MDPI

*Article*

# Reasoning Clinical Decision Support System with Large Language Model Agents: A Case Study in Pulmonary Embolism

**Antonio Giménez-López [1,2]*, Andrés Piñeiro-Martín [1,2], Iago Mosquera-Fajardo [3,2], Laura Docío-Fernández [1], Carmen García-Mateo [1]**

[1] GTM research group, atlanTTic Research Center, Universidade de Vigo, Vigo, Spain;
[2] Balidea Consulting & Programming S.L., Santiago de Compostela, Spain;
[3] ICU Department, University Hospital of Ferrol, Ferrol, Spain;
* Correspondence: antoniogl@Balidea.com

**Abstract:** This study presents a novel agent-based architecture that equips Clinical Decision Support Systems (CDSSs) with advanced reasoning capabilities through the integration of Large Language Model (LLM) agents and Retrieval-Augmented Generation (RAG). The system was designed to autonomously interpret official clinical guidelines on pulmonary embolism (PE), integrate structured patient data, and generate relevant outputs to support decision-making in clinical practice. A multi-agent CDSS prototype was developed to address two core functions: dynamic consultation of clinical guidelines and personalized clinical case evaluation. The system was evaluated using two custom-designed datasets: 45 realistic guidelines queries and 20 simulated patient cases. Performance was assessed through automated RAGAS metrics and expert clinical validation. The system demonstrated high performance across both functionalities, achieving 97.78% accuracy in solving guidelines queries and mean expert ratings above 95% for clinical accuracy, risk score computation, and interpretability in case evaluations. These results indicate that agentic CDSSs can reason effectively over clinical data and guidelines, producing structured outputs that align with clinical standards and expert expectations. The findings support the potential of this approach to enhance clinical workflows and decision-making, while underscoring the importance of further validation and collaborative integration with human clinicians.

**Keywords:** Clinical decision support systems; large language models; agents; artificial intelligence; e-health; healthcare

## 1. Introduction

In today's complex healthcare landscape, Clinical Decision Support Systems (CDSSs) have emerged as essential tools to assist clinicians and mitigate the burden of information overload they have to manage. CDSSs are specialized software tools designed to enhance clinical decision-making by integrating medical knowledge, patient-specific data, and computational algorithms to provide evidence-based recommendations. These systems act as a cognitive assistant for clinicians, offering insights that range from diagnostic suggestions to treatment optimization, thus improving the precision, robustness, and safety of medical practices [1]. Consequently, they significantly contribute to improving healthcare quality by diagnostic accuracy, minimizing human error, and promoting adherence to evidence-based clinical guidelines [2].

The main purpose of CDSSs is to bridge the gap between the large and rapidly expanding repository of medical knowledge and the clinical need for informed decisions at the point of care. By synthesizing data from various sources, such as Electronic Health

Received:
Revised:
Accepted:
Published:

**Citation:** . Title. *Appl. Sci.* **2025**, *1*, 0. https://doi.org/

Records (EHRs), clinical guidelines, and medical center procedures, these systems empower healthcare providers to make decisions aligned with the best available evidence [3,4]. Modern solutions transcend traditional decision support by incorporating patient-centered functionalities, including personalized health recommendations, educational resources, and interactive tools that facilitate communication between clinicians and patients. These capabilities aim to enhance patient engagement, improve understanding of treatment options, and foster shared decision-making [5,6].

Over the past decades, CDSSs have evolved significantly, transitioning from early rule-based systems, such as MYCIN [7] or HELP [8], which relied on expert-defined logic and deterministic rules, toward more flexible ML-driven solutions [9]. While the rule-based approach was the basis for demonstrating the feasibility of computerized decision making, their limitations in scalability, adaptability, and context understanding made them inadequate for modern clinical environments. The emergence of ML-based CDSSs enabled data-driven predictions, improving diagnostic performance in various domains [10,11], including cardiovascular or pulmonary disease risk assessment [12–14]. However, these models operate as "black boxes", limiting their adoption due to poor interpretability, and limited trust among clinicians [15,16].

More recently, the adoption of Large Language Models (LLMs) and Transformer-based architectures has redefined the potential of CDSSs by by enabling more effective handling of unstructured clinical data [17,18]. Their integration into Retrieval-Augmented Generation (RAG) pipelines has allowed CDSSs to combine structured patient data with unstructured medical guidelines, improving contextual relevance and reducing the risk of hallucinations, leading to more reliable personalized recommendations [19]. Even with this potential, these systems still face challenges, including their limited capacity for managing complex multi-step queries and the persistent risk of hallucinations when the outputs are not adequately grounded in verified knowledge sources [20,21].

Current efforts seek to refine these CDSSs by improving their ability to handle complex queries, enable user interaction and integrate seamlessly into clinical workflows, thereby meeting the dynamic demands of modern healthcare. However, the effective utilization of these tools demands human oversight to avoid over-reliance on algorithmic outputs, which can lead to serious harm to the patients. Despite advancements in Artificial Intelligence (AI), these systems are not immune to errors, biases, or lack of context. Clinicians must remain the ultimate decision-makers, interpreting and validating the recommendations provided by the system to ensure patient safety and high-quality care [4].

Beyond issues of safety and reliability, critical concerns also arise around data privacy, interpretability, and compliance with emerging regulations. These systems rely on sensitive clinical data, which must be handled in strict accordance with regulations such as HIPAA in the United States and GDPR in the European Union. Furthermore, the European *AI Act* introduces further requirements for high-risk AI systems, such as those used in healthcare, demanding transparency, robustness, and traceability of decision making processes [22]. Technical challenges also persist: hallucinations remain difficult to detect and mitigate, and the inherently opaque nature of large generative models continues to hinder explainability and trust [23]. Additionally, while advanced approaches like Retrieval-Augmented Generation (RAG) and agent-based architectures offer exciting potential, they are still in the early stages of research. Their clinical applicability is limited by a lack of empirical validation, standardized evaluation frameworks, and scalable design principles [9,10]. Addressing these issues is essential to foster trust among clinicians and patients, and to facilitate the safe and effective integration of AI-driven CDSSs into clinical workflows.

A promising yet underexplored paradigm is the use of LLMs as *reasoning engines*, enabling them with capabilities for dynamic planning, tool utilization, and autonomous

decision-making. When combined with access to dynamic and up-to-date medical knowledge, these systems can manage many of the limitations of earlier CDSS approaches [24]. Although initial applications indicate that such agentic CDSSs are capable of delivering personalized and context-aware recommendations [25], their deployment in real-world clinical settings remains constrained by the lack of standardized evaluation frameworks and comprehensive clinical validation.

To the best of our knowledge, no existing system integrates agent-based reasoning with guideline-grounded RAG pipelines in a validated CDSS. This work sought to demonstrate the feasibility, practical value, and future potential of integrating advanced AI methodologies, such as LLMs, RAG, and agent-based architectures into CDSSs.

To explore the potential of agent-based CDSSs, this work uses Pulmonary Embolism (PE) management as a representative and clinically challenging use case. PE is a life-threatening condition, affecting approximately 1 in 1000 individuals annually, caused by the obstruction of pulmonary arteries, most commonly due to emboli originating from deep vein thrombosis [26]. Despite significant advances in diagnostic and therapeutic approaches, PE continues to pose serious clinical challenges due to its nonspecific symptoms, rapid progression, and potential for acute complications such as hypoxia, cardiac arrest, and chronic thromboembolic pulmonary hypertension, in the long term.

Accurate diagnosis and effective management require timely consultation of extensive clinical guidelines and the use of validated risk assessment metrics [27]. These include the Pulmonary Embolism Severity Index (PESI) and its simplified version (sPESI), which estimate 30-day mortality, as well as the guideline-based early mortality risk. However, these metrics are often difficult to compute manually in time-constrained settings, and guideline recommendations must be adapted to patient-specific conditions, increasing the cognitive burden on physicians. This makes PE a particularly suitable scenario for the development of AI-powered CDSSs capable of integrating structured patient data, guideline knowledge, and risk metrics to support real-time clinical decision making.

The main contributions of this work are as follows:

1. The design of an innovative CDSSs architecture, based on LLM agents, that provides dynamic reasoning capabilities for interpreting medical knowledge and patient information, as well as autonomous decision-making to support physicians in daily clinical practice.
2. The integration of a RAG module into this architecture, enabling agents to ground their reasoning, actions and conclusions in validated knowledge extracted from official clinical guidelines and evidence-based standards.
3. The design and implementation of a comprehensive evaluation framework that combines automated performance metrics with expert-based clinical validation, assessing key dimensions such as clinical relevance, factual consistency, guideline adherence, interpretability, and clinical safety.
4. The development of a fully functional agent-based CDSS prototype, evaluated through the designed framework, demonstrating the potential, feasibility, and clinical relevance of this approach for supporting real-world medical decision-making.

In summary, this study offers preliminary evidence supporting the feasibility and potential benefits of integrating LLM agents to provide reasoning capabilities to CDSSs. It contributes with a structured architecture, an open evaluation methodology, and insights for future improvements in AI-assisted clinical workflows.

The rest of the document is organized as follows: Section 2 details the design and implementation of the agent-based CDSSs prototype; Section 3 explains how it will be evaluated; Section 4 details its results on the proposed framework; and finally, Section 5 summarizes key findings and addresses challenges and research directions of this approach.

## 2. Methods

To investigate the feasibility and clinical potential of autonomous CDSSs based on LLM agents, a prototype system was designed and implemented, specialized in the management of PE. The system acts as an intelligent assistant capable of integrating structured clinical data, interpreting complex medical guidelines, and providing real-time diagnostic and therapeutic recommendations.

This prototype combines LLM agents with a RAG architecture to address the complexities of clinical practice associated with PE. The integration of these technologies allows the system to reason over clinical information and autonomously plan and execute appropriate actions. Furthermore, by retrieving relevant content from clinical guidelines as needed, the system ensures that its decisions, actions, and responses remain firmly grounded in established medical knowledge.

The two core services provided by the system are:

1. **Guidelines Consultation Service**: This functionality allows medical practitioners to ask natural language questions directly related to the official clinical guidelines on the diagnosis, risk stratification, and treatment of PE. This enables rapid access to complex guideline documents without the need for manual navigation, thus facilitating informed decision-making even in time-constrained environments.
2. **Clinical Case Evaluation Service**: This component enables the system to process structured clinical data from a designated patient, calculate risk stratification metrics and provide personalized recommendations for diagnosis or treatment. This assists clinicians in validating clinical pathways, and optimizing care decisions.

The overall architecture is designed to support autonomous reasoning by allowing the LLM agent to plan its actions, retrieve relevant evidence, ask for missing patient clinical information, and generate well-grounded, context-aware responses. By specifically focusing on a specific but clinically significant condition, such as PE, the system can be evaluated in a controlled yet realistic setting, capturing the complexities and uncertainties inherent in real-world clinical scenarios. This allows for a realistic and focused assessment of the reasoning capabilities and reliability of LLM-based autonomous agents in healthcare.

Figure 1 illustrates the high-level architecture of the proposed CDSS, designed to assist physicians in the management of PE cases. Clinicians can access the system via a conversational interface, submitting natural language queries through the virtual assistant. After receiving a request, this virtual assistant forwards the query to the multi-agent system (blue in Figure 1). Within this module, the designed workflow analyzes the user request to determine its nature, classifying it into one of the services provided by the system.

