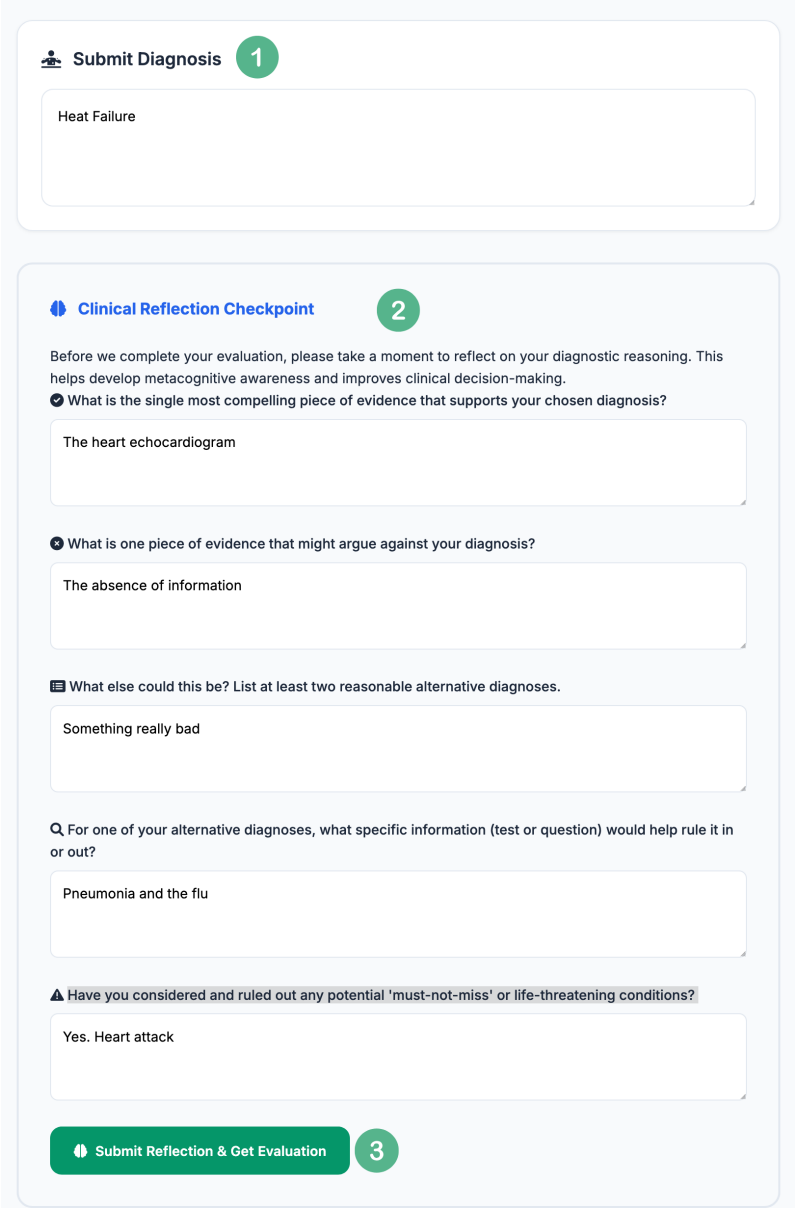

Figure 4.7: Diagnosis and reflection submission. Legend: (1) primary diagnosis entry; (2) metacognitive and reasoning questions that support diagnosis or rule out differential diagnosis; (3) final submission button.

4.4.2 Evaluation results interface

After submission, Smart-Doc generates an automated summary with scores and narrative feedback. The overview card presents an overall score together with the main dimensions used in this study, and a comprehensive narrative highlights specific strengths, opportunities for improvement, key recommendations, and information gathering evaluation, finishing with feedback on diagnostic accuracy and cognitive bias awareness. This pairing of quantitative and qualitative feedback is designed to be actionable during debrief.

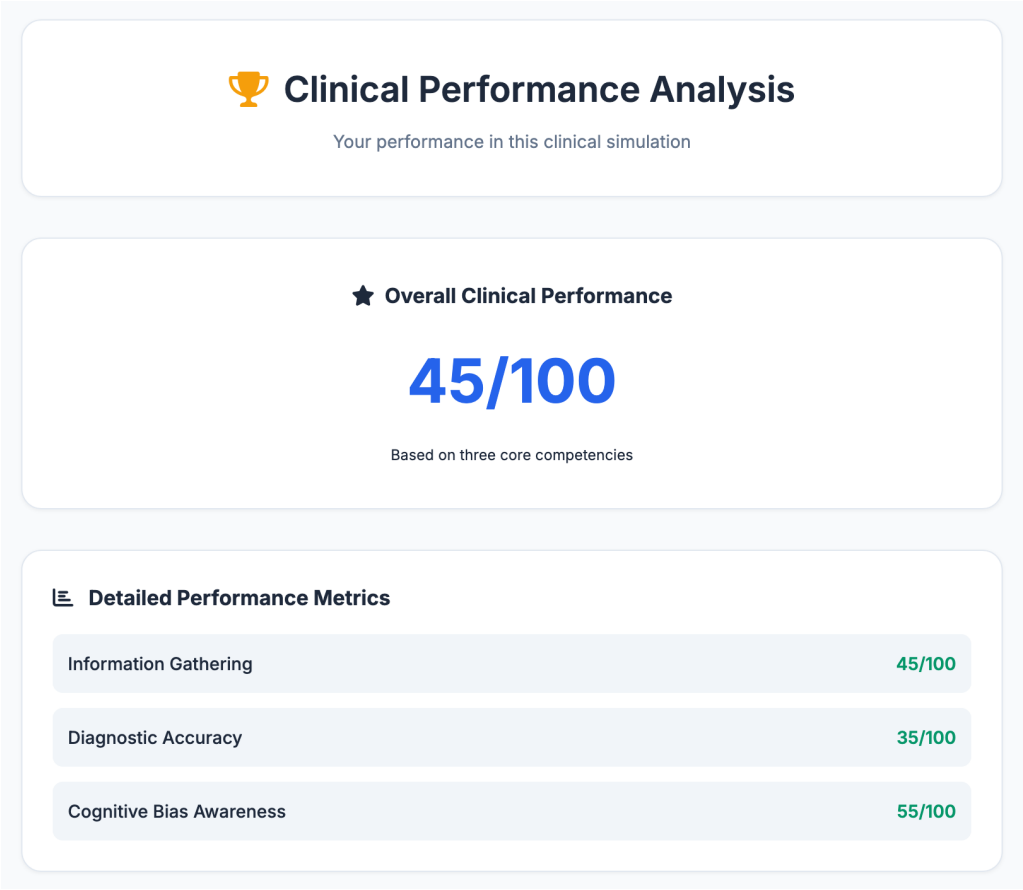

Figure 4.8: Overall performance summary. Legend: (1) overall score; (2) dimension scores for information gathering, diagnostic accuracy, and bias awareness.

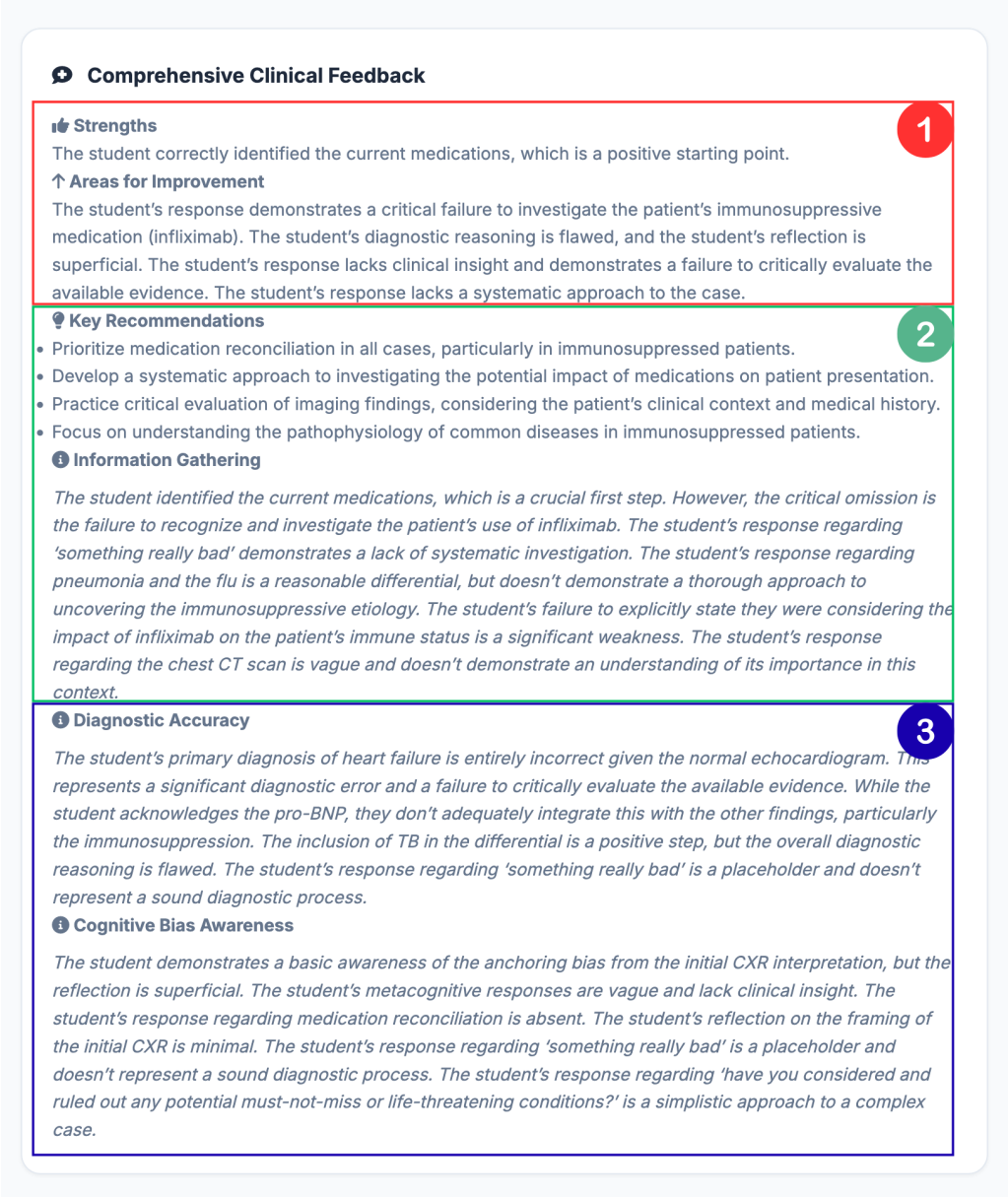

Figure 4.9: Comprehensive narrative clinical feedback. Legend: (1) strengths and areas for improvement; (2) key recommendations and information gathering; (3) diagnostic accuracy and cognitive bias awareness.

4.4.3 Administrative dashboard

System administrators use the dashboard to manage cohorts and to adjust experimental settings. User accounts can be created and assigned to groups, model profiles can be selected for different modules, and prompt templates can be reviewed and versioned. Bias card visibility can be toggled for controlled comparisons, and session data can be exported for analysis.

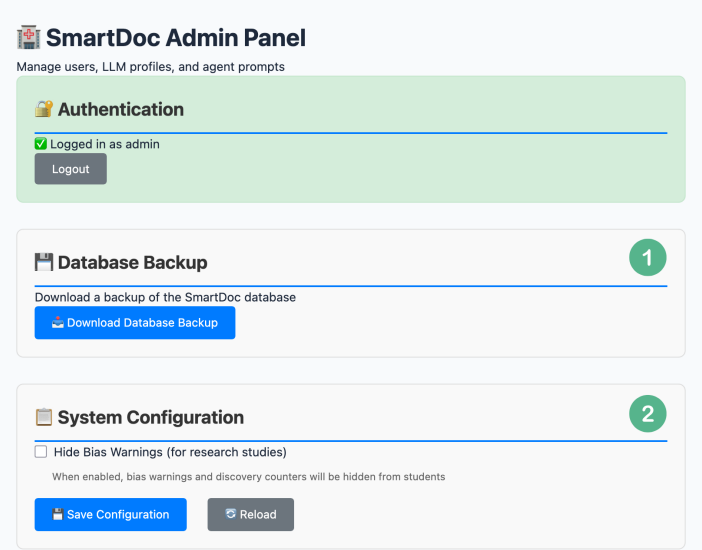

Figure 4.10: Administrative controls for study configuration. Legend: (1) bias card visibility toggle; (2) database backup.

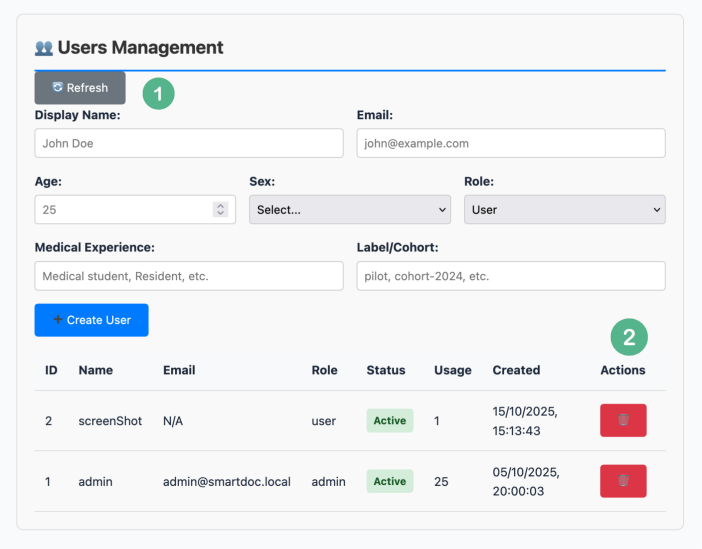

Figure 4.11: User and cohort management. Legend: (1) create and configure user: name, email, age, sex, role, medical experience, and label/cohort identification; (2) user list with actions.

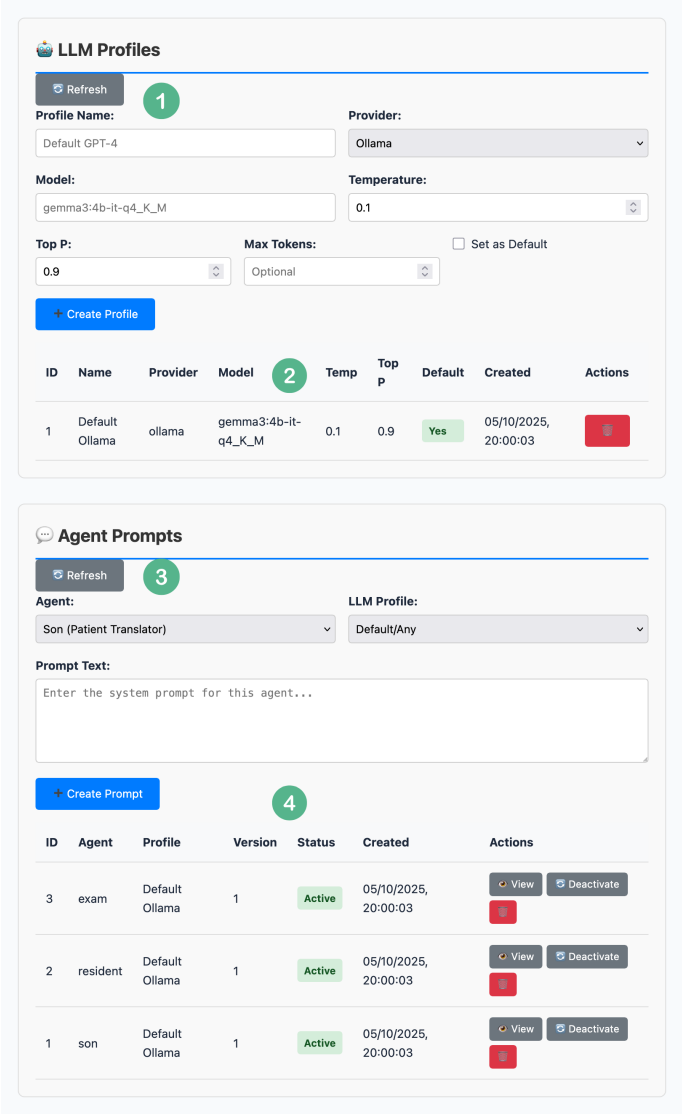

Figure 4.12: Model and prompt configuration. Legend: (1) model profile management: profile name, provider, model, and model configuration; (2) List of configured LLM profiles and actions; (3) Prompt profile management: agent, LLM profile map, and prompt text; (4) list of prompt profiles and actions;

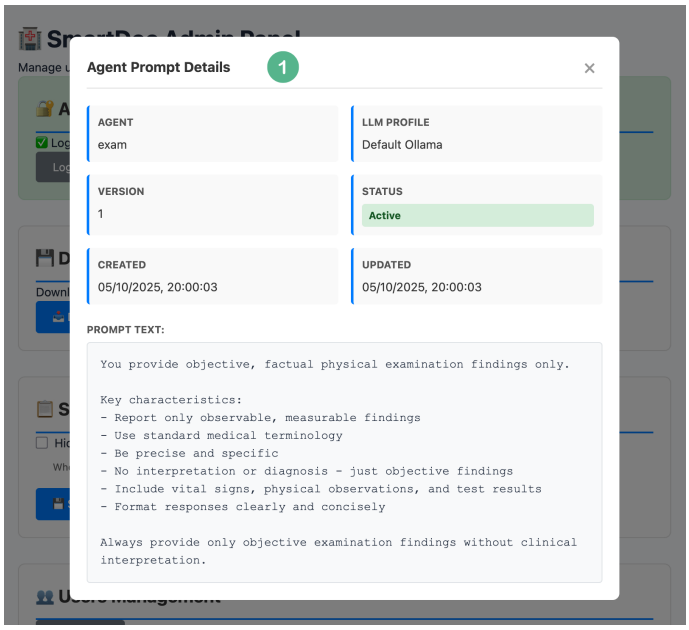

Figure 4.13: Prompt viewer and version history. Legend: (1) prompt text with variables; (2) version timeline; (3) revert and duplicate; (4) preview with sample inputs.

The interface design reflects the platform's pedagogy. Reasoning is made visible through structured discovery, bias prompts invite meta-cognitive oversight at the right moment, and the reflection form requires a concise account of evidence and alternatives. The evaluation view closes the loop with targeted feedback, and the dashboard supports controlled, reproducible studies. In combination, these screens connect the model-driven backend to an intuitive, human-centered experience for learners and educators.

4.5 Summary

This chapter described the development of **Smart-Doc**, translating cognitive theory and prior evidence into a working, bias-aware virtual patient. The design centered on intent-driven progressive disclosure to make the diagnostic process observable, real-time bias monitoring to surface heuristic pitfalls, and structured reflection to externalize reasoning for feedback. Together, these mechanisms operationalize "how to think" rather than only "what to diagnose."

We detailed the system design and its technical realization, including a modular architecture, algorithms for progressive information release with phase-aware intent classification, hybrid bias pattern checks, and LLM integration with schema-validated outputs. The exemplar miliary tuberculosis workflow demonstrated how these components interact during an authentic case, while the user interfaces showed how discoveries, bias prompts, and reflective entries are presented without interrupting immersion.

To avoid pre-empting the empirical analysis, this chapter confined itself to *what the system is* and *how it works*. The next chapter evaluates *how well it works* with clinical interns—reporting diagnostic and educational outcomes, usability, and workload—followed by a discussion of implications, limitations, and future work in Chapter 6. Readers seeking aggregate results, interpretation, and the broader contribution should consult Chapters 5 and 6.