Once a user query is classified, the corresponding service is activated, orchestrating the necessary tasks to process and resolve the query. To achieve this, the system can retrieve stored clinical information from the patient, access available clinical metrics, and use the RAG module (purple in Figure 1), which enables services to retrieve specific information from the PE official clinical guidelines from the ESC [27]. When required clinical information is missing, the assistant prompts the clinician for clarification. If certain values remain unavailable, such as pending lab results, the system transparently acknowledges the limitation and adapts its reasoning accordingly, without making assumptions that could compromise clinical safety. Finally, once the final response of a service is generated, it is passed through the Hallucination Detector (green in Figure 1). This module verifies that each sentence in the response is grounded in the retrieved evidence from the guidelines; identifying and flagging any unsupported or misleading statements before delivering the final output to the clinician.

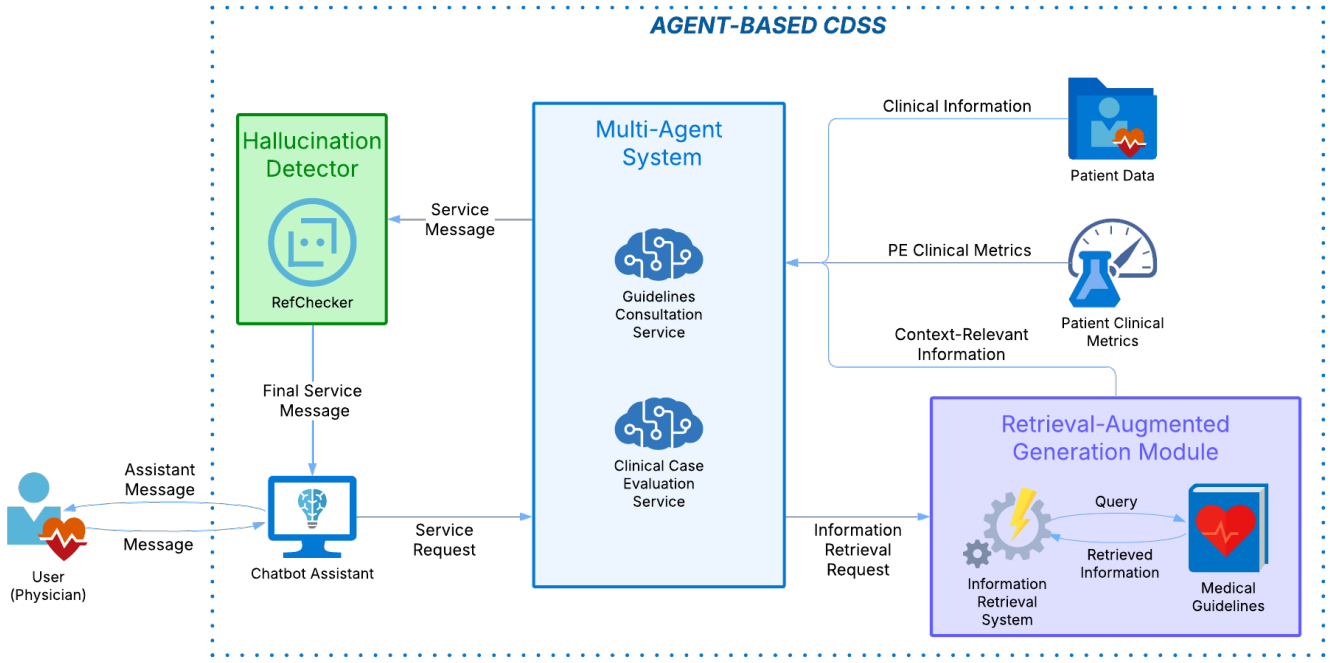


**Figure 1.** High-level architecture of the CDSS prototype. Physician queries are submitted through a chatbot assistant, which interacts with a system of LLM agents (blue) to answer questions based on PE clinical guidelines or generate patient-specific clinical recommendations. These agents can access available patient data and PE metrics, and use the RAG module (purple) to retrieve relevant information from PE guidelines to accomplish their tasks. The response messages from these services are analyzed by the Hallucination Detector (green) before being finally displayed to the user.

Within each service, multiple **agents** operate in a structured and collaborative manner, each performing a specific task to ensure the successful execution of the main functionality. By coordinating their outputs, these agents collectively achieve their final goal. All internal agents are powered by a **LLM**, which serves as the core engine for reasoning and response generation across services. These agents leverage the RAG module to improve their decision-making process with retrieved medical knowledge. This integration allows the system to understand and apply the guidelines for PE management.

The prototype also incorporates short-term memory throughout the entire interaction with the clinician, preserving relevant information such as user queries, intermediate decisions, and patient-specific data. This memory enables coherent and self-consistent reasoning across multiple steps, allowing the system to refine its outputs as new information becomes available or as the clinical context evolves.

The final design of the prototype allows clinicians to select a specific patient and retrieve their clinical data from a structured database. It is capable of answering a wide range of queries related to the PE clinical guidelines, from simple factual questions with directly extractable answers to more complex queries that require handling exceptions, applying clinical criteria, and reasoning over the content of the guidelines. In addition, the system can analyze structured patient cases by automatically calculating risk stratification scores (PESI, sPESI, and risk of early mortality) based on the available clinical data. Using this information, it provides physicians with diagnostic and therapeutic recommendations adapted to the patient's clinical status and aligned with guideline-based decision logic.

### 2.1. Multi-Agent System

This section aims to provide a deeper understanding of how the multi-agent system (blue rectangle in Figure 1) has been designed and how all the agent cooperate to deliver the described services to system's users.

The multi-agent system has been developed using the LangGraph [28] framework, which provides a set of specialized libraries that enable the implementation of graph-based agent workflows, efficient inter-agent communication, and dynamic task coordination. Through LangGraph, the system is structured as a graph-based multi-agent framework, where each node is executed in a sequential and organized manner, ensuring that each agent fulfills its designated role effectively.

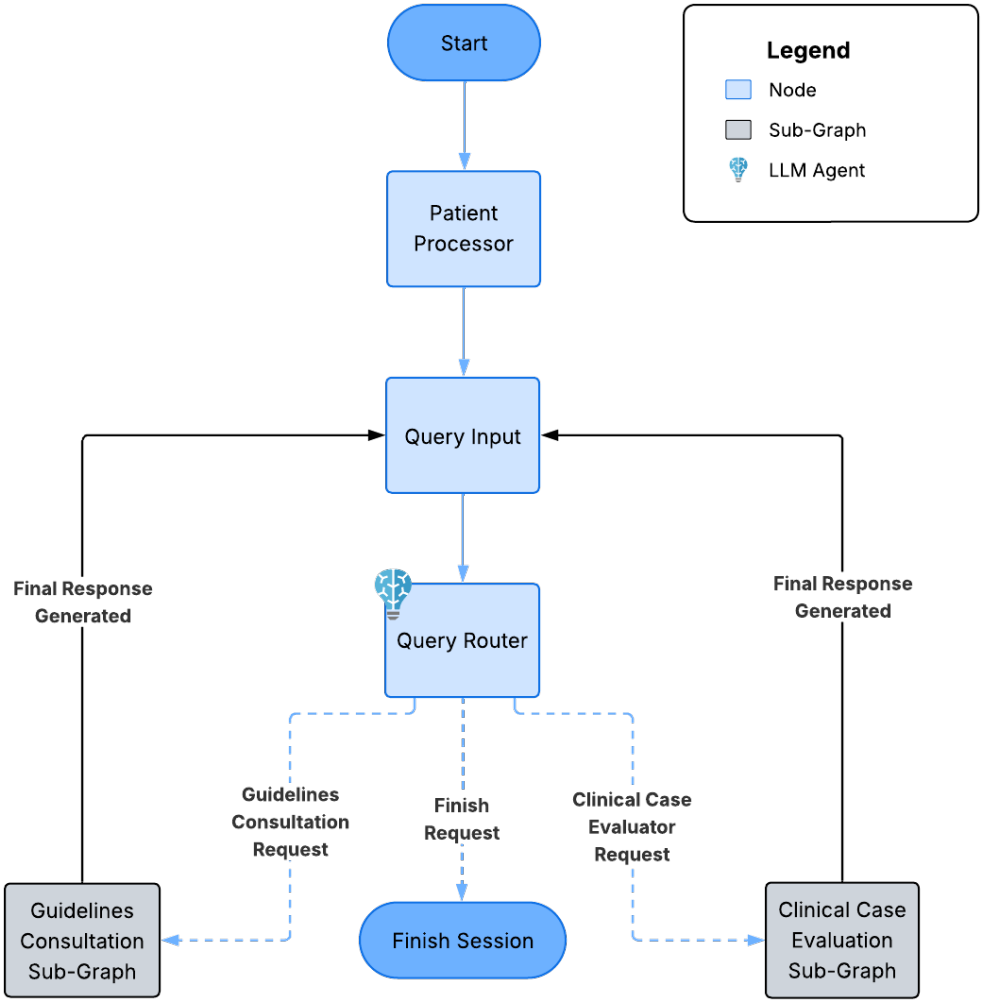


**Figure 2.** General workflow of the Agent-Based CDSS. Following patient selection, the system captures the clinician's query, routes it to the appropriate subgraph and returns the generated final response. Then, the system returns to the input state, allowing follow-up interactions until the clinician ends the session.

The graph-based architecture of the system is composed of a set of interconnected nodes, each responsible for executing a specific function that contributes to the overall objective of the CDSS. Depending on the nature of the task, nodes may differ in complexity and required capabilities.

Those nodes that involve, advanced interpretability or intelligent and dynamic behavior are implemented as LLM agents. These agents are equipped with three key capabilities: (i) sophisticated reasoning capabilities via direct access to PE guideline content; (ii) contextual adaptability driven by the current system state and the clinician's query; and (iii) the ability to invoke specialized *tools* to perform required actions.

These tools are callable functions or modules that allow the agent to perform concrete actions, thereby overcoming one of the inherent limitations of LLMs: their inability to act on the environment by themselves.

By combining internal reasoning with tool invocation, each LLM-based node is capable of autonomously completing its assigned task and generating a meaningful, context-aware response that contributes to resolving the main task.

All implemented agents have been designed to explain their decisions and actions in natural language, improving the interpretability of the system. Moreover, each agent has been instructed to reason through its process before reaching a conclusion, following the chain-of-thought prompting strategy [29]. This approach ensures that decisions are based on step-by-step logical reasoning rather than direct predictions.

The graph representing the main workflow is shown in Figure 2. The pipeline commences with three main nodes that constitute the basis of the CDSS. These nodes work sequentially to initialize the system, manage user interactions, and determine the appropriate workflow based on the user's query.

First, the *Patient Processor Node* is activated upon the initialization of the virtual assistant. Here, the system asks the user to select a patient by entering the patient's unique ID. Once the ID is provided, it retrieves the corresponding patient data from the clinical case database and loads it into memory for further processing. A detailed description of the database, including the patient data it contains, will be provided in ***Section 3.1.2***. Consider that to prevent overloads, the interaction is limited to a single patient per conversation.

Subsequently, the pipeline continues to the *Query Input Node*, at which point the user is prompted to enter a query. Pending the submission of the request, the prototype enters a state of standby. Upon receipt of the query, it is stored in the system's memory and forwarded to the next nodes of the pipeline.

Finally, the processed query reaches the *Query Router Node*, where an LLM agent analyzes it to determine which of the system's services is best suited to handle the request. In this way, the query is classified as either a guidelines consultation or a clinical case analysis request. Based on this classification, the corresponding subgraph is activated to resolve the user's query. The objective of this node is to ensure correct routing of requests, facilitating an appropriate response from the system.