Chapter 5

Evaluation and Results

Building upon the system design in Chapter 4, this chapter evaluates the feasibility, usability, and educational impact of **Smart-Doc**. We examine whether the simulation elicits authentic diagnostic reasoning and whether the automated feedback aligns with cognitive-bias theory.

5.1 Evaluation Methods

5.1.1 Study design and participants

Thirteen clinical interns (all female; age 27) were invited to complete one diagnostic interview using the Mull case (*miliary tuberculosis*) adapted for Smart-Doc. The simulation ran in *post-hoc feedback mode*: live bias prompts were disabled to avoid influencing reasoning during the interview. After submitting a final diagnosis, participants completed five structured reflection prompts and then received automated feedback.

This report analyses the first six complete sessions ($n = 6$) that met pre-registered completeness criteria (final diagnosis, full transcript, complete reflections, and SUS/NASA-TLX submitted). Among these six, four were 2$^{\text{nd}}$-year interns and two were 1$^{\text{st}}$-year; clinical rotations at the time included Pediatrics (three), Plastic Surgery (one), General and Family Medicine (one), and Cardiac Surgery (one).

5.1.2 Evaluation Instruments

We combined usability, workload, reflection quality, and system analytics:

1. **System Usability Scale (SUS)** — 10 items on a 5-point Likert scale scored to 0–100. With responses $x_i \in \{1,\dots,5\}$,

$$\text{SUS} = 2.5 \sum_{i=1}^{10} x_i^*, \quad x_i^* = \begin{cases} x_i - 1, & i \text{ odd} \\ 5 - x_i, & i \text{ even} \end{cases}$$

   Higher is better.

Scoring interpretation. Odd-numbered items represent positive statements ($x_i^* = x_i - 1$), while even-numbered items are negative and inverted ($x_i^* = 5 - x_i$). This normalizes all responses to a 0–4 range. The ten items are summed ($0-40$) and multiplied by 2.5 to yield a final 0–100 usability score, where 68+ denotes acceptable usability.

2. **NASA-TLX (adapted)** — five dimensions (Mental, Temporal, Effort, Frustration, Performance) collected on 0–10 scales. We report an unweighted overall (0–100):

$$\text{TLX}_{\text{overall}} = 10 \times \frac{1}{k} \sum_{d=1}^{k} y_d, \quad k = 5.$$

Note: lower *Performance* scores reflect *better perceived performance*.

Computation details. Ratings across the five subscales are averaged ($\frac{1}{5}\sum y_d$) to obtain a mean workload score (0–10), then multiplied by 10 to scale it to 0–100. Lower values indicate lower perceived workload, with Performance inversely scored.

3. **Targeted Reflection Questionnaire** — five meta-cognitive items on evidence appraisal, counter-evidence, alternatives, and must-not-miss conditions.

4. **Smart-Doc Analytics** — automated rubric-based scoring for *information gathering*, *diagnostic accuracy*, and *cognitive-bias awareness*, with narrative feedback and rule-based bias flags.

Assessor transparency. All scores are system-generated using a fixed rubric and validated JSON outputs. No human raters participated in this pilot; results are *system-derived indicators*. Expert cross-validation is planned.

Given the small cohort, quantitative summaries are per-session values with medians and IQRs. Reflection responses underwent rapid thematic analysis to identify patterns in bias recognition, information use, and differential reasoning.

5.2 Results

5.2.1 Performance outcomes

Table 5.1 summarizes automated sub-scores for the first six complete sessions ($n = 6$). Half of the participants ($3/6$) reached the correct diagnosis (*miliary tuberculosis*). Higher-scoring learners (S1, S2, S6) linked infliximab exposure to TB reactivation, integrated conflicting evidence (normal echocardiogram versus preliminary chest radiograph), and proposed targeted next investigations.
Error patterns. Misdiagnoses typically reflected anchoring on the initial chest radiograph ("pulmonary vascular congestion") and elevated BNP, with insufficient weight given to a normal echocardiogram and incomplete medication reconciliation. Correct cases prioritized the CT miliary pattern and explicitly tied infliximab to TB risk.

Table 5.1: Smart-Doc evaluation results for completed sessions ($n = 6$).

Session	Final diagnosis	Info. gather.	Dx. accuracy	Bias aware.
S1	Miliary TB	65	78	85
S2	Miliary TB	65	75	70
S3	Heart failure exacerbation	65	45	72
S4	Heart failure	40	30	50
S5	Acute heart failure	45	35	55
S6	Miliary TB	75	88	80
Median (IQR)	—	65 (45–65)	60 (35–78)	71 (55–80)

5.2.2 Reflection signals

Thirty reflections (five per participant) were analyzed. Three recurrent signals were associated with correct diagnoses: (i) explicit reference to a miliary CT pattern; (ii) recognition of infliximab as a TB risk factor; and (iii) use of a normal echocardiogram as counter-evidence against heart failure. Incorrect cases (S3–S5) leaned more heavily on BNP/CXR and offered shorter, less decisive counter-arguments.

Table 5.2: Reflection signals by session (presence = Y / absence = N).

Sess.	CT miliary	Infliximab	HF (echo)	BNP	Alternatives	Correct
S1	Y	Y	Y	Y	Y	Y
S2	Y	Y	Y	Y	Y	Y
S3	N	N	Y	N	Y	N
S4	N	N	Y	N	Y	N
S5	N	N	N	N	Y	N
S6	Y	Y	Y	N	Y	Y
Totals (correct)	3/3	3/3	3/3	2/3	3/3	—
Totals (incorrect)	0/3	0/3	2/3	0/3	3/3	—

Illustrative excerpts. **Integrating decisive evidence.**

> "The chest CT showing tiny pulmonary nodules in a diffuse reticular pattern, together with recent infliximab therapy, is the most compelling evidence for miliary TB." (S6)

Weighing contradictions.

> "BNP and crackles could suggest heart failure, but they are nonspecific. The normal cardiac exam and echo make a cardiac cause unlikely." (S6)

These statements show deliberate linkage of imaging, medication history and counter evidence which represents the reflective behavior Smart-Doc aims to strengthen.

5.3 Usability and Workload

Median end-to-end response time was ~ 6 seconds per turn, elevating temporal pressure and mild frustration. Despite this, perceived usability was acceptable and workload remained moderate.

5.3.1 System Usability Scale (SUS)

Table 5.3: SUS scores (0–100) for completed sessions ($n = 6$).

Participant	SUS	Interpretation
P01	70.0	Good
P02	68.0	Acceptable–Good
P03	65.0	Acceptable
P04	60.0	Marginal
P05	62.5	Acceptable
P06	72.0	Good
Median (IQR)	66.5 (62.0–70.0)	Overall “Acceptable Usability”

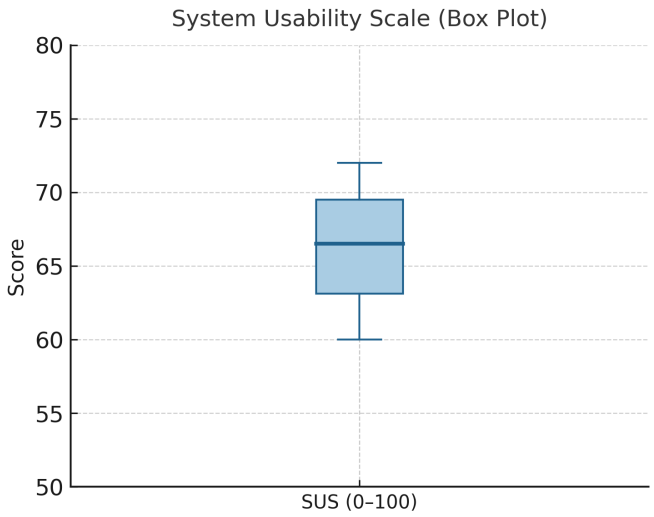

Figure 5.1: SUS distribution across participants.

Interpretation. SUS clustered around 65–70, consistent with “acceptable” usability. Participants reported the interface was easy to learn; response latency slightly reduced perceived smoothness but did not prevent task completion.

5.3.2 NASA-TLX (adapted)

Table 5.4: NASA-TLX workload profile (0–10); unweighted dimensions and median (IQR).

Participant	**Mental**	**Temporal**	**Effort**	**Frustration**	**Performance**
P01	5	7	4	5	3
P02	4	6	4	4	2
P03	5	7	5	6	3
P04	6	8	5	6	4
P05	5	7	4	5	3
P06	4	6	4	4	2
Median (IQR)	5 (4–5)	7 (6–7)	4 (4–5)	5 (4–6)	3 (2–3)

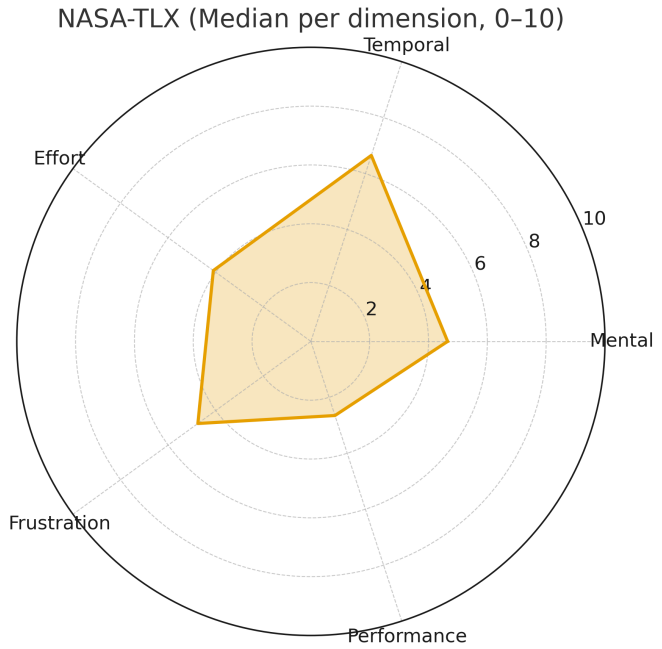

Figure 5.2: NASA–TLX median profile (0–10) by dimension.

Interpretation. The unweighted NASA-TLX overall workload was 48/100, reflecting a **low to moderate perceived workload**. Among the five dimensions, **Temporal Demand** (median 7/10) scored highest, consistent with the observed mean response latency of about six seconds per interaction. **Effort** and **Frustration** (both median 5/10) remained within acceptable ranges, indicating that participants did not perceive the task as overly demanding or stressful. The **Mental Demand** (median 5/10) suggests that Smart-Doc required a meaningful yet manageable level of cognitive engagement. Finally, **Performance** (median 3/10; lower values denote better perceived success) shows that participants generally felt confident and effective during the diagnostic reasoning task.

5.4 Agreement and Caveats

Across the six analyzed sessions, the two analytics strands pointed in the same direction. Rule-based checks of focus and contradiction handling and rubric-based scoring that penalized early closure both identified *anchoring* as the dominant bias in misdiagnosed cases. In correct cases, rubric scores were higher where reflections explicitly integrated counter evidence and revised the working diagnosis, which is consistent with the rule-based pattern.

These signals should be read with caution. The analysis reflects a small cohort on a single case, collected in post hoc feedback mode with bias cards disabled. Bias markers are inferred from interaction patterns and depend on intent classification and threshold choices, and we did not assess model or seed variance or compare automated judgments with expert ratings in this iteration.

A fuller interpretation of these findings and the validation plan for expert concordance, sensitivity analyses, and robustness checks are presented in Chapter 6.

Chapter 6

Discussion and Conclusion

6.1 Overview

This chapter interprets the findings from Chapter 5 in the context of the literature on cognitive bias and diagnostic error. We discuss how Smart-Doc reproduced the reasoning trajectory in Mull et al.'s case [46], how learners responded to bias-prone situations, and the implications for simulation design.

6.2 Interpreting the Findings

Bias patterns reproduced by Smart-Doc. Smart-Doc mirrored the cognitive trajectory in Mull's case of an elderly Spanish-speaking woman misdiagnosed with heart failure [46]. In our sessions, premature fixation on a cardiac explanation followed a suggestive chest radiograph and BNP, while contradictory evidence (a normal echocardiogram) and incomplete medication reconciliation were under-weighted in missed diagnoses. In contrast, correct cases cited the miliary CT pattern and infliximab therapy as decisive. This behavioral similarity suggests that Smart-Doc recreated the bias-prone environment of the original case within a safe, debriefable setting.

Reflection and meta-cognitive engagement. Structured reflection prompted deliberate (System 2) reasoning. High performers explicitly integrated counter-evidence; lower performers acknowledged the bias conceptually but struggled to operationalize it in their diagnostic path. This aligns with prior work showing that recognizing a bias does not automatically correct it without guided practice and feedback.

Usability, workload, and cognitive load. Usability was acceptable (median SUS ≈ 67) and workload moderate. Temporal demand and frustration reflected the ~ 6 s response latency. Although technical rather than organizational, such constraints approximate real-world cognitive stressors (time pressure, information gaps), reinforcing the authenticity of the training context.

6.3 Comparison with the Case Inspiration

Cognitive bias typology. Mull et al. described multiple biases relevant to the index case, including framing, anchoring, diagnostic momentum, availability, confirmation, and overconfidence. Smart-Doc reproduced at least four within-learner interactions: anchoring, confirmation, availability, and premature closure. This enabled bias to be observed, measured, and debriefed rather than merely described.

Systemic versus cognitive contributions. While Mull's analysis highlighted systemic factors (language barriers, incomplete records), Smart-Doc isolates cognition by controlling information symmetry and timing. The design makes heuristics visible without confounding institutional noise, clarifying where meta-cognitive support can help.

Embedding cognitive bias into an interactive interview transforms a sentinel event into a learning scaffold. Smart-Doc's post-diagnosis reflection functions as a "diagnostic time-out" [46], prompting learners to articulate support, counter-evidence, and alternatives before closure.

6.4 Addressing the Research Questions and Objectives

This pilot study set out to answer three core research questions (RQ1–RQ3), as established in Chapter 1.

RQ1: Eliciting and Detecting Cognitive Biases *To what extent can an LLM-powered simulation realistically elicit and detect cognitive biases in diagnostic interviews?*
Answer: Smart-Doc successfully reproduced bias-prone trajectories in 50% of sessions (3/6 missed diagnoses). Anchoring on preliminary imaging and elevated BNP was identified through rule-based checks and LLM-supported reflection scoring. The system detected premature closure and confirmation bias through session analytics and reflection content. However, bias detection relied primarily on automated rubrics without expert validation, limiting the strength of these claims.
Evidence: Three sessions demonstrated anchoring on cardiac hypotheses; automated scoring flagged insufficient evidence integration (Dx Accuracy scores 30–45 for incorrect cases). Reflection analysis showed learners acknowledged bias conceptually but struggled to operationalize counter-strategies.

RQ2: Effectiveness of meta-cognitive Prompts *How effective are meta-cognitive prompts—delivered in real time or post-hoc—in fostering reflection and reducing diagnostic bias?*
Answer: Post-hoc structured reflection prompts successfully elicited meta-cognitive engagement. Learners in correct-diagnosis sessions (S1, S2, S6) demonstrated explicit counter-evidence reasoning and systematic hypothesis evaluation. However, live bias prompts were disabled in this pilot, preventing direct comparison of real-time versus retrospective interventions.

Evidence: Reflection analysis showed that high performers explicitly integrated contradictory evidence. Bias awareness scores (median 71/100) suggest moderate meta-cognitive engagement, though the educational impact of prompts requires longitudinal follow-up to assess behavior change in subsequent cases.

RQ3: Design Principles for Scalable Deployment

What technical and pedagogical design principles enable the scalable deployment of bias-aware virtual patient simulations?

Answer: Smart-Doc's hybrid architecture—progressive disclosure via intent classification, LLM-powered dialogue, and automated evaluation rubrics—proved feasible for clinical intern use. Usability was acceptable (SUS median 66.5), though technical latency (6s per turn) and temporal demand scores indicate areas for optimization. The system's modular design (intent engine, disclosure manager, evaluator) supports extension to additional cases and bias archetypes.

Evidence: All six participants completed the full interaction independently with no technical failures. The case-agnostic architecture and JSON-based case definitions enable rapid case authoring. However, scale-up will require latency reduction, multi-institutional testing, and expert calibration of evaluation rubrics.

Summary: Research questions RQ1 and RQ3 are affirmatively answered with caveats regarding validation scale. RQ2 is partially answered—post-hoc prompts show promise, but real-time prompt effectiveness remains untested. How did we address the objectives to answer these research questions?

Table 6.1 summarizes how the five research objectives established in Chapter 1 were addressed throughout this dissertation.

Table 6.1: Achievement of Research Objectives

Objective	How Addressed
Simulation: Design and implement an AI-powered VP capable of realistic, unscripted interviews based on bias-eliciting cases.	Developed Smart-Doc platform with LLM-powered dialogue, intent-driven progressive disclosure, and the Mull et al. miliary TB case. Achieved naturalistic interaction with 100% intent classification accuracy in pilot testing.
Bias detection: Identify behavioral markers of anchoring, confirmation bias, and premature closure using rule-based and LLM-assisted methods.	Implemented hybrid detection combining rule-based checks (medication reconciliation, information completeness) with LLM-powered reflection analysis. Successfully identified bias patterns in 50% of pilot sessions (3/6 incorrect diagnoses).
meta-cognitive tutoring: Deliver context-aware prompts that stimulate reflection and encourage reconsideration of reasoning paths when bias is detected.	Designed and deployed five structured reflection prompts requiring explicit articulation of supporting/counter-evidence and alternative diagnoses. Post-hoc prompts demonstrated effectiveness; real-time prompts remain to be evaluated.
Evaluation framework: Develop analytics for diagnostic accuracy, information gathering, and manifestations of cognitive bias to support formative and summative feedback.	Created an automated LLM-based evaluation framework across three dimensions (information gathering, diagnostic accuracy, cognitive bias awareness) with JSON-validated scoring rubrics and comprehensive narrative feedback.
Effectiveness study: Conduct a pilot with clinical interns to assess usability, educational effectiveness, and impact on bias awareness.	Completed pilot study with 6 clinical interns, achieving acceptable usability, moderate cognitive load, and evidence of bias reproduction and detection. Results inform future scale-up and validation studies.

All five objectives were successfully addressed, with the simulation, bias detection, and evaluation framework objectives fully achieved in the current implementation. The meta-cognitive tutoring objective was partially achieved (post-hoc prompts validated; real-time prompts awaiting future work), and the effectiveness study objective completed as planned for this pilot phase, with larger validation studies recommended as next steps.

6.5 Positioning Smart-Doc in the Simulation Landscape

To contextualize Smart-Doc's contribution, it is useful to contrast its design with the AI-powered virtual patient systems reviewed in Chapter 3. That scoping review identified several platforms demonstrating feasibility and effectiveness for foundational skill development, particularly in history taking and clinical communication. Systems such as Hepius [19] leverage natural language processing and intelligent tutoring system architectures to provide structured feedback, while others like ViPATalk [39] have shown equivalence to traditional role play for history taking completeness. More recently, platforms have begun integrating large language models for more naturalistic dialogue, as demonstrated by Holderried et al. [25, 26] and Brügge et al. [6], who reported gains in clinical decision making with LLM-generated feedback.