After the execution of any sub-graph, the workflow resets to the Query Processor, where the system awaits the next input from the user. This design ensures that the system remains ready to handle subsequent queries efficiently, maintaining flexibility and responsiveness in supporting clinical tasks.

The following subsections describe the two designed sub-graphs to provide the described services, highlighting their function in guiding the system's decision-making and response workflow.

#### 2.1.1. Guidelines Consultation Sub-Graph

The sub-graph workflow is illustrated in Figure 3. It is responsible for handling the *Guidelines Consultation Service* described in Section 2, whose mission is to provide clinicians with accurate answers to questions related to the diagnosis, risk stratification and treatment of PE, as described in the official clinical guidelines.

To achieve this, the sub-graph is composed of a specialized LLM agent responsible for analyzing the clinician's question, determining what information is needed from the clinical guidelines, and retrieving that information. This agent autonomously sends requests to the *Dynamic Retrieval Tool* in order to obtain the most relevant passages of information which are then used as contextual input to support the reasoning process.

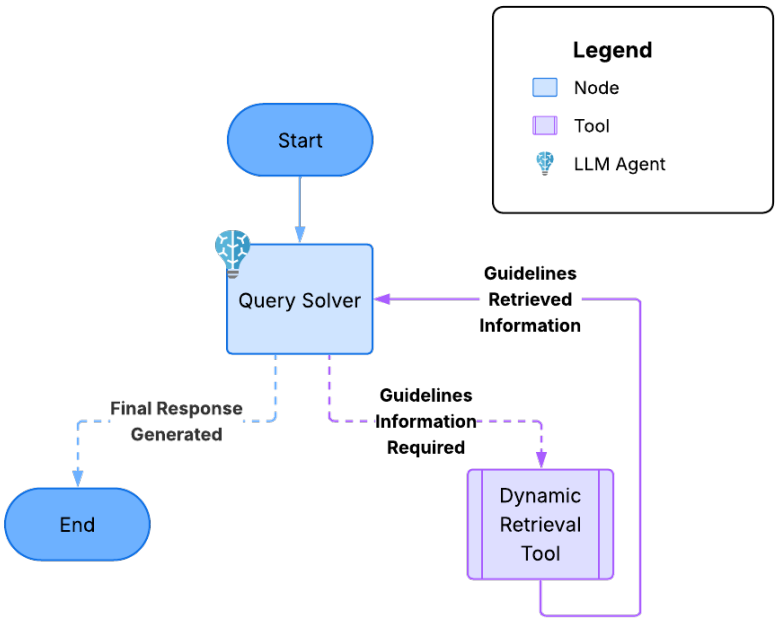


**Figure 3.** Internal workflow of the *Guidelines Consultation Sub-graph*. The *Query Solver Node* analyzes the user's question and autonomously identifies the information required from the clinical guidelines. It can perform one or more adapted queries to the *Dynamic Retrieval Tool*, a RAG-based component that returns semantically relevant passages. The retrieved evidence is then used by the node to iteratively refine its reasoning and generate the final response.

The agent operates dynamically, issuing as many retrieval calls as necessary and refining its information needs based on the evolving context of the query. Once sufficient contextual evidence has been gathered, the node generates a detailed response, explicitly explaining its reasoning process. The final answer is grounded in the retrieved guidelines content, which is used as contextual input to support the agent's generation process.

Using the *Dynamic Retrieval Tool*, the agent can perform natural language searches over the guidelines and specify how many relevant passages should be retrieved. When the tool receives a request, the underlying RAG module performs a search over the vector database and returns the top-ranked fragments that best match the agent's query. This design allows the agent to iteratively refine its information needs across multiple retrieval steps, adapting its strategy based on previously retrieved content. To prevent infinite loops or excessive querying, the number of retrieval calls allowed per query has been capped at three.

The objective of this sub-graph was to explore the capabilities of LLM agents in assisting clinicians with navigating complex medical guidelines. By providing a simple interface through which medical practitioners can submit their questions in natural language, the service aims to facilitate quick and structured access to relevant and complex information.

### 2.1.2. Clinical Case Evaluation Sub-Graph

The sub-graph workflow is shown in Figure 4. It is responsible for handling the *Clinical Case Evaluation Service* outlined in Section 2, whose purpose is to generate diagnostic and therapeutic recommendations based on a structured patient case. This process requires

the prototype to integrate multiple sources of information, including patient clinical data, clinical risk metrics, and medical knowledge, to deliver a coherent and context-aware response with the evaluation of the clinical case.

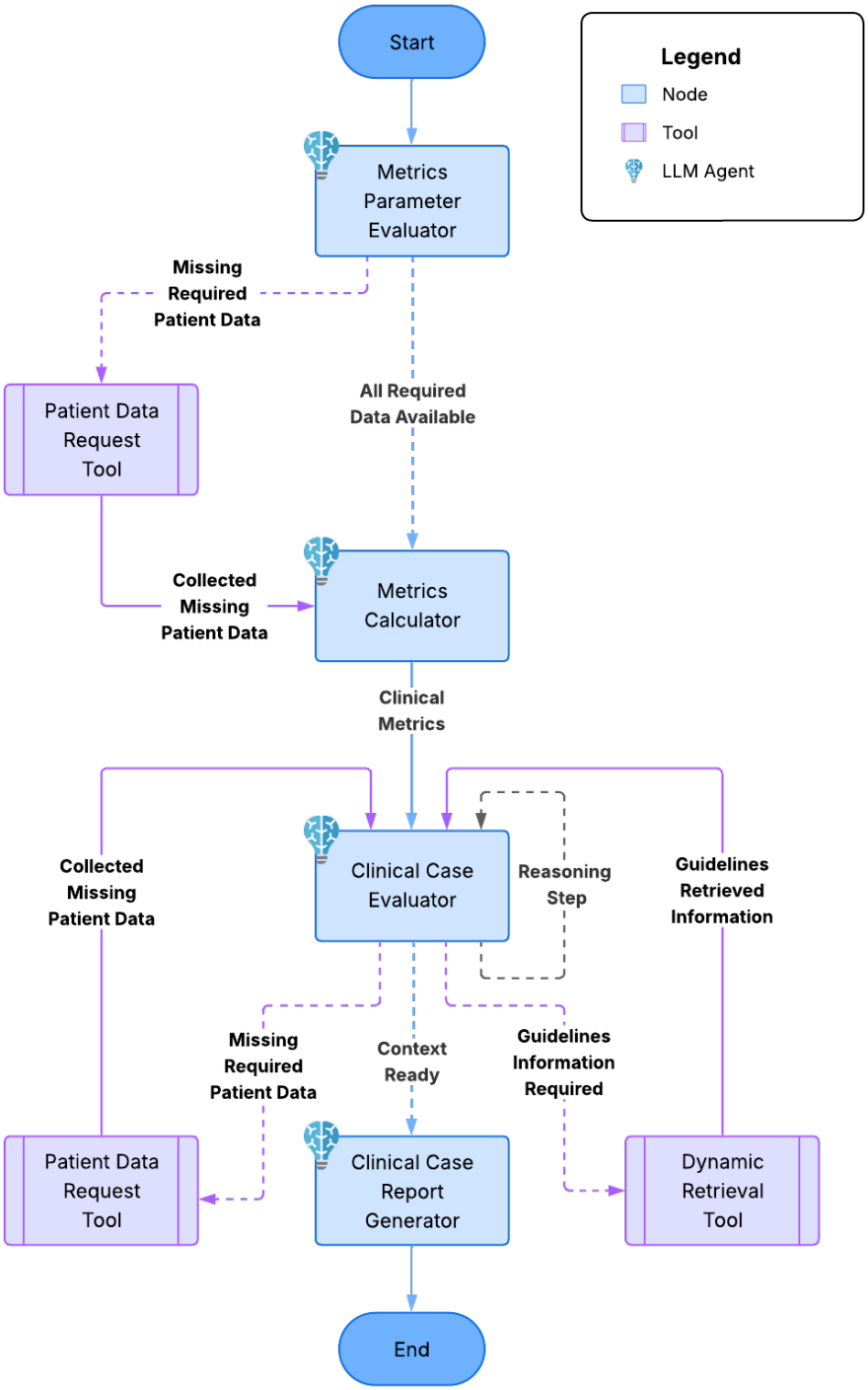


**Figure 4.** Internal workflow of the Clinical Case Evaluation Sub-Graph. First, the *Metrics Parameter Evaluator Node* identifies the necessary patient data for calculating the clinical metrics and validates whether all the required data is available. The *Metrics Calculator Node* applies the appropriate scoring criteria to obtain the PESI, sPESI and risk of early mortality. The *Clinical Case Evaluator Node* then iteratively reasons over the case, invoking tools to retrieve additional patient data or relevant guideline information as needed. Once sufficient context is available, the *Clinical Case Report Generator Node* produces a structured summary including the clinical assessment and final recommendations.

The first stage of the subgraph is responsible for computing the clinical metrics required to evaluate the patient's condition. This stage is composed of two sequential nodes: the *Metrics Parameter Evaluator* and the *Metrics Calculator*, both implemented as LLM agents. Their combined objective is to ensure that all necessary input data is available and to compute risk stratification metrics (PESI, sPESI, and Risk of Early Mortality), which are essential for the clinic case evaluation.

The process begins with the activation of the *Metrics Parameter Evaluator Node*, an agent that first analyzes the relevant sections of the clinical guidelines to determine which patient variables are required for computing each clinical score. Once this information has been extracted, the agent inspects the patient's stored data to verify whether all necessary variables are available. If any necessary information is missing, the agent invokes the *Patient Data Request Tool*, prompting the physician to manually provide the missing values if available. Next, the workflow proceeds to the *Metrics Calculator Node*, where an agent applies the corresponding computational logic to the patient data to compute the clinical scores. As in the previous node, the criteria for each score are provided as contextual input retrieved via the RAG module, allowing the agent to interpret and apply the correct decision rules. The resulting metrics are then stored in memory for subsequent nodes.

The metrics computation logic has been designed to operate even in scenarios where some patient data is missing. In such cases, the system adopts a cautious approach, avoiding the risk of underestimating severity in the presence of incomplete information. When uncertainty arises due to unavailable data, the prototype conservatively opts to classify the patient into the highest plausible risk category, based on the data at hand.

Although clinical metrics are derived from structured data, this study intentionally explored whether agents could effectively handle this task instead of relying on rule-based logic implemented through traditional code. While hardcoding the decision criteria for each score is feasible, it quickly becomes impractical when scaling to multiple conditions or incorporating frequent updates to clinical guidelines. The use of LLM agents in this context is motivated by the hypothesis that such models can autonomously interpret the necessary criteria from clinical guidelines, apply them correctly, and explain their reasoning. This approach lays the foundation for a more scalable and flexible architecture, where expanding the system to cover additional diseases or incorporate new clinical scores simply requires the ingestion of new guideline documents, without extensive manual reprogramming.

Once the clinical scores have been computed and stored in memory, the workflow proceeds to the clinical evaluation of the case and the reporting phase. This second part of the sub-graph is composed of two sequential agents: the *Clinical Case Evaluator Node* and the *Clinical Case Report Generator Node*.

The *Clinical Case Evaluator Node* implements an agent responsible for autonomously generating the complete evaluation of the patient's case. This agent operates through a self-iterative reasoning loop in which it progressively builds the necessary clinical context to generate accurate recommendations. At each step of the loop, the agent receives the current state of the reasoning process, consisting of retrieved information, patient data, and previous reflections, and decides either to invoke a tool or to continue reasoning internally.

The reasoning process begins with an initial analysis of the available patient data and the previously computed clinical metrics. Based on this initial assessment, the agent determines whether additional information from the PE clinical guidelines is required to support its recommendation. When necessary, it invokes the *Dynamic Retrieval Tool*, issuing a natural language query to retrieve the most relevant fragments from the vectorized guideline corpus. In parallel, if the agent detects that specific patient variables are missing but potentially available, it uses the *Patient Data Request Tool* to ask the clinician for the missing data. This self-reflective reasoning loop continues until the agent concludes that it

has acquired sufficient contextual knowledge to complete the case evaluation. To ensure computational efficiency and prevent infinite loops, the number of calls to the *Dynamic Retrieval Tool* is limited to a maximum of three per. Once the agent determines that the reasoning process is complete, the workflow proceeds to the following node, where the final output is generated and presented to the clinician.

This node has also been specifically designed with patient safety as a primary concern. During its reasoning process, the agent must identify any clinical conditions that may represent contraindications to certain diagnostic procedures or therapeutic interventions. When such risks are detected, the agent is expected to adapt its recommendations accordingly, avoiding unsafe suggestions and proposing alternative strategies when necessary. In addition, the system is designed to provide detailed and actionable outputs: beyond general advice, the agent is encouraged to specify concrete diagnostic techniques or treatment options, and to include medication types and dosages when clinically appropriate.

The final stage of the subgraph is handled by the *Clinical Case Report Generator Node*, an agent that is responsible for composing the final report of the evaluation of the clinical case that is presented to the clinician. The generated report consists of four main sections: (i) a summary of the most relevant patient data for the case at hand, (ii) a diagnostic report based on the current clinical evidence, (iii) a set of recommended actions adapted to the patient's condition, which may include further diagnostic tests or therapeutic interventions, and (iv) a final summary that highlights the key insights and priorities. All recommendations are accompanied by an explicit justification, including references to the specific sections of the clinical guidelines from which they are derived, along with the corresponding level of evidence when available.

The objective of this subgraph, within the scope of this study, is to analyze the potential of agents in supporting clinical decision-making in real-world healthcare settings. The system has been designed to generate structured evaluation reports for individual clinical cases by leveraging available patient data and evidence drawn from official clinical guidelines. The aim is not to replace physicians, but to assist them by providing high-quality, recommendations that may enhance decision-making efficiency and consistency. By examining the effectiveness of this approach, the study seeks to assess whether such systems can offer meaningful support to healthcare professionals in their daily practice.

### *2.2. Hallucination Detector*

In order to address the challenge of hallucinations in generative AI systems, a hallucination detection module has been integrated into the architecture to evaluate the responses generated by the agents before they are delivered to physicians (green rectangle in Figure 1). In the context of agent-based CDSSs, hallucinations refer to outputs that are not supported by the medical evidence of the clinical guidelines. This type of error poses a critical risk in healthcare applications, where the reliability of each response must be verifiable and traceable to validated clinical sources. The hallucination detection module is based on the *RefChecker* framework [30], which decomposes each response into structured claim triples and compares them against the retrieved guidelines context. Each statement is then classified as "Entailment", "Contradiction", or "Neutral". This verification step enables the system to highlight potentially hallucinated content prior to presenting the final response to the clinician. Importantly, this component does not alter the original response. Instead, it flags unsupported statements and displays a visible alert, allowing the clinician to review the content and decide whether the highlighted fragment constitutes an actual hallucination.

The inclusion of this module is intended to investigate whether LLM-based agents are more prone to hallucinations when confronted with complex clinical cases that require reasoning over large context windows, compared to other application domains.

### 2.3. Retrieval-Augmented Generation Module

The RAG module (purple rectangle in Figure 1) was developed to enable both the agents and the *Dynamic Retrieval Tool* to dynamically retrieve and integrate relevant information from clinical guidelines as contextual input for response generation. This component is essential for grounding the agent's reasoning in up-to-date medical knowledge, ensuring that responses are not solely based on the model's internal parameters but also on validated external sources of information.

This module has been implemented using the LangChain framework [31], which offers high-level abstractions for building retrieval pipelines and managing semantic search operations. The architecture of the module comprises two main stages: (i) ingestion of the clinical guidelines into a vector database, and (ii) retrieval of relevant passages in response to agent queries. Figure 5 illustrates the overall workflow of the module, where the ingestion process is depicted in red and the retrieval pipeline in green.

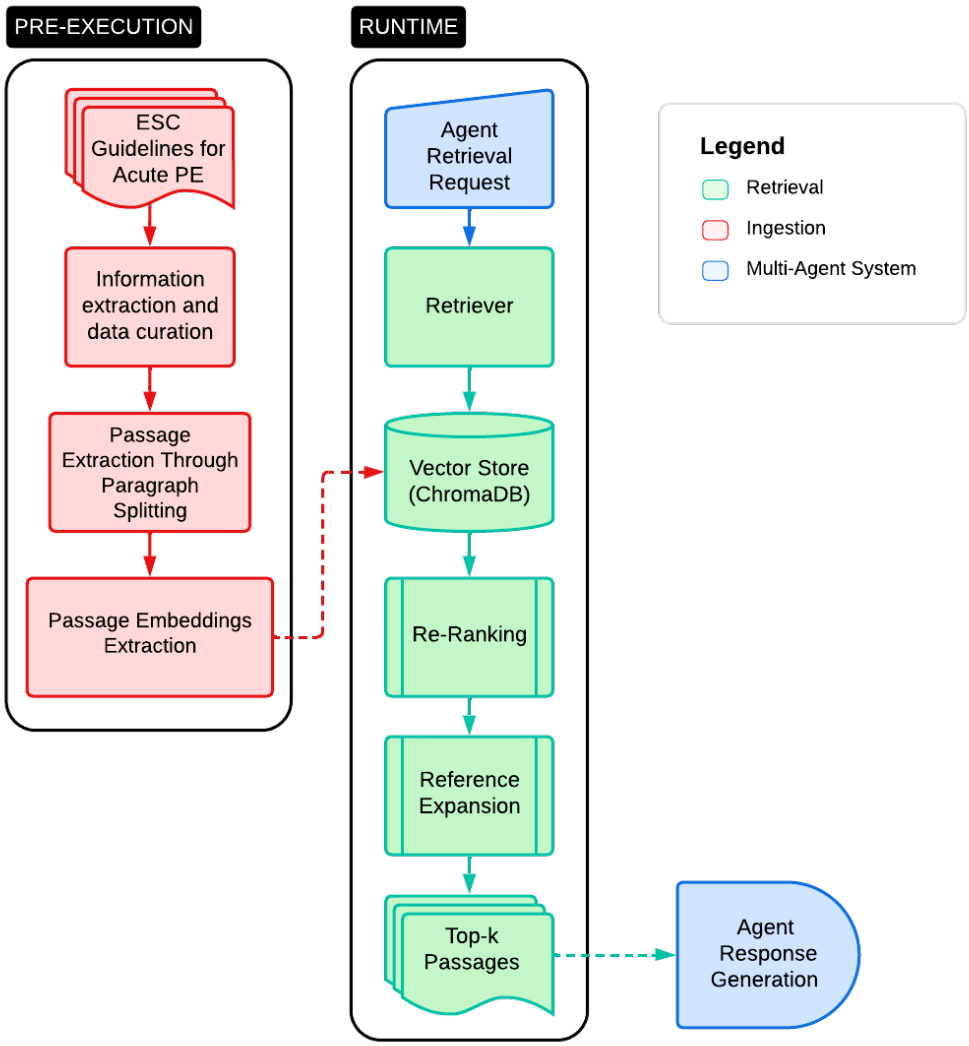


**Figure 5.** Architecture of the RAG module. The system is divided into two main phases: the Pre-execution phase (left), in which official clinical guidelines are processed, segmented, and embedded into a vector store; and the Runtime phase (right), during which LLM agents dynamically query the system to retrieve relevant information. The retrieval pipeline includes a retriever, a re-ranking stage, and a reference expansion mechanism to return the top-k most relevant passages. These passages are then provided as contextual input to the agent responsible for response generation.

#### 2.3.1. Ingestion Process

Before the execution of the system, the ingestion process is initiated with the transformation of official clinical guidelines (in PDF format) into a structured, machine-readable format. This step is performed before the deployment of the prototype and ensures that the knowledge from the source documents is efficiently indexed and accessible for retrieval process. In this study, the *ESC Guidelines for the diagnosis and management of acute pulmonary embolism 2019* and its Supplementary Data [27] documents have been selected as the primary reference materials.

First, the Docling tool [32] was used to convert the PDF documents into Markdown, allowing for more consistent parsing and text segmentation. Given the complexity of the original documents, additional manual curation was performed to ensure accurate content representation. Images were replaced with textual summaries, and malformed table structures were corrected. Once converted, the content was split into coherent passages (or "chunks") based on paragraph structure. Each passage retained metadata identifying its source section, referenced tables or figures, and other contextual anchors. This metadata was preserved to enable traceability and improved ranking during retrieval. Finally, the embeddings for each passage were generated using an embedding model, and stored in a ChromaDB vector store [33].

#### 2.3.2. Retrieval Process

During the runtime of the system, when a query is issued by an agent, the retrieval pipeline begins by encoding the query into a vector representation and computing its similarity against all stored passage embeddings. The top-k most relevant chunks are retrieved by the retriever based on cosine similarity. To further refine the result set, a re-ranking stage is applied using a LLM, which jointly considers all retrieved passages and reorders them based on their contextual relevance to the original query. Finally, the retrieval process includes a reference expansion mechanism. If any retrieved passage refers to additional sections, tables, or figures, the system attempts to locate and include those linked fragments in the final context. This ensures that the agent receives a more complete and coherent evidence base before generating a response.

From a research perspective, the inclusion of the RAG module plays a central role in evaluating the applicability of current information retrieval techniques in a highly complex and safety-critical domain such as healthcare. This component enables the system to study whether RAG-based pipelines can effectively provide a reliable knowledge base for multi-agent systems tasked with answering clinical questions and evaluating patient cases. By decoupling knowledge acquisition from hardcoded logic, the module allows agents to reason over dynamically retrieved medical evidence, mirroring how clinicians interact with textual guidelines. Furthermore, the system can be rapidly extended to support additional pathologies by simply ingesting new clinical guidelines, thus enabling the development of flexible, domain-adaptable CDSSs with minimal manual intervention.

## 3. Evaluation Framework

This section describes the evaluation framework used to assess the performance, reliability, and practical applicability of the proposed multi-agent system. The objective of this evaluation is to analyze how effectively the system responds to clinical information needs in two main scenarios: direct consultation of clinical guidelines and evaluation of structured clinical cases. To support this analysis, two custom evaluation datasets were created. The first consists of clinical questions designed to simulate the types of queries a physician might pose when consulting official guidelines in routine practice. The second includes realistic patient cases that reflect common diagnostic and therapeutic scenarios in

pulmonary embolism management. Together, these datasets provide a controlled framework for testing the capabilities of the LLM-based agents and exploring the potential of agent-based architectures to assist in clinical decision-making under real-world conditions. The following subsections detail the dataset construction and the evaluation protocol, including the metrics and criteria used.

### *3.1. Evaluation Datasets*

For the evaluation of the capabilities proposed in this study, it has been necessary to design and construct the two specific datasets from scratch, given that no public datasets currently exist that enable the measurement of these particular qualities in agents. The following section outlines the characteristics and construction methodology of both datasets.

#### 3.1.1. Guidelines Consultation Dataset

The first evaluation dataset was created to assess the system's ability to respond accurately to clinician queries based on the content of the official PE guidelines. This dataset consists of realistic, manually constructed questions that reflect the types of information a physician might seek when consulting the guidelines in a real clinical context.