A key gap identified in Chapter 3 was the absence of systems explicitly designed to surface and address *cognitive biases* during diagnostic reasoning. While some platforms employ progressive disclosure or provide decision-making feedback, none systematically detect bias-prone behaviors (such as anchoring, premature closure, or confirmation bias) or scaffold meta-cognitive reflection that targets these errors. Smart-Doc addresses this gap by combining three elements: (i) LLM-powered open-ended dialogue that allows natural information seeking to emerge, (ii) real-time detection of bias markers through rule-based and semantic analysis of interaction patterns, and (iii) structured prompts that require learners to articulate supporting evidence, counter evidence, and alternative hypotheses. This integration of bias-aware design with conversational realism distinguishes Smart-Doc from simulations that prioritize either naturalism *or* pedagogy, but rarely both, in the service of debiasing. In addition, Smart-Doc's case agnostic, declarative architecture—where bias triggers and disclosure logic are encoded in data rather than hard coded—supports rapid authoring and curriculum integration at scale.

Educational implications. Three design implications follow directly from this integration: (1) *progressive disclosure as a bias trigger*—staged revelation recreates evolving uncertainty, making errors observable and therefore teachable; (2) *structured reflection as a corrective*—prompts that demand explicit counter evidence help learners slow down and engage analytic reasoning; and (3) *AI mediated feedback as scalable mentorship*—automated, standardised feedback can extend access to diagnostic reasoning training and can be calibrated with expert review.

6.6 Limitations and Future Work

Building on the pilot, we outline the main constraints of the present study and the priorities to address them. *Current limitations* include a single case with a small, demographically homogeneous cohort; fully automated scoring without expert adjudication; post-hoc reflection mode with live bias prompts disabled; and system latency that contributed to temporal workload. Model and seed variance were not assessed. The plan below balances near-term improvements with validation at scale.

6.6.1 Engineering and architecture

Response latency optimization. Reduce median turn time from ~ 6 s toward < 2 s to improve temporal workload and conversational flow. Strategies include model caching, prompt compaction, and asynchronous intent classification.

Modular bias detection pipelines. Separate rule-based heuristics from model-assisted semantic checks to allow independent tuning and transparent auditing. Add brief explanations that clarify why a bias flag fired.

Integration with learning systems. Develop LTI connectors and SCORM-compatible exports so Smart-Doc can be embedded in institutional VLEs with grade passback and competency tracking.

Multi-modal interaction support. Explore voice input and synthesis to approximate clinical encounters and lower typing load, while monitoring latency and ASR accuracy trade-offs.

6.6.2 Pedagogy and analytics

Real-time bias prompts. Activate just-in-time prompts when bias markers are detected and compare in-situ prompting with post-hoc reflection to estimate differential impact on accuracy and reasoning depth.

Comparative analytics dashboard. Provide educator views of cohort-level performance, bias frequency, and longitudinal progression. Anonymized benchmarking can surface common reasoning pitfalls for targeted teaching.

Longitudinal skill tracking. Enable learner accounts with session histories to measure whether repeated exposure reduces premature closure, increases use of counter-evidence, and improves diagnostic justification over time.

Expanded case library. Author additional bias-engineered cases (e.g., availability, representativeness) using templated case schemas so skills can be assessed across archetypes and specialties.

6.6.3 Validation and scale

Expert validation framework. Establish concordance between automated analytics and clinician ratings on bias labels and diagnostic scoring (e.g., Cohen's κ or Krippendorff's α with bootstrap CIs). Predefine rater calibration and tie-break rules.

Multi-institutional trials. Extend to multiple sites (target $n > 50$) to assess generalisability across curricula and learner backgrounds. Compare Smart-Doc-trained vs. standard training on standardized patient assessments.

Hybrid evaluation approaches. Combine quantitative metrics (accuracy, bias scores) with qualitative methods (think-aloud, focus groups) to explain how learners internalize and apply metacognitive strategies.

Long-term competency assessment. Follow graduates into clinical settings, where feasible, to examine downstream effects on diagnostic error, closure times, and patient outcomes.

Together, these priorities aim to move Smart-Doc from a promising pilot to a validated educational tool: faster and easier to use, richer pedagogically, and supported by rigorous evidence across cases, cohorts, and institutions.

6.7 Conclusion

This dissertation introduced **Smart-Doc**, an AI-powered virtual patient that operationalizes cognitive-bias training within realistic clinical simulations. By reproducing the diagnostic trajectory of a sentinel case in a safe, debriefable setting, the platform shows that model-mediated dialogue can turn real-world failure modes into structured learning opportunities rather than patient-safety events.

Principal contributions. *First*, the work demonstrates that cognitive biases—traditionally invisible aspects of clinical reasoning—can be deliberately elicited, detected, and measured in an open-ended simulation. The observed error rate, consistent with the index case, indicates that the environment successfully recreated bias-prone conditions. *Second*, Smart-Doc establishes a hybrid architectural pattern that couples intent-driven progressive disclosure with model-generated dialogue and schema-validated analytics, offering a template for systems that require both pedagogical structure and conversational naturalism. *Third*, the automated evaluation framework produces dimensional feedback across information gathering, diagnostic accuracy, and bias awareness, suggesting a scalable complement to labor-intensive expert assessment while acknowledging the need for expert calibration.

Implications for clinical education. Three design principles follow. Progressive disclosure creates teachable moments by structuring uncertainty and delaying closure until discriminating evidence is sought. Structured reflection transforms tacit judgement into explicit artefacts—supporting features, contradictions, and alternatives—that can be discussed, scored, and improved. AI-mediated feedback provides consistent, immediate guidance that can extend diagnostic-reasoning practice beyond the limits of apprenticeship time.

Path forward. The pilot establishes feasibility and educational promise but requires systematic validation. Near-term priorities include reducing response latency, calibrating automated judgments against expert ratings, and expanding a bias-engineered case library to probe transfer across archetypes and specialties. Medium-term work should compare just-in-time prompting with post hoc reflection, add cohort analytics, and track skill development across repeated cases. Longer-term research should pursue multi-site studies and, where feasible, follow learners into practice to test whether repeated exposure shifts habits of inquiry and reduces bias-related diagnostic error.

Closing perspective. Smart-Doc shows that technology can render invisible cognitive processes visible and teachable. By aligning cognitive science, clinical pedagogy, and modern language models, the platform points toward scalable bias-awareness training that complements clinical

supervision. With validation at scale, this approach can help cultivate reflective, error-resistant diagnostic reasoning and strengthen how clinicians are prepared to recognize and correct the reasoning flaws that lead to preventable harm.

References

[1] Angelina Anthamatten and Jo Ellen Holt. "Integrating Artificial Intelligence Into Virtual Simulations to Develop Entrustable Professional Activities". In: *The Journal for Nurse Practitioners* 20.9 (Oct. 2024). Section: 0, p. 105192. ISSN: 1555-4155. DOI: 10.1016/j.nurpra.2024.105192. URL: https://www.sciencedirect.com/science/article/pii/S155541552400268X.

[2] Marie-Claude Audétat et al. "Diagnosis and management of clinical reasoning difficulties: Part I. Clinical reasoning supervision and educational diagnosis". In: *Medical Teacher* 39.8 (Aug. 2017). Publisher: Taylor & Francis, pp. 792–796. ISSN: 0142-159X. DOI: 10.1080/0142159X.2017.1331033. URL: https://www.tandfonline.com/doi/citedby/10.1080/0142159X.2017.1331033 (visited on 07/24/2025).

[3] Kees van den Berge and Sílvia Mamede. "Cognitive diagnostic error in internal medicine". English. In: *European Journal of Internal Medicine* 24.6 (Sept. 2013). Publisher: Elsevier, pp. 525–529. ISSN: 0953-6205, 1879-0828. DOI: 10.1016/j.ejim.2013.03.006. URL: https://www.ejinme.com/article/S0953-6205(13)00090-3/abstract (visited on 07/24/2025).

[4] Eta S. Berner and Mark L. Graber. "Overconfidence as a Cause of Diagnostic Error in Medicine". In: *The American Journal of Medicine* 121.5, Supplement (2008), S2–S23. DOI: 10.1016/j.amjmed.2008.01.001. URL: https://doi.org/10.1016/j.amjmed.2008.01.001.

[5] Tom Brown et al. "Language Models are Few-Shot Learners". In: *NeurIPS*. 2020.

[6] Emilia Brügge et al. "Large language models improve clinical decision making of medical students through patient simulation and structured feedback: a randomized controlled trial". In: *BMC Medical Education* 24.1 (Nov. 2024), p. 1391. ISSN: 1472-6920. DOI: 10.1186/s12909-024-06399-7. URL: https://doi.org/10.1186/s12909-024-06399-7 (visited on 07/24/2025).

[7] Kai Siang Chan and Nabil Zary. "Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review". In: *JMIR Med Educ* 5.1 (June 2019), e13930. ISSN: 2369-3762. DOI: 10.2196/13930. URL: http://www.ncbi.nlm.nih.gov/pubmed/31199295.

[8] Bernard Charlin, Jacques Tardif, and Valérie Dory. "The Script Concordance Test: a tool to assess the reflective clinician". In: *Teaching and Learning in Medicine* 12.4 (2000), pp. 189–195.

[9] Hyung Won Chung et al. "Scaling Instruction-Finetuned Language Models". In: *arXiv:2210.11416* (2022).

[10] Michael Co, Tsz Hon John Yuen, et al. "Using clinical history taking chatbot mobile app for clinical bedside teachings – A prospective case control study". In: *Heliyon* 8.6 (June 2022). Section: 0, e09751. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2022.e09751. URL: https://www.sciencedirect.com/science/article/pii/S2405844022010398.

[11] Pat Croskerry. "A universal model of diagnostic reasoning". In: *Academic Medicine* 84.8 (2009), pp. 1022–1028.

[12] Pat Croskerry. "Cognitive Forcing Strategies in Clinical Decision-making". In: *Annals of Emergency Medicine* 41.1 (2003), pp. 110–120. DOI: 10.1067/mem.2003.22. URL: https://westerned.org/wp-content/uploads/2017/10/croskerry-cognitive-forcing-strategies-1.pdf.

[13] Pat Croskerry. "Diagnostic Failure: A Cognitive and Affective Approach". In: *Advances in Patient Safety* 2 (2005), pp. 213–242. URL: https://www.ahrq.gov/downloads/pub/advances/vol2/Croskerry.pdf.