The dataset comprises 45 carefully curated questions, divided into three distinct difficulty categories (basic, intermediate, and complex), each containing 15 questions, along with their corresponding reference answers. This approach enables a granular evaluation of the *Guidelines Consultation Service* across varying degrees of query complexity, thus providing insights into the agent's capability to handle both routine clinical inquiries and more challenging scenarios that require integration of multiple clinical parameters.

The basic category comprises unambiguous, factual queries that can be addressed almost directly from the guidelines text. The intermediate category incorporates questions requiring synthesis of multiple guideline sections and consideration of specific factors. The complex category presents complex clinical scenarios that demand sophisticated reasoning and integration of contraindications, special conditions, and advanced decision-making.

To illustrate the diversity and clinical relevance of the queries included in each difficulty category, the following examples are provided:

**Basic Level Queries**

1. "In which patients is plasma D-dimer measurement recommended to reduce the need for unnecessary imaging?"
2. "When is rescue thrombolytic therapy recommended in patients with pulmonary embolism?"

**Intermediate Level Queries**

1. "Why is CTPA not sufficient to exclude PE in some patients despite its high sensitivity and specificity?"
2. "In which clinical situations is unfractionated heparin (UFH) preferred over low-molecular-weight heparin (LMWH) or NOACs in the initial treatment of acute PE?"

**Advanced Level Queries**

1. "A 45-year-old woman presents with acute shortness of breath. She is stable, has a sPESI score of 0, and a normal chest X-ray. The hospital does not have access to CTPA, but V/Q SPECT is available. Is it appropriate to proceed with V/Q SPECT as the initial diagnostic imaging test if PE is suspected?"
2. "How do comorbidities such as cancer or heart failure affect the interpretation of PE incidence and risk?"

All queries were developed and validated by an experienced physician with extensive expertise in PE management, ensuring clinical authenticity and relevance to real-world practice scenarios. This expert validation guarantees that the evaluation framework accurately reflects the information-seeking behaviors and query formulation patterns characteristic of practicing clinicians, thus enhancing the validity of our evaluation methodology.

#### 3.1.2. Clinical Case Evaluation Dataset

The second evaluation dataset comprises 20 designed clinical cases of PE developed by the specialist physician. The cases have been systematically structured to represent the vast majority of clinical scenarios encountered in real-world practice, including the full spectrum from typical presentations to more complex or atypical cases that healthcare professionals may encounter in their daily clinical practice.

The simulated clinical cases consist of structured clinical data from fictitious patients, incorporating the following comprehensive information:

1. Demographic Data: Age, Sex, Weight and Height.
2. Patient Clinical History: Previous cancer, Previous heart failure, COPD and Chronic cardiopulmonary disease.
3. Absolute and Relative Contraindications for Thrombolysis: Active bleeding, Major surgery in the previous 3 weeks, Hemorragic stroke in the previous 6 months, Pregnancy, Traumatic Resucitation, ...
4. Physical Examination: Systolic/Diastolic BP, Hypoperfusion, Cardiac Frequency, Respiratory Frequency, Altered mental status, Glasgow scale and Oxygen saturation.
5. Echocardiography: RV dysfunction, RV/LV diameter radio, Tapse, McConnell sign, Right heart thrombi, IVC diameter, IVC inspiratory collaspsibility and TAPSV.
6. CTPA: Detected PE, RV/LV diameter ratio and IVC contrast reflux.
7. Laboratory Tests: Troponin elevated, NtroBNP elevated, Creatinine and CrCL.

The dataset comprises a combination of numerical, binary, and textual data, reflecting the heterogeneous nature of clinical information. Importantly, each patient profile includes only the minimum set of data required to enable a clinical case evaluation. However, not all relevant variables are available for every case. Some values has been deliberately marked as "Missing", indicating that they may still be accessible through user input if the system requests them. Others has been marked as "Not Available", representing diagnostic results or clinical parameters that cannot be obtained in real time, such as pending imaging studies or lab results. This setup is designed to replicate realistic clinical constraints, where decision-making must often proceed with incomplete information, and the system must determine whether to request additional data or issue a recommendation based on the evidence at hand. As an example, a clinical case is summarized below:

*The patient is an 18-year-old female patient with a confirmed diagnosis of pulmonary embolism, as evidenced by CT pulmonary angiography. She exhibits signs of hemodynamic and respiratory compromise, including tachycardia (111 bpm), low systolic blood pressure (100 mmHg), and increased respiratory rate (24 breaths per minute). Echocardiographic findings indicate right ventricular dysfunction, with a TAPSE value below the threshold and a positive McConnell sign. A notable clinical feature of this case is that the patient is pregnant, which introduces specific considerations when formulating diagnostic or therapeutic recommendations. Pregnancy constitutes a relative contraindication for certain interventions, particularly thrombolytic therapy, and requires the system to reason carefully about risk–benefit trade-offs. In addition, several fields relevant to risk stratification and contraindication assessment, such as cancer history, recent major trauma, or prior hemorrhagic stroke, are marked as missing. Other values, like the NT-proBNP level or echocardiographic RV/LV ratio, are marked as not available.*

This case exemplifies the type of complex scenario the system must handle, balancing partial information, comorbid conditions, and clinical risks while generating a safe and guideline-consistent recommendation.

### *3.2. Evaluation Methodology*

The following section outlines the methodology used to evaluate the system's performance across its two core functionalities. For each task, we describe the evaluation criteria, metrics, and procedures applied, providing a structured framework to assess the reliability, guideline adherence, and clinical relevance of the system's outputs. The evaluation was conducted using the two datasets introduced in Section 3.1, and combines automated metrics with expert-based analysis to assess the CDSS's performance from both objective and subjective perspectives.

#### 3.2.1. Guidelines Consultation Evaluation Procedure

To evaluate the system's performance in responding to clinical queries based on the official PE guidelines, the 15 questions from each predefined difficulty category were processed using the *Guidelines Consultation Service*. Two complementary evaluation strategies were applied: one based on automated metrics using the RAGAS framework [34], to analyze how effectively the system retrieves and uses relevant guideline content to support its answers; and another involving manual expert validation to assess the clinical correctness of the responses.

For the automated evaluation, the RAGAS framework was used to compute four quantitative metrics: *Context Precision*, *Context Recall*, *Context Relevance*, and *Response Groundedness*. The first three metrics serve to evaluate how well the agent is able to identify its information needs and how accurately the RAG module retrieves relevant content from the PE guidelines. The final metric, evaluates whether the generated answers are meaningfully supported by that retrieved evidence, thus measuring the degree to which the agent grounds its output in verified clinical information. Since RAGAS metrics rely on LLM-based inference, and such models may introduce slight stochastic variation, even when using temperature = 0, each metric was computed three times per question. The final results are reported as the mean and standard deviation across the three runs, to ensure consistency and reduce the influence of random fluctuations in model behavior.

Each metric is presented below with its specific purpose and interpretive context.

1. *Context Precision (CP)*

   This metric evaluates the proportion of the retrieved context that is actually useful for generating the system's final response. It reflects how focused and relevant the retrieved passages are in relation to the content of the answer. For each retrieved passage, an LLM determines whether the information it contains contributed meaningfully to the response.

   The score is calculated as the average usefulness of the retrieved passages, defined as:

$$Context\ Precision = \frac{1}{n}\sum_{i=1}^{n} PrecisionScore(r, c_i) \tag{1}$$

   where $n$ is the number of retrieved context passages, $r$ is the generated response, $c_i$ is the $i$-th passage and $PrecisionScore(c_i, r)$ is an LLM-evaluated score indicating whether $c_i$ was used to produce $r$.

   A high *Context Precision* score implies that the system retrieves concise and highly relevant information, minimizing unnecessary content, while a low score suggests that a significant portion of the retrieved context was irrelevant or unused, potentially introducing noise.

2. *Context Relevance (CRel)*
This metric measures the overall alignment between the retrieved context and the original physician query, independently of whether that context was ultimately used in the system's response. It assesses whether the retrieved passages are topically appropriate and meaningfully related to the user's question.
The metric is computed as the mean *RelevanceScore* across all retrieved context chunks for a given query:

$$Context\ Relevance = \frac{1}{n}\sum_{i=1}^{n} RelevanceScore(q, c_i) \tag{2}$$

where $n$ is the total number of retrieved chunks, $q$ is the input query, and $c_i$ is the $i$-th retrieved chunk.
Each *RelevanceScore* is obtained through an LLM-based classification process. Specifically, two independent prompts are used to assess how relevant each retrieved passage is to the query. Each prompt categorizes the passage as relevant, partially relevant, or not relevant, which are then mapped to numerical values of 1, 0.5, and 0, respectively. The final score for a given passage is the average of the two prompt-based evaluations:

$$ContextScore(q, c_i) = \frac{1}{2}(LLMScore_1(q, c_i) + LLScore_2(q, c_i)) \tag{3}$$

Unlike *Context Precision* or *Recall*, *Context Relevance* does not rely on the generated answer; it focuses exclusively on the relationship between the question and the retrieved evidence. As such, it provides insight into how well the system interprets clinician queries, formulates retrieval prompts, and selects the appropriate guideline content. A high Context Relevance score indicates that the agent successfully identifies the clinician's information need and that the RAG module is capable of retrieving concise, high-quality evidence to support downstream reasoning.
3. *Context Recall (CR)*
This metric measures the extent to which the system has retrieved all the necessary information needed to support the key elements of a correct answer. It focuses on completeness: whether the retrieved context includes all relevant pieces of evidence needed to produce a fully accurate response.
In RAGAS, rather than analyzing the passages directly, *Context Recall* examines the claims made in the reference answer and determines which of those are supported by the retrieved evidence.
Formally, the metric is defined as:

$$Context\ Recall = \frac{|SupportedClaims(C, ref)|}{|Claims(ref)|} \tag{4}$$

where $C$ represents the set of retrieved context passages, $ref$ is the reference answer, $Claims(ref)$ represents the set of factual assertions extracted from the reference, and the subset of those claims that can be verified using the retrieved passages is $SupportedClaims(ref, C)$.
This metric is computed using LLM-based inference to determine whether each reference claim is entailed by the retrieved context. A high *Context Recall* score indicates that the system has successfully gathered all essential information to support a complete response. In contrast, a low score reveals gaps in the retrieved context, suggesting that the system may have missed relevant evidence from the guidelines.

4. *Response Groundedness (RG)*
   This metric evaluates whether the content of the response generated by the system is explicitly supported by the retrieved context. It assesses whether each claim in the response can be found, either wholly or partially, in the retrieved evidence from the guidelines. Unlike other retrieval-focused metrics, *Response Groundedness* evaluates the final answer itself, ensuring that it does not include unsupported statements.
   The metric is computed as the mean *GroundednessScore* across all factual claims extracted from the generated response, evaluated against the retrieved context $C$:

$$\textit{Response Groundedness} = \frac{1}{n}\sum_{i=1}^{n} \textit{GroundednessScore}(C, r_i) \tag{5}$$

   where $n$ is the total number of factual claims in the response, $C$ represents the set of retrieved context passages, and $r_i$ denotes the $i$-th factual claim in the response.
   Each *GroundednessScore* is computed through LLM-based inference. Specifically, two independent prompt templates are used to assess how well each claim is supported by the context. Each prompt categorizes the claim groundedness as fully grounded, partially grounded, or not grounded, which are them mapped to numerical values of 1, 0.5, and 0, respectively. The final score for a given claim is the average of the evaluation of the two promtps:

$$\textit{GroundednessScore}(C, r_i) = \frac{1}{2}(\textit{LLMScore}_1(C, r_i) + \textit{LLMScore}_2(C, r_i)) \tag{6}$$

   This metric provides a critical measure of factual consistency, ensuring that the system's outputs are not only coherent but also verifiably based on retrieved medical evidence. In the context of this study, a high *Response Groundedness* score indicates that the CDSS is able to generate trustworthy answers that align explicitly with clinical guidelines, thus supporting transparency, traceability, and safety in real-world usage.

In parallel, the experienced physician reviewed the system-generated answers for all 45 questions. For each case, the expert was presented with the original query and the system's final response. Each answer was assigned a binary score (1 or 0), where a score of 1 was given only when the response was clinically accurate, relevant, and fully addressed the query. This evaluation strategy, referred to hereafter as *Binary Clinical Accuracy (BCA)*, was adopted due to the inherent difficulty of constructing a single ground-truth answer per question and the variability in acceptable medical responses. A binary scoring system was preferred over a graded scale to reduce ambiguity and simplify the expert judgment process, particularly given the heterogeneity of the questions. The BCA score serves as the primary indicator of the system's clinical validity, complementing the automated retrieval-focused metrics described above.

### 3.2.2. Clinical Case Evaluation Procedure

To evaluate the system's performance in the generation of clinical case evaluations and personalized diagnostic and therapeutic recommendations, the 20 simulated patient cases described in Section 3.1 were processed using the *Clinical Case Evaluation Service*. Each case presents a different clinical profile, with varying levels of data completeness and patient complexity. The evaluation focused on assessing whether the system was able to correctly interpret the clinical case, to manage missing or unavailable information in an appropriate manner, and to produce recommendations that are medically sound and in alignment with official PE clinical guidelines.

Due to the complexity and clinical nuance involved in evaluating full patient cases, applying the same automated metrics used for guideline-based question answering was

considered inappropriate for this task. These metrics are not designed to assess clinical case evaluations, as they lack the capacity to interpret which information should be retrieved or how the context contributes to case-specific clinical reasoning. Moreover, no standard benchmark currently exists for this type of evaluation in agent-based CDSSs.

Given these limitations, a dedicated external review process was designed involving ten consultant-level physicians with specializations in critical care and cardiology. The objective was to gather expert clinical judgment on the quality, reliability, and safety of the recommendations produced by the system for real-world clinical practice. Each physician reviewed the full reasoning process carried out by the CDSS for three randomly assigned cases from a total pool of 20 simulated patient profiles. The distribution ensured that each case was evaluated by at least one expert, while the random assignment preserved the diversity and balance of clinical conditions reviewed across participants. After examining their assigned cases, each expert completed a structured and anonymous questionnaire composed of six targeted questions. Each physician was given access to the complete case report, including intermediate reasoning steps and tool calls, risk scores, and any relevant system output generated during the evaluation process.

Each question corresponds to a specific evaluation criterion aimed at exploring a distinct dimension of the system's clinical relevance and functional quality. The six criteria assessed were: *Clinical Accuracy*, *Specificity and Completeness*, *Risk Metrics Computation*, *Clinical Safety*, *Reasoning and Interpretability*, and *Trust in the System*. Experts rated each criterion using a 7-point Likert scale, where 1 indicated strong disagreement and 7 indicated strong agreement with a specific statement describing the behavior of the system. To reduce ambiguity, each item included a short description of the dimension being evaluated along with a clearly defined statement for the expert to rate.

The scores collected for each evaluation criterion will be analyzed both individually and in aggregate to assess the overall perception of the system's clinical performance in all reviewed cases. This analysis will provide insights into the perceived strengths and limitations of the *Clinical Case Evaluation Service*, as judged by experienced physicians in the context of realistic, case-based scenarios.

The six evaluation criteria used in the expert review are described below, along with the specific statements rated by the clinicians:

1. *Clinical Accuracy*
   Description: Assesses whether the recommendations provided are clinically correct and aligned with official guidelines and the specific clinical case, without contradictions or factual errors.
   Statement: *"The clinical recommendations provided by the system are correct, aligned with official clinical guidelines, and clearly adapted to the specific characteristics of each evaluated clinical case."*
2. *Specificity and Completeness*
   Description: Evaluates whether the recommendations address in sufficient detail the actions to be taken and whether they cover all the important aspects for correct clinical decision making, avoiding partial or incomplete answers.
   Statement: *"The recommendations provided by the system are specific, detailed, and complete for managing the clinical case, including all necessary diagnostic and therapeutic measures without omitting important aspects."*
3. *Risk Metrics Computation*
   Description: Evaluates whether the system performs calculations accurately and rigorously, without errors in formulas or classifications, and whether it correctly reflects the patient's clinical condition.

Statement: *"The system correctly calculates the clinical risk stratification metrics (PESI, sPESI, and Early Mortality Risk) based on the patient's available data, applying the formulas correctly, interpreting the variables properly, and assigning the appropriate risk categories according to the patient's condition."*

4. *Clinical Safety*
   Description: Assesses whether the system ensures patient safety by avoiding contraindicated or unsafe recommendations and by adhering to clinical safety standards in all suggested actions.
   Statement: *"The system avoids recommendations that could put the patient at risk, correctly identifies contraindications, and ensures that all recommendations strictly follow clinical safety standards to minimize risks."*
5. *Reasoning and Interpretability*
   Description: Evaluates whether the system's reasoning process is logical, transparent, and understandable, enabling clinicians to follow how and why the recommendations were made.
   Statement: *"The system uses logical and coherent clinical reasoning, and is able to clearly explain how it arrives at its conclusions and recommendations."*
6. *Trust in the System*
   Description: Measures the clinician's overall level of trust in the system as a decision-support tool, based on the perceived reliability, usefulness, and safety of its recommendations.
   Statement: *"I trust the system as a reliable support tool for clinical decision-making, considering the accuracy, usefulness, and safety of the responses provided."*

## 4. Results & Discussion

This section presents and analyzes the results obtained from the evaluation methodology described in Section 3. The aim is to assess the performance, reliability, and clinical applicability of the proposed multi-agent CDSS in its two core functionalities: guideline-based question answering and patient-specific clinical case evaluation. Quantitative results from automated metrics are combined with qualitative insights derived from expert assessments, offering a comprehensive perspective on the system's behavior. The discussion is structured to highlight key findings, interpret their implications in the context of clinical decision support, and identify limitations and potential directions for future improvements.

All experiments were conducted using models available through the official OpenAI API [35]. First, for the computation of the LLM-based metrics from RAGAS, it was used the GPT-4o model [36]. For the document ingestion process, it was employed the *text-embedding-3-large* model [37], OpenAI's most advanced embedding model at the time of the study. This model was used to generate the dense vector representations of the guideline passages that populate the vector database. For all LLM agents, the *GPT-4.1 mini*[38] model was used. This version was chosen not only for its favorable balance between generation quality and inference cost, but also for methodological reasons. Relying on the most powerful model available could lead to results that are heavily dependent on raw model capacity, potentially masking weaknesses in system design. In contrast, *GPT-4.1 mini* provides a performance level that is more representative of models commonly used in applied or open-source environments. This makes the evaluation more robust and the results more generalizable to other agent-based CDSSs implementations. In particular, *GPT-4.1* models provide a significantly larger context window than *GPT-4o*, which is particularly beneficial in agent-based reasoning tasks that involve long input chains and require persistent memory across multiple reasoning steps. Finally, for the computation of LLM-based RAGAS metrics, it was used *GPT-4o*.

### 4.1. Guidelines Consultation Service Results

This subsection presents the results obtained from the evaluation of the system's performance in answering clinical queries based on the official PE guidelines. As described in Section 3.2.1, the analysis includes both automated metrics, computed using the RAGAS framework, and manual expert validation through the *BCA* score. This results will serve to assess how effectively the CDSS retrieves relevant information form the clinical guidelines, grounds its responses in that information, and provides clinically accurate responses. The results for each difficulty level are presented in Table 1.

**Table 1.** Experimental results for the *Guidelines Consultation Service* by difficulty level. All metrics are expressed as percentages (%). RAGAS metrics are reported as mean ± standard deviation across three runs. *BCA* indicates the percentage of answers rated as fully correct by the clinical expert, out of a total of 15 questions per difficulty level.

| Difficulty Level | CP (%) | CR (%) | CRel (%) | RG (%) | BCA (%) |
|---|---|---|---|---|---|
| Basic | 86.58 ± 0.74 | 86.67 ± 0.00 | 93.33 ± 00.00 | 100.00 ± 0.00 | 100.00 |
| Intermediate | 95.09 ± 0.24 | 91.48 ± 2.31 | 91.11 ± 0.96 | 100.00 ± 0.00 | 100.00 |
| Complex | 93.89 ± 0.87 | 71.48 ± 0.07 | 87.22 ± 0.96 | 98.89 ± 0.96 | 93.33 |

The results obtained across the three levels of question difficulty reveal a strong overall performance of the system in addressing clinical questions related to PE based on the official guidelines. All automated metrics remain at high levels, exhibiting only moderate variation across categories, and the system maintains a high rate of fully correct responses according to expert judgment. These findings indicate that the CDSS is capable not only of retrieving relevant content but also of effectively reasoning over that information to produce clinically accurate answers. In the following analysis, the quality of context preparation is examined, along with its relationship with clinical accuracy, and the grounding of the system's outputs.

The *BCA* results confirm the clinical accuracy and reliability of the system in handling clinical questions. The system attained perfect scores in both the basic and intermediate categories, successfully addressing queries that necessitated either direct extraction or moderate integration of information from the guidelines. In the complex category, which involved more abstract reasoning and contextual interpretation, the system still performed very well, correctly answering 14 of 15 queries (93.33%). These results highlight the agent's capacity to adapt general clinical knowledge to specific cases, even when the answer is not explicitly delineated in the guidelines. Notably, the clinical expert observed that the responses were not only precise but also surprisingly well-structured and detailed, reflecting a coherent reasoning process and a comprehensive grasp of the guidelines content. In the single complex case where the system failed, a mandatory but rare recommendation was omitted. Although the response was not clinically unsafe, this isolated instance suggests that, in low-frequency, high-importance scenarios, additional user prompting may be needed. However, this is not considered a significant issue due to the CDSS's interactive design, which allows follow-up queries to refine or complete the response.

The CDSS demonstrated high *CP*, *CR*, and *CRel* across all levels of difficulty. This suggests that the RAG module is effective in retrieving appropriate passages from the clinical guidelines, even in more complex conditions where queries require broader interpretation rather than straightforward factual lookup. The significant decrease in *CR* for complex queries (71.5%) can be attributed to the inherent characteristics of these inquiries. These queries are not addressed explicitly in the guidelines, but require the system to apply general recommendations to a specific clinical context. Consequently, the LLM responsible for computing the recall score identified that none of the retrieved passages directly contained the answer, leading to lower recall values despite the quality of the agent's reasoning.

Despite this decline in recall, the *BCA* for complex queries remained high (93.3%), thus confirming the system's capacity to extrapolate general clinical knowledge from the guidelines and apply it effectively to specific situations. In other words, the agent's reasoning abilities allow it to translate general medical principles into query-adapted responses, even when the answer must be inferred rather than retrieved verbatim. This is indicative of a key advantage of agent architectures compared to traditional approaches: they enable the LLM to reason flexibly over incomplete or indirect evidence, rather than relying solely on exact matches between the query and the text.

A further unanticipated finding is that *CP* and *CR* scores for basic queries are slightly lower than for intermediate ones. This behavior is explained by how the retrieval system works, although the answers to basic questions may require only one or two sentences, the RAG module retrieves full passages (paragraphs) from the guidelines. As a result, while the correct information is present in the retrieved context, much of the surrounding content may not be contributory to the final answer, thus lowering the precision and recall metrics. Conversely, intermediate questions typically require an extensive information base, meaning that a larger portion of the retrieved context contributes directly to the final answer, raising these scores. This effect is given that the guideline documents were segmented in paragraphs during the ingestion process, resulting in some passages containing excessive context for basic cases. However, this becomes beneficial at higher difficulty levels, as this ensures more complete context is available to support complex reasoning.

In terms of faithfulness in responses, *RG* remained nearly perfect across all levels, confirming that the generated responses, regardless of difficulty, were consistently supported by the content retrieved from the guidelines. Even in the most challenging queries, requiring multi-step reasoning, the system avoided fabricating or hallucinating unsupported content. Instead, it exhibited the ability to construct its responses on the retrieved evidence. The slight reduction in *RG* for complex queries likely reflects the increased need to combine multiple passages and reasoning steps, rather than a loss in factual quality. This observation is further supported by the fact that the hallucination detection module did not identify any hallucinated statements during the entire evaluation process.

Taken together, these findings demonstrate the promising capabilities of the agent-based CDSS to support clinicians in: (i) retrieving relevant information from official guidelines, and (ii) resolving complex queries that require interpreting general recommendations and applying them to specific clinical contexts. The system has shown to be reliable and consistent, grounding all of its responses in verified content drawn directly from the clinical guidelines. Furthermore, the evaluation confirms the agent's capacity for autonomously and dynamically retrieving the clinical information it needs to formulate its answers. The results across *CP*, *CR*, and *CRel* indicate a robust performance in building a relevant contextual foundation for reasoning. Nevertheless, there remains room for improvement in optimizing both the ingestion and retrieval processes. Enhancing these components could help reduce unnecessary contextual noise, lower the cost of inference, and ensure that the agent consistently receives concise, high-value clinical input, ultimately improving the quality and interpretability of its responses.

These capabilities underscore the potential of agent-based architectures in real-world clinical workflows. The system has proved not only the ability to retrieve exact content from official guidelines, but also to reason over that information to answer more complex and elaborated queries posed in natural language. This capability could be particularly valuable in time-sensitive environments, such as emergency departments or inpatient wards, where clinicians often need rapid, flexible, and trustworthy access to guideline-based knowledge. By allowing unstructured, context-aware consultations, the system provides an accurate and interpretable solution to support day-to-day clinical decision-making.

**Illustrative Case**

The following is an example of a final response generated by the *Guidelines Consultation Service*. For brevity, only the physician introduced query and the final response is included:

**User Query:**
*"A 78-year-old woman with newly diagnosed PE is found to have a creatinine clearance (CrCl) of 25 mL/min. What anticoagulant regimen is preferred and why?"*

**System Response:**
*In a 78-year-old patient with newly diagnosed with PE and a creatinine clearance (CrCl) of 25 mL/min, the preferred anticoagulant regimen is unfractionated heparin (UFH) rather than low-molecular-weight heparin (LMWH) or non-vitamin K antagonist oral anticoagulants (NOACs).*

*This preference is based on the following guideline recommendations and rationale:*

1. *Patients with serious renal impairment, defined as CrCl <30 mL/min, are recommended to receive UFH for initial anticoagulation. UFH dosing can be adjusted based on activated partial thromboplastin time (aPTT) monitoring, allowing for safer management in the context of reduced renal clearance.*
2. *LMWH and fondaparinux, although generally preferred over UFH due to lower bleeding risk and heparin-induced thrombocytopenia, require dose adjustment in patients with CrCl between 15–30 mL/min and are used with caution. However, in severe renal impairment, UFH is favored because it is not renally cleared and its anticoagulant effect can be more readily controlled and reversed.*
3. *NOACs are generally avoided in patients with severe renal impairment. Specifically, dabigatran is not recommended for CrCl <30 mL/min; edoxaban is not recommended for CrCl <15 mL/min and requires dose adjustment for CrCl 15–50 mL/min; rivaroxaban and apixaban should be used with caution in CrCl 15–29 mL/min and are not recommended below 15 mL/min. Given the patient's CrCl of 25 mL/min, NOACs pose a higher risk and are not the preferred choice.*

*Summary: For this elderly patient with PE and CrCl of 25 mL/min, UFH is the preferred initial anticoagulant due to its safety profile in severe renal impairment, ability to be monitored and dose-adjusted via aPTT, and avoidance of accumulation risks associated with LMWH and NOACs in this setting.*

This example demonstrates the value of the *Guidelines Consultation Service* in addressing complex, clinically relevant queries that go beyond direct factual lookups. The system is able to analyze the structure and intent of the user's question, retrieve the most relevant sections of the guidelines, and apply that information to provide a reasoned and interpretable response. Rather than merely extracting text, the agent synthesizes dispersed knowledge and articulates a clinical rationale that reflects guideline-based decision-making. This level of reasoning enables physicians to engage with the system in a flexible and natural way, posing elaborate or nuanced questions and receiving structured answers that they can critically evaluate, validate, and integrate into their own clinical judgment. In this way, the system acts not just as a static reference tool, but as a dynamic assistant capable of supporting real-world medical reasoning.

### 4.2. Clinical Case Evaluation Service Results

This subsection presents the results of the evaluation conducted on the *Clinical Case Evaluation Service*. As detailed in Section 3.2.2, the evaluation was based on a structured expert review, in which ten consultant-level physicians assessed the system's ability to analyze patient clinical cases and generate appropriate clinical recommendations. The analysis focuses on six key dimensions, *clinical accuracy*, *specificity and completeness*, *risk metrics computation*, *clinical safety*, *reasoning and interpretability*, and *trust*, captured through a structured questionnaire. These results aim to provide insight into the system's clinical reliability, reasoning quality, and overall acceptance by medical professionals. The results of the questionnaire are displayed in Table 2.

**Table 2.** Evaluation results for the *Clinical Case Evaluation Service* based on expert review. Each criterion was rated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree). Results are reported as mean ± standard deviation (SD), both on the original scale and as a normalized percentage.

| Criterion | Mean (1–7) ± SD | Mean (%) ± SD |
|---|---|---|
| Clinical Accuracy | 6.70 ± 0.48 | 95.00 ± 8.05 |
| Specificity and Completeness | 6.70 ± 0.48 | 95.00 ± 8.05 |
| Risk Score Computation | 7.00 ± 0.00 | 100.00 ± 0.00 |
| Clinical Safety | 6.30 ± 0.82 | 88.33 ± 13.72 |
| Reasoning and Interpretability | 6.70 ± 0.48 | 95.00 ± 8.05 |
| Trust in the System | 6.00 ± 1.05 | 83.33 ± 17.56 |

The results obtained from the expert-based evaluation reflect a strong performance of the *Clinical Case Evaluation Service* across all six assessed dimensions. All criteria received mean scores above 6 on a 7-point Likert scale, corresponding to levels of agreement above 83%, and most criteria exceeded the 90% threshold. These results surpassed initial expectations, especially considering the inherent complexity of clinical case evaluation and the variability of judgment among healthcare experts. The consistency and strength of the scores across all dimensions underscore the potential of agent-based systems that are capable of autonomously interpreting official medical guidelines and applying that knowledge to real-world patient scenarios. These results provide compelling evidence that such architectures can support complex clinical decision-making with a level of accuracy and relevance that aligns closely with expert medical standards.

The highest-rated dimension was *Risk Score Computation*, which achieved a perfect score (7.00), indicating unanimous agreement among the clinicians that the system correctly calculated and applied the PESI, sPESI, and Early Mortality Risk metrics. This result validates the system's ability to interpret structured patient data and apply established scoring methodologies consistently and without error.

Similarly, *Clinical Accuracy*, *Specificity and Completeness*, and *Reasoning and Interpretability* stood out with particularly strong results, all receiving mean scores of 6.7 out of 7 (above 95% agreement) and low standard deviations. These values indicate not only a high degree of confidence among experts regarding the correctness of the system's recommendations, but also that those recommendations were perceived as complete, clinically accurate, and well articulated. Taken together, these three metrics provide direct evidence of the system's ability to reason over complex clinical cases and generate recommendations that are both compliant with clinical guidelines and specifically adapted to the patient at hand. The prototype also demonstrated the agents' capacity to dynamically interpret guideline content, extract the most relevant information based on the clinical case at hand, and apply it effectively to generate precise and patient-specific responses. Furthermore, by

expressing its reasoning process in natural language, the system offers physicians detailed and transparent justifications for each of its actions and conclusions. This enhances the interpretability of the CDSS and gives clinicians the necessary insight to assess and validate the output, promoting a collaborative decision-making process in clinical practice.

*Clinical Safety* received a slightly lower score (6.30 ± 0.82), though still indicating a high level of agreement that the system avoids unsafe or contraindicated recommendations and adheres to accepted clinical standards. The observed variability does not appear to stem from inconsistencies in system performance, but rather from differences in how individual experts perceive the broader implications of relying on AI-based decision support in high-stakes contexts. This interpretation is reinforced by the results for *Trust in the System*, which, despite being positive overall (6.00 ± 1.05), exhibited the greatest variability among all dimensions. These scores reflect the diversity of opinions among clinicians regarding the safety and reliability of AI systems in clinical practice, not necessarily doubts about specific outputs, but rather about the general readiness of such technologies. This suggests that ongoing efforts to validate the safety of these systems will be essential to build trust and ensure safe adoption in real-world clinical environments.

In summary, the results of this evaluation suggest that agent-based architectures hold substantial promise for the development of next-generation CDSSs capable of assisting physicians in real-world clinical workflows. The high scores obtained in *Clinical Accuracy*, *Specificity and Completeness*, *Clinical Safety*, and *Risk Score Computation* demonstrate that this approach can produce recommendations that are not only correct and well-adapted to the specific clinical case, but also detailed, clinically actionable, and safe for the patient. Furthermore, the strong results observed in *Reasoning and Interpretability* confirm that this architecture is capable of emulating clinical reasoning in a manner similar to how a human expert would operate. The system not only applies relevant knowledge to each situation, but also articulates its conclusions through clear and comprehensible explanations, allowing clinicians to understand, assess, and validate the recommendations provided.

Despite this overall positive assessment, it is worth noting that some experts remain cautious regarding the integration of AI-driven systems into routine practice. This was reflected in the slightly lower and more variable scores for *Trust in the System*, which likely stem from broader concerns about the safety and reliability of autonomous decision-support tools in high-stakes environments. These concerns highlight the need for continued research into the robustness, safety, and practical viability of agent-based CDSSs. More importantly, they point to the importance of fostering a collaborative human-AI relationship, in which such systems are positioned not as autonomous decision-makers, but as assistive tools designed to enhance clinical practice. In this model, the physician remains the final arbiter, responsible for validating and interpreting the system's outputs within the full context of the patient's condition and care environment.