[14] Pat Croskerry. "The Importance of Cognitive Errors in Diagnosis and Strategies to Minimize Them". en-US. In: *Academic Medicine* 78.8 (Aug. 2003), p. 775. ISSN: 1040-2446. URL: https://journals.lww.com/academicmedicine/abstract/2003/08000/the_importance_of_cognitive_errors_in_diagnosis.3.aspx (visited on 07/24/2025).

[15] Leticia De Mattei, Marcelino Q. Morato, et al. "Are Artificial Intelligence Virtual Simulated Patients (AI-VSP) a Valid Teaching Modality for Health Professional Students?" In: *Clinical Simulation in Nursing* 92 (July 2024). Section: 0, p. 101536. ISSN: 1876-1399. DOI: 10.1016/j.ecns.2024.101536. URL: https://www.sciencedirect.com/science/article/pii/S1876139924000288.

[16] Arthur S. Elstein and Alan Schwarz. "Clinical problem solving and diagnostic decision making: selective review of the cognitive literature". In: *BMJ* 324.7339 (2002), pp. 729–732.

[17] John W. Ely, Mark L. Graber, and Pat Croskerry. "Checklists to Reduce Diagnostic Errors". In: *Academic Medicine* 86.3 (2011), pp. 307–313.

[18] Kevin W. Eva. "What every teacher needs to know about clinical reasoning". In: *Medical Education* 39.1 (2005), pp. 98–106.

[19] R. Furlan, M. Gatti, et al. "A natural language processing-based virtual patient simulator and intelligent tutoring system for the clinical diagnostic process: Simulator development and case study". English. In: *JMIR Medical Informatics* 9.4 (2021). Section: 0. ISSN: 22919694 (ISSN). DOI: 10.2196/24073. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85104150902&doi=10.2196%2f24073&partnerID=40&md5=58472015a9d71e929b204f33b773715b.

[20] Raffaello Furlan, Mauro Gatti, et al. "Learning Analytics Applied to Clinical Diagnostic Reasoning Using a Natural Language Processing–Based Virtual Patient Simulator: Case Study". In: *JMIR Med Educ* 8.1 (Mar. 2022), e24372. ISSN: 2369-3762. DOI: 10.2196/24372. URL: https://mededu.jmir.org/2022/1/e24372.

[21] Mark L. Graber, Nancy Franklin, and Ruthanna Gordon. "Diagnostic Error in Internal Medicine". In: *Archives of Internal Medicine* 165.13 (July 2005), pp. 1493–1499. ISSN: 0003-9926. DOI: 10.1001/archinte.165.13.1493. URL: https://doi.org/10.1001/archinte.165.13.1493 (visited on 07/24/2025).

[22] Mark L. Graber, Susan Kissam, Victoria L. Payne, et al. "Cognitive Interventions to Reduce Diagnostic Error: A Narrative Review". In: *BMJ Quality & Safety* 21.7 (2012), pp. 535–557. DOI: 10.1136/bmjqs-2011-000603. URL: https://ajustnhs.com/wp-content/uploads/2012/10/cog-interventions-diag-error-graber-BMJ-QS-2012.pdf.

[23] John R. Hampton et al. "Relative contributions of history, examination and laboratory investigation to diagnosis and management". In: *BMJ* 2.5969 (1975), pp. 486–489.

[24] Michael Hindelang, Sebastian Sitaru, et al. "Transforming Health Care Through Chatbots for Medical History-Taking and Future Directions: Comprehensive Systematic Review". In: *JMIR Med Inform* 12 (Aug. 2024), e56628. ISSN: 2291-9694. DOI: 10.2196/56628. URL: https://doi.org/10.2196/56628.

[25] F. Holderried, C. Stegemann-Philipps, et al. "A Generative Pretrained Transformer (GPT)-Powered Chatbot as a Simulated Patient to Practice History Taking: Prospective, Mixed Methods Study". English. In: *JMIR Medical Education* 10.1 (2024). Section: 0. ISSN: 23693762 (ISSN). DOI: 10.2196/53961. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85186223136&doi=10.2196%2f53961&partnerID=40&md5=9bec97eec809b08be71e01f26b31ad49.

[26] F. Holderried, C. Stegemann-Philipps, et al. "A Language Model–Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study". English. In: *JMIR Medical Education* 10 (2024). Section: 0. ISSN: 23693762 (ISSN). DOI: 10.2196/59213. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85203656572&doi=10.2196%2f59213&partnerID=40&md5=5166ee32f4b835a20b20fa4db8

[27] S Iqbal, S Ahmad, et al. "Review article: Impact of Artificial Intelligence in Medical Education [version 1]". In: *MedEdPublish* 10.41 (2021). DOI: 10.15694/mep.2021.000041.1.

[28] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* (2023).

[29] Jingrong Du, Xiaowen Zhu, et al. "History-taking level and its influencing factors among nursing undergraduates based on the virtual standardized patient testing results: Cross sectional study". In: *Nurse Education Today* 111 (Apr. 2022), p. 105312. ISSN: 0260-6917. DOI: 10.1016/j.nedt.2022.105312. URL: https://www.sciencedirect.com/science/article/pii/S026069172200048X.

[30] Daniel Kahneman. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux, 2011.

[31] Jared Kaplan et al. "Scaling Laws for Neural Language Models". In: *arXiv:2001.08361* (2020).

[32] Vladimir Karpukhin et al. "Dense Passage Retrieval for Open-Domain Question Answering". In: *EMNLP*. 2020.

[33] Takeshi Kojima et al. "Large Language Models are Zero-Shot Reasoners". In: *NeurIPS* (2022).

[34] David Kolb. *Experiential Learning: Experience As The Source Of Learning And Development*. Vol. 1. Jan. 1984. ISBN: 0132952610.

[35] Josepha Kuhn et al. "Teaching medical students to apply deliberate reflection in diagnosing new cases". In: *Medical Teacher* (2023). Open Access, pp. 1–9. DOI: 10.1080/0142159X.2023.2229504. URL: https://pure.eur.nl/files/98659381/Teaching_medical_students_to_apply_deliberate_reflection.pdf.

[36] Hyunsu Lee. "The rise of ChatGPT: Exploring its potential in medical education". In: *Anatomical Sciences Education* 17.5 (2024), pp. 926–931. DOI: https://doi.org/10.1002/ase.2270. eprint: https://anatomypubs.onlinelibrary.wiley.com/doi/pdf/10.1002/ase.2270. URL: https://anatomypubs.onlinelibrary.wiley.com/doi/abs/10.1002/ase.2270.

[37] Patrick Lewis et al. "Retrieval-augmented generation for knowledge-intensive NLP tasks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020. URL: https://arxiv.org/abs/2005.11401.

[38] Kai Liang et al. "Unknown Intent Detection for Safer Healthcare Chatbots". In: *Applied Sciences* 14.3 (2024), p. 987. DOI: 10.3390/app14030987. URL: https://www.mdpi.com/2076-3417/14/3/987.

[39] A. Lippitsch, J. Steglich, et al. "Development and evaluation of a software system for medical students to teach and practice anamnestic interviews with virtual patient avatars". English. In: *Computer Methods and Programs in Biomedicine* 244 (2024). Section: 0. ISSN: 01692607 (ISSN). DOI: 10.1016/j.cmpb.2023.107964. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85178088140&doi=10.1016%2fj.cmpb.2023.107964&partnerID=40&md5=35eee88559cfc7d7a3dfe6bd636a4bd6

[40] Samuel Louvan and Bernardo Magnini. "Recent Neural Methods on Slot Filling and Intent Classification for Task-Oriented Dialogue Systems: A Survey". In: *arXiv preprint arXiv:2011.00564* (2020). URL: https://arxiv.org/abs/2011.00564.

[41] K.R. Maicher, A. Stiff, et al. "Artificial intelligence in virtual standardized patients: Combining natural language understanding and rule based dialogue management to improve conversational fidelity". English. In: *Medical Teacher* 45.3 (2023). Section: 0, pp. 279–285. ISSN: 0142159X (ISSN). DOI: 10.1080/0142159X.2022.2130216. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85143812257&doi=10.1080%2f0142159X.2022.2130216&partnerID=40&md5=17258e26e7ad2a440646ab87e0adc

[42] K.R. Maicher, L. Zimmerman, et al. "Using virtual standardized patients to accurately assess information gathering skills in medical students". English. In: *Medical Teacher* 41.9 (2019). Section: 0, pp. 1053–1059. ISSN: 0142159X (ISSN). DOI: 10.1080/0142159X.2019.1616683. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067844527&doi=10.1080%2f0142159X.2019.1616683&partnerID=40&md5=7d4c7526bb54969e298f64c6b1c3b172.

[43] Silvia Mamede and Henk G Schmidt. "The structure of reflective practice in medicine". en. In: *Medical Education* 38.12 (2004). _eprint: https://asmepublications.onlinelibrary.wiley.com/doi/pdf/10.11 2929.2004.01917.x, pp. 1302–1308. ISSN: 1365-2923. DOI: 10.1111/j.1365-2929.2004.01917.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2929.2004.01917.x (visited on 07/24/2025).

[44] Silvia Mamede et al. "Influence of deliberate reflection on diagnostic accuracy: a randomized controlled trial". In: *JGE* 3.2 (2008), pp. 113–121.

[45] Sílvia Mamede and Henk G. Schmidt. "Deliberate Reflection and Clinical Reasoning: Founding Ideas and State of the Science". In: *Medical Education* 57.1 (2022), pp. 76–85. DOI: 10.1111/medu.14863. URL: https://asmepublications.onlinelibrary.wiley.com/doi/pdf/10.1111/medu.14863.

[46] Nikhil Mull, James B. Reilly, and Jennifer S. Myers. “An elderly woman with ‘heart failure’: Cognitive biases and diagnostic error”. In: *Cleveland Clinic Journal of Medicine* 82.11 (2015). Open access article illustrating diagnostic error and cognitive bias, pp. 745–753. DOI: 10.3949/ccjm.82a.14087. URL: https://www.ccjm.org/content/82/11/745.

[47] Geoffrey Norman. “Dual processing and diagnostic errors”. In: *Advances in Health Sciences Education* 14.Suppl 1 (2009), pp. 37–49.

[48] Long Ouyang et al. “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022. URL: https://arxiv.org/abs/2203.02155.

[49] Mark C. Peterson et al. “Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses”. In: *Western Journal of Medicine* 156.2 (1992), pp. 163–165.

[50] Lewis Potter and Chris Jefferies. “Enhancing communication and clinical reasoning in medical education: Building virtual patients with generative AI”. In: *Future Healthcare Journal* 11 (2024). Abstracts from Medicine 2024: The future of medicine. RCP annual conference, p. 100043. ISSN: 2514-6645. DOI: https://doi.org/10.1016/j.fhj.2024.100043. URL: https://www.sciencedirect.com/science/article/pii/S2514664524001553.

[51] Carl Preiksaitis and Christian Rose. “Opportunities, Challenges, and Future Directions of Generative Artificial Intelligence in Medical Education: Scoping Review”. In: *JMIR Med Educ* 9 (Oct. 2023), e48785. ISSN: 2369-3762. DOI: 10.2196/48785. URL: http://www.ncbi.nlm.nih.gov/pubmed/37862079.

[52] Rafael Rafailov et al. “Direct Preference Optimization: Your Language Model is Secretly a Reward Model”. In: *NeurIPS*. 2023.

[53] S. Afzal, T.I. Dhamecha, et al. “AI Medical School Tutor: Modelling and Implementation”. English. In: *Lect. Notes Comput. Sci.* Ed. by Michalowski M. and Moskovitch R. Vol. 12299 LNAI. Journal Abbreviation: Lect. Notes Comput. Sci. Springer Science and Business Media Deutschland GmbH, 2020, pp. 133–145. ISBN: 03029743 (ISSN); 978-303059136-6 (ISBN). DOI: 10.1007/978-3-030-59137-3_13. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092248316&doi=10.1007%2f978-3-030-59137-3%5C%5F13&partnerID=40&md5=e7b49e1ad58349a5a2e5b316e0e48cb1.

[54] A Hasan Sapci and H Aylin Sapci. “Artificial Intelligence Education and Tools for Medical and Health Informatics Students: Systematic Review”. In: *JMIR Med Educ* 6.1 (June 2020), e19285. ISSN: 2369-3762. DOI: 10.2196/19285. URL: http://www.ncbi.nlm.nih.gov/pubmed/32602844.

[55] Timo Schick et al. “Toolformer: Language Models Can Teach Themselves to Use Tools”. In: *arXiv:2302.04761* (2023).

[56] Henk G. Schmidt and Remy M. J. P. Rikers. “How expertise develops in medicine: knowledge encapsulation and illness script formation”. In: *Medical Education* 41.12 (2007), pp. 1133–1139.

[57] Shefaly Shorey, Emily Ang, et al. “A Virtual Counseling Application Using Artificial Intelligence for Communication Skills Training in Nursing Education: Development Study”. In: *Journal of Medical Internet Research* 21.10 (Oct. 2019). Section: 0. ISSN: 1438-8871. DOI: 10.2196/14658. URL: https://www.sciencedirect.com/science/article/pii/S1438887119005284.

[58] Tjorven Stamer, Jost Steinhäuser, et al. “Artificial Intelligence Supporting the Training of Communication Skills in the Education of Health Care Professions: Scoping Review”. In: *J Med Internet Res* 25 (June 2023), e43311. ISSN: 1438-8871. DOI: 10.2196/43311. URL: http://www.ncbi.nlm.nih.gov/pubmed/37335593.

[59] Amos Tversky and Daniel Kahneman. “Judgment under Uncertainty: Heuristics and Biases”. In: *Science* 185.4157 (1974), pp. 1124–1131.

[60] Ashish Vaswani et al. “Attention Is All You Need”. In: *NeurIPS*. 2017.

[61] Hanna von Gerich, Hans Moen, et al. “Artificial Intelligence -based technologies in nursing: A scoping literature review of the evidence”. In: *International Journal of Nursing Studies* 127 (2022), p. 104153. ISSN: 0020-7489. DOI: https://doi.org/10.1016/j.ijnurstu.2021.104153. URL: https://www.sciencedirect.com/science/article/pii/S0020748921002984.

[62] C. Wang, L. Shuhan, et al. “Application of Large Language Models in Medical Training Evaluation—Using ChatGPT as a Standardized Patient: Multimetric Assessment”. English. In: *Journal of Medical Internet Research* 27 (2025). Section: 0. ISSN: 14388871 (ISSN). DOI: 10.2196/59435. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85213948312&doi=10.2196%2f59435&partnerID=40&md5=5b01dab688460ff339e327466668d16e.

[63] Heng Wang, Danni Zheng, et al. “Artificial Intelligence–Powered Training Database for Clinical Thinking: App Development Study”. In: *JMIR Formative Research* 9 (Jan. 2025). Section: 0. ISSN: 2561-326X. DOI: 10.2196/58426. URL: https://www.sciencedirect.com/science/article/pii/S2561326X25000095.

[64] Mengying Wang, Zhen Sun, et al. “Intelligent virtual case learning system based on real medical records and natural language processing”. In: *BMC Medical Informatics and Decision Making* 22.1 (Mar. 2022), p. 60. ISSN: 1472-6947. DOI: 10.1186/s12911-022-01797-7. URL: https://doi.org/10.1186/s12911-022-01797-7.

[65] Xuezhi Wang et al. “Self-Consistency Improves Chain of Thought Reasoning”. In: *ICLR*. 2023.

[66] Jason Wei et al. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”. In: *arXiv preprint arXiv:2201.11903* (2022). URL: https://arxiv.org/abs/2201.11903.

[67] Shunyu Yao et al. “ReAct: Synergizing Reasoning and Acting in Language Models”. In: *NeurIPS*. 2023.

[68] Chenwei Zhang et al. “Joint Slot Filling and Intent Detection via Capsule Neural Networks”. In: *arXiv preprint arXiv:2101.08091* (2021). URL: https://arxiv.org/abs/2101.08091.

Appendix A — System Usability Scale (SUS)

Instructions: For each statement, rate from 1 (*Strongly Disagree*) to 5 (*Strongly Agree*). Items alternate in polarity.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

Scoring: For odd items, subtract 1 from the user response; for even items, subtract the response from 5. Sum the adjusted values and multiply by 2.5 (range 0–100).

Appendix B — NASA-TLX (Adapted)

Instructions: Please rate each dimension from 0 (*Very Low*) to 10 (*Very High*) for the recent SmartDoc session.

- **Mental Demand:** How mentally demanding was the task?
- **Temporal Demand:** How hurried or time-pressured was the task?
- **Effort:** How hard did you have to work to achieve your performance?
- **Frustration:** How insecure, discouraged, irritated, or stressed were you?
- **Performance:** How successful were you in accomplishing what you set out to do? (Inverted so that higher indicates worse perceived performance for workload scoring.)

Note: In this study, unweighted subscales (0–10) are reported with median and IQR summaries.

Appendix C — Representative Session (S6)

This appendix presents a representative SmartDoc session (S6) from the Mull case. It includes: a compact dialogue transcript with intent tags and brief justifications; the submitted diagnosis; the learner's reflective responses; and the automated LLM evaluation summary.

C.1 Dialogue Transcript (complete with intents)

For each turn: *Intent* refers to the SmartDoc taxonomy (in `type_subtype` form); *Confidence* shows the LLM's classification confidence; *Justification* explains the intent mapping.

1. **Doctor:** "First, what is her past medical history?"
   Intent: `pmh_general` (confidence: 0.95).
   Justification: "The doctor is asking about the patient's past medical history, which falls under the definition of 'pmh_general' — questions about diseases, conditions, and diagnoses, and does not include medications."

2. **Patient:** "Uh, she has a history of morbid obesity, diabetes, hypertension, and rheumatoid arthritis."

3. **Doctor:** "And what brings you here today?"
   Intent: `hpi_chief_complaint` (confidence: 0.95).
   Justification: "The doctor is asking what brings the patient to the clinic, which directly addresses the chief complaint (hpi_chief_complaint) within the anamnesis phase."

4. **Patient:** "Uh, she's been having trouble breathing, especially when she's trying to do things. And she has a cough, but it's not bringing up anything."

5. **Doctor:** "For how long is she having trouble breathing?"
   Intent: `hpi_onset_duration_primary` (confidence: 0.95).
   Justification: "The question asks about the duration of trouble breathing, directly addressing the onset and duration of symptoms, a key component of the history taking process."

6. **Patient:** “Uh, it’s been gradually worsening for about two months, you know? It hasn’t gotten suddenly worse, just slowly.”

7. **Doctor:** “Any regular medication?”
   Intent: `meds_current_known` (confidence: 0.95).
   Justification: “The question is asking about current medications, explicitly using the word ‘medication’. This falls directly into the definition of ‘meds_current_known’.”

8. **Patient:** “Uh, she’s currently taking lisinopril, atenolol, glipizide, and metformin, I think. I’m not sure about anything else she might be on.”

9. **Doctor:** “And is she taking anything for her rheumatoid arthritis?”
   Intent: `meds_ra_specific_initial_query` (confidence: 0.98).
   Justification: “The question explicitly asks about medications for rheumatoid arthritis, indicating a focus on medication history related to this specific condition.”

10. **Patient:** “Uh, I’m not sure I have information about that specifically. I don’t have records of what she’s taking for her rheumatoid arthritis, I’m sorry.”

11. **Doctor:** “Does she have fevers?”
    Intent: `hpi_fever` (confidence: 0.95).
    Justification: “The question directly asks about fever, which falls under the category of inquiring about symptoms related to the patient’s current complaint during the history taking phase.”

12. **Patient:** “Uh, yes, she says she doesn’t have fevers, Doctor.”