**Illustrative Case**

The following example illustrates a final report generated by the *Clinical Case Evaluation Service*. For brevity, only a patient summary and the final report is included:

> **Patient Summary:**
> *Male, 67 years old. History of cancer and chronic cardiopulmonary disease. Recently underwent major surgery within the last three weeks. On physical examination, the patient presents with marked tachycardia (127 bpm), hypotension (SBP: 97 mmHg), high respiratory rate (45 breaths/min), and signs of hypoperfusion, with oxygen saturation at 89%. No echocardiographic or laboratory data are available at this time. Absolute contraindications for thrombolysis are partially unconfirmed due to missing information regarding recent cerebrovascular events. Relative contraindications were not identified.*

**System Report:**

***Patient state:*** *The patient presents with clinical signs of hemodynamic instability, including borderline hypotension, tachycardia, severe tachypnea, low oxygen saturation, and evidence of hypoperfusion. Both PESI and sPESI scores are markedly elevated—Class V (207) and High Risk (5), respectively—indicating a high risk of early mortality. No imaging or laboratory data are currently available to confirm pulmonary embolism or right ventricular dysfunction. The patient has an absolute contraindication to systemic thrombolysis due to recent major surgery.*

***Diagnosis:*** *Suspected high-risk pulmonary embolism with hemodynamic instability. The clinical presentation strongly suggests high-risk PE, but confirmatory imaging and laboratory tests are currently unavailable. The patient is unstable, requiring urgent diagnostic and therapeutic interventions.*

***Recommendations:***

1. ***Immediate Diagnostic Workup:***
    - ***Bedside Transthoracic Echocardiography (TTE):*** *Urgently assess for RV dysfunction and exclude alternative causes of shock. This is critical as echocardiographic evidence of RV dysfunction in an unstable patient with suspected PE justifies emergency reperfusion therapy. (Class I, Level C)*
    - ***Bilateral Venous Compression Ultrasonography (CUS):*** *To detect proximal DVT, which supports venous thromboembolism diagnosis and may guide urgent management if CTPA is not feasible (Class I, Level C)*
    - ***CT Pulmonary Angiography (CTPA):*** *Perform as soon as feasible to confirm PE diagnosis and assess RV/LV ratio, but do not delay treatment if the patient is too unstable to transfer. (Class I, Level C)*
    - ***Laboratory Tests:*** *Obtain cardiac biomarkers (troponin, NT-proBNP) and renal function tests to assist risk stratification and guide anticoagulation dosing. (Class I, Level C)*
2. ***Therapeutic Management:***
    - ***Anticoagulation:*** *Initiate immediately with intravenous UFH including a weight-adjusted bolus (80 units/kg) and continuous infusion (18 units/kg/h), monitored via aPTT; unless contraindications develop. UFH is preferred due to its short half-life and reversibility, especially given recent surgery. (Class I, Level C)*
    - ***Reperfusion Therapy:*** *Systemic thrombolysis is contraindicated due to recent major surgery. Therefore, consider alternative strategies:*
        - ***Surgical Pulmonary Embolectomy:*** *First-line option if expertise and resources available, given the high-risk PE and contraindication to thrombolysis. (Class I, Level C)*
        - ***Percutaneous Catheter-Directed Treatment:*** *Alternative if surgery is not feasible. (Class IIa, Level C)*
    - ***Hemodynamic Support:*** *Initiate vasopressor support with norepinephrine to maintain adequate MAP ($\geq$ 65 mmHg); consider dobutamine in case of RV failure with low output. (Class IIa, Level C)*
    - ***Extracorporeal Membrane Oxygenation (ECMO):*** *Consider ECMO as a bridge to reperfusion in case of refractory collapse or cardiac arrest. (Class IIb, Level C)*

3. ***Monitoring and Follow-up:***
    - *Continuous hemodynamic and respiratory monitoring in intensive care settings is mandatory.*
    - *After reperfusion and stabilization, plan transition to long-term anticoagulation, considering cancer status and renal function.*
    - *Given the patient's cancer history, after stabilization, LMWH or DOACs may be considered; following oncologic thrombosis guidelines and renal function results.*

***Summary:*** *This patient presents with suspected high-risk PE and hemodynamic instability, compounded by recent major surgery contraindicating systemic thrombolysis. Immediate bedside echocardiography and venous ultrasonography are essential to confirm RV dysfunction and DVT, guiding urgent management. Anticoagulation with UFH should be started without delay. Given the absolute contraindication to thrombolysis, surgical pulmonary embolectomy is the preferred reperfusion strategy if available; otherwise, catheter-directed treatment is a reasonable alternative. Hemodynamic support with vasopressors is critical. Close monitoring in ICU and prompt diagnostic confirmation with CTPA when feasible are necessary. This approach balances the urgent need for reperfusion with the patient's bleeding risk and clinical instability.*

This example illustrates the potential value of the *Clinical Case Evaluation Service* as a practical decision-support tool that can assist physicians by transforming complex clinical cases into an organized, actionable report. The system provides a complete overview of the patient's condition, a likely working diagnosis, and a structured plan of care that integrates both diagnostic and therapeutic recommendations. Each proposed action is justified with its corresponding guideline strength and level of evidence, offering clinicians not only a prioritized roadmap but also the necessary context to evaluate each decision. Importantly, the final summary condenses this information into a format that facilitates rapid understanding. In routine clinical settings, where physicians are often under pressure, managing multiple patients, and navigating large volumes of information, this type of intelligent support can help mitigate clinicians' cognitive overload and improve clinical outcomes. Tools like this, which can reason over patient data, apply up-to-date guidelines, and clearly communicate rationale and priorities, represent a meaningful step toward safer, more efficient, and more supported clinical workflows.

## 5. Conclusions

This work presents an innovative approach to the design of AI-based CDSSs that integrates LLM agents with RAG pipelines to enhance the clinical decision support capabilities of these systems. This agentic architecture makes CDSS to move beyond static or rule-based approaches, towards systems that reason about the steps required to complete each clinical task and dynamically orchestrate sub-tasks toward the final objective. Through the use of the LLM agents, the system gains the ability to act autonomously and the capacity to reason over clinical information. The RAG module provides this agents the access to the content from official clinical guidelines, allowing them to retrieve the information necessary to accomplish their tasks. This ensures that all reasoning processes and decisions are grounded in validated medical knowledge issued by official health organizations.

To explore the clinical viability and relevance of using LLM agents to enhance CDSSs with reasoning capabilities, we implemented a prototype system based on this agentic architecture, specialized in the management of PE. The system was designed to provide two core services through which the reasoning abilities of LLM agents were evaluated: (i) a guideline consultation service, where the agents retrieve and interpret guideline

information to answer clinical questions posed in natural language, and (ii) a clinical case evaluation service, where the agents analyze structured patient data to generate a report of the clinical case with diagnostic and therapeutic recommendations adapted to it. These two complementary functionalities enabled a comprehensive assessment of the system's ability to support real-world clinical decision-making across both knowledge retrieval and case-based reasoning tasks.

The evaluation results obtained from both system functionalities demonstrate the promising potential of this approach. In the guideline consultation service, the system consistently produced accurate, well-grounded responses across all levels of query complexity, as confirmed by automated metrics and expert validation. The agents were able to dynamically retrieve relevant content, reason over complex clinical questions, and articulate their conclusions in a form usable by physicians. Notably, even when queries required integrating dispersed guideline information or handling ambiguous contexts, the system maintained high accuracy and interpretability, highlighting the value of agent-based orchestration in handling complex informational needs.

On the other hand, the clinical case evaluation service showed strong performance in generating personalized, guideline-compliant recommendations based on realistic, structured patient data. A group of ten physicians experts rated the system highly across all evaluation dimensions (clinical accuracy, completeness, safety, interpretability, and trust), demonstrating that agents are able to synthesize relevant patient information, apply appropriate guidelines, and produce clinically relevant evalations of clinical cases. These findings suggest that agent-based CDSSs can serve as effective cognitive aids for physicians, particularly in managing complex or high-risk cases where information overload and time constraints may otherwise hinder optimal decision-making.

Despite these encouraging results, certain limitations and challenges remain. The variability observed in expert trust ratings underscores the need for further validation and user-centered refinement to foster broader acceptance of agent-based CDSSs in clinical practice. Additionally, while the system achieved strong performance in a single-domain context (PE management), broader generalization to other clinical areas and more heterogeneous patient populations will require further development and testing. Improving the efficiency and precision of the retrieval and reasoning pipelines also remains a priority, to ensure that agents consistently receive concise, high-value context and can operate at scale within real-world clinical workflows. Future work will focus on expanding the system to support additional clinical conditions, integrating more advanced retrieval methods, and conducting prospective studies to evaluate the impact of this architecture on clinical outcomes, workflow efficiency, and physician experience. Ultimately, the goal is to develop CDSSs that operate as collaborative, transparent, and trustworthy tool for clinicians, to support human expertise in complex decision-making environments.

**Author Contributions:** Conceptualization, A.G., A.P. and I.M.; methodology, A.G., A.P. and I.M.; software, A.G.; validation, A.G., and I.M; formal analysis, A.G.; investigation, A.G.; resources, A.G., and I.M; data curation, A.G., and I.M; writing—original draft preparation, A.G.; writing—review and editing, A.P., L.D and C.G; visualization, A.G.; supervision, A.P., L.D and C.G; project administration, A.P., L.D. and C.G. All authors have read and agreed to the published version of the manuscript.

**Data Availability Statement:** The raw data and source code supporting the conclusions of this article are available on request from the corresponding author. The materials are currently being prepared for public release in a dedicated repository (GitHub), which is under development due to time constraints.

**Funding:** This research has received funding from the Consellería de Cultura, Educación, Formación profesional e Universidades of the Xunta de Galicia for the "Centro singular de investigación de Gali-

cia" accreditation 2019-2022 and by the "Axudas para a consolidación e estructuración de unidades de investigación competitivas do Sistema Universitario de Galicia -ED431B 2024/36", and the European Union for the "European Regional Development Fund - ERDF".

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| AI | Artificial Intelligence |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| CDSS | Clinical Decision Support System |
| RAG | Retrieval-Augmented Generation |
| EHR | Electronic Health Record |
| LLM | Large Language Model |
| SotA | State of the Art |
| PE | Pulmonary Embolism |
| PESI | Pulmonary Embolism Severity Index |
| sPESI | Simplified Pulmonary Embolism Severity Index |
| ESC | European Society of Cardiology |
| SD | Standard Deviation |
| CP | Context Relevance |
| CR | Context Recall |
| CRel | Context Relevance |
| BCA | Binary Clinical Accuracy |
| CrCl | Creatinine Clearance |
| UFH | Unfractionated Heparin |
| LMWH | Low-Molecular-Weight Heparin |
| NOACs | Non-vitamin K antagonist Oral Anticoagulants |
| aPTT | activated Partial Thromboplastin Time |
| SBP | Systolic Blood Pressure |
| TTE | Transthoracic Echocardiography |
| CUS | Compression Ultrasound |
| DVT | Deep Vein Thrombosis |
| CTPA | Creatinine Clearance |
| CTPA | Computed Tomography Pulmonary Angiography |
| RV | Right Ventricle |
| LV | Left Ventricle |
| MAP | Mean Arterial Pressure |
| ECMO | Extracorporeal Membrane Oxygenation |
| DOACs | Direct Oral Anticoagulants |

## References

1. Beeler, P.E.; Bates, D.W.; Hug, B.L. Clinical decision support systems. *Swiss medical weekly* **2014**, *144*, w14073.
2. Khalifa, M.; Albadawy, M.; Iqbal, U. Advancing clinical decision support: the role of artificial intelligence across six domains. *Computer Methods and Programs in Biomedicine Update* **2024**, p. 100142.
3. Sloane, E.B.; Silva, R.J. Artificial intelligence in medical devices and clinical decision support systems. In *Clinical engineering handbook*; Elsevier, 2020; pp. 556–568.
4. Kumar, A.; Joshi, S. Applications of AI in healthcare sector for enhancement of medical decision making and quality of service. In Proceedings of the 2022 International Conference on Decision Aid Sciences and Applications (DASA). IEEE, 2022, pp. 