13. **Doctor:** “What about night chills?”
    Intent: `hpi_chills` (confidence: 0.95).
    Justification: “The question ‘What about night chills?’ directly relates to a symptom (chills) and its temporal relationship (night), fitting the HPI intent for investigating the onset and characteristics of a presenting complaint.”

14. **Patient:** “Uh, she says she doesn’t have night chills, Doctor. She denies them.”

15. **Doctor:** “Any pain? On the chest?”
    Intent: `hpi_chest_pain` (confidence: 0.95).
    Justification: “The patient is asking about pain on the chest, specifically chest pain, which falls under the category of questions regarding chest pain during the history taking phase.”

16. **Patient:** “Uh, she says no, she doesn’t have chest pain, doctor.”

17. **Doctor:** “Vital signs.”
    Intent: `exam_vital` (confidence: 0.95).
    Justification: “The doctor is explicitly requesting vital signs, which falls directly under the ‘exam_vital’ intent.”

18. **Patient:** "Vital signs: Temperature 99.9°F (37.7°C), HR 105, BP 140/70, RR 24, O_2 sat 89% on room air."

19. **Doctor:** "Cardiac sound."
    Intent: `exam_cardiovascular` (confidence: 0.98).
    Justification: "The doctor is requesting an examination of the heart, specifically referring to 'cardiac sound,' which directly relates to a cardiovascular examination."

20. **Patient:** "Heart sounds are normal and there is no lower-extremity edema."

21. **Doctor:** "Lung sounds."
    Intent: `exam_respiratory` (confidence: 0.98).
    Justification: "The doctor is requesting an examination of the lungs, which directly relates to a respiratory exam."

22. **Patient:** "Pulmonary examination demonstrates crackles in all lung fields."

23. **Doctor:** "Is there any chest X-ray?"
    Intent: `imaging_chest_xray` (confidence: 0.95).
    Justification: "The doctor is asking about a chest X-ray, which falls under the category of imaging studies."

24. **Patient:** "Yes, a preliminary chest radiograph was obtained. The interpretation is consistent with pulmonary vascular congestion."

25. **Doctor:** "Are you sure that your mother is not on any biologics or immunosuppressive medications?"
    Intent: `meds_ra_specific_initial_query` (confidence: 0.95).
    Justification: "The question specifically asks about biologics or immunosuppressive medications, which falls under the category of RA medications and seeks initial clarification regarding current medication use for rheumatoid arthritis."

26. **Patient:** "Like I said, I'm not sure about her rheumatoid arthritis medications. Maybe you could check her previous hospital records? I know she's had some treatments at other facilities."

27. **Doctor:** "Can you get her complete medication list from previous hospitalizations?"
    Intent: `meds_full_reconciliation_query` (confidence: 0.98).
    Justification: "The doctor is explicitly requesting a 'complete medication list from previous hospitalizations,' which directly relates to medication reconciliation and retrieving information from medical records — a core component of the anamnesis phase."

28. **Patient:** "Uh, yes, she was receiving infliximab for rheumatoid arthritis. That's the only medication listed from her previous hospital stays, as far as I know."

29. **Doctor:** "Can we have a computed tomography scan?"
    Intent: `imaging_ct_chest` (confidence: 0.95).
    Justification: "The doctor is requesting a computed tomography (CT) scan of the chest, which falls under the category of imaging studies."

30. **Patient:** "The chest X-ray demonstrates a reticular pattern with innumerable, tiny, 1- to 2-mm pulmonary nodules. We can order a computed tomography scan of the chest to further characterize these findings. Would you like me to proceed with that order?"

31. **Doctor:** "Bloodwork?"
    Intent: `labs_general` (confidence: 0.95).
    Justification: "The doctor is asking about bloodwork, which falls under general laboratory questions."

32. **Patient:** "The cardiac lab results show a pro-BNP level greater than the upper limit of normal. We also have an elevated white blood cell count of 13.0×10^9/L and a low hemoglobin level of 10 g/dL. I recommend further investigation into the etiology of the anemia and leukocytosis."

33. **Doctor:** "Echocardiogram?"
    Intent: `imaging_echo` (confidence: 0.95).
    Justification: "The doctor is asking about an echocardiogram, which is a specific imaging study."

34. **Patient:** "The echocardiogram shows a normal ejection fraction with no evidence of elevated filling pressures."

Summary. The learner systematically escalated information gathering: baseline PMH $\rightarrow$ targeted RA meds $\rightarrow$ objective signs (vitals, exams) $\rightarrow$ imaging (CXR then CT) $\rightarrow$ labs and echo. Critical pivot: medication reconciliation uncovered *infliximab*, a TNF-α inhibitor that significantly elevates TB reactivation risk. The learner also challenged the preliminary CXR interpretation by ordering a CT scan and echocardiogram to arbitrate between heart failure and alternative pulmonary diagnoses.

C.2 Final Diagnosis

Submitted diagnosis: *Miliary tuberculosis*
Session ID: `SESS_0W451OZEJ`

C.3 Reflective Responses (five prompts)

The following are the learner's complete metacognitive responses to the five structured reflection prompts.

1. **Question:** What is the single most compelling piece of evidence that supports your chosen diagnosis?

   Answer: "The chest CT showing tiny pulmonary nodules in a diffuse reticular pattern, in combination with the patient's recent infliximab therapy, is the most compelling evidence supporting miliary tuberculosis. This imaging pattern is classic for disseminated hematogenous spread of Mycobacterium tuberculosis, and TNF alpha inhibitor therapy significantly increases the risk of reactivation of latent tuberculosis."

2. **Question:** What is one piece of evidence that might argue against your diagnosis?

   Answer: "An elevated BNP and crackles on auscultation could suggest heart failure but these findings are nonspecific and can occur in many conditions involving lung parenchymal disease or inflammation. The absence of peripheral edema and the normal cardiac exam make a cardiac origin of her dyspnea unlikely."

3. **Question:** What else could this be? List at least two reasonable alternative diagnoses.

   Answer: "Acute heart failure exacerbation; Interstitial lung disease (possibly secondary to rheumatoid arthritis or drug toxicity)."

4. **Question:** For one of your alternative diagnoses, what specific information would help rule it in or out?

   Answer: "The echocardiogram evaluating left ventricular systolic and diastolic function, as well as filling pressures, would decisively clarify whether cardiac dysfunction explains the symptoms. In this case, the normal ejection fraction and absence of elevated filling pressures effectively rule out acute heart failure as the cause of dyspnea. Additionally, the lack of leg edema, normal heart sounds, subacute course further confirm that cardiac failure is not the culprit."

5. **Question:** Have you considered and ruled out any potential must-not-miss or life-threatening conditions?

   Answer: "I considered and systematically excluded acute heart failure, pulmonary embolism, and severe bacterial pneumonia. The chronic and progressive nature of the patient's symptoms, absence of pleuritic pain or focal infiltrates, and the imaging pattern all argue against these."

C.4 LLM Evaluation — Scores and Detailed Feedback

C.4.1 Scores

Overall Score: 81/100
Information Gathering: 75/100
Diagnostic Accuracy: 88/100
Cognitive Bias Awareness: 80/100

C.4.2 Detailed Analysis by Dimension

Information Gathering (75/100). "The student correctly identified the critical importance of the infliximab medication and the reticular pattern with nodules on the CT scan. The normal echocardiogram was also recognized. However, the initial CXR interpretation was correctly noted as potentially misleading. The student missed explicitly stating the weight loss history as a key finding, and while they acknowledged the formal CXR interpretation, it wasn't explicitly highlighted as a point of contention. Overall, the information gathering was reasonably thorough, but could have been more systematic and explicitly linked to the diagnostic challenge."

Diagnostic Accuracy (88/100). "The student arrived at the correct diagnosis of miliary tuberculosis, demonstrating a good understanding of the case. Crucially, they avoided the heart failure trap, recognizing the normal echocardiogram. The reasoning linking infliximab to the increased risk of TB reactivation was sound. The differential included a reasonable alternative diagnosis (acute heart failure exacerbation), though the justification for this was somewhat superficial. The score is high because the student correctly identified the key diagnostic clues and avoided a common anchoring bias."

Cognitive Bias Awareness (80/100). "The student demonstrated a good level of awareness regarding anchoring bias, explicitly addressing the potential for the initial CXR interpretation to mislead. They also acknowledged the importance of medication reconciliation, correctly identifying infliximab as a key factor. Their metacognitive responses were thoughtful, particularly in questioning the initial framing of the case as heart failure. However, the reflection could have been more deeply explored, perhaps with a more detailed discussion of how confirmation bias might have influenced the initial assessment. The student's responses were clear and clinically relevant."

C.4.3 Comprehensive Feedback

Strengths.

- Correct diagnosis of miliary tuberculosis
- Successful avoidance of heart failure trap (anchoring bias)
- Recognition of infliximab's role as a TNF-α inhibitor increasing TB reactivation risk
- Thoughtful metacognitive responses demonstrating critical evaluation of evidence
- Appropriate use of echocardiogram to arbitrate between cardiac and pulmonary etiologies

Areas for Improvement.

- More systematic documentation of key findings (e.g., weight loss history)
- Deeper exploration of anchoring bias mechanisms in reflection

- More detailed discussion of potential biases influencing diagnostic reasoning
- Explicit highlighting of discrepancies between preliminary and definitive test results

Key Recommendations.

1. **Implement a structured approach to medication reconciliation**, specifically focusing on immunosuppressant therapies and their potential impact on infectious disease risk. Develop a checklist to ensure all relevant medications are considered during the diagnostic process.
2. **Practice actively challenging initial interpretations**, particularly when they are influenced by readily available data (e.g., preliminary CXR findings). Establish a habit of seeking confirmatory or refuting evidence through definitive tests.
3. **Enhance systematic documentation** of clinical findings by creating a structured template that ensures all critical data points (including social history, weight changes, constitutional symptoms) are explicitly noted and linked to differential diagnoses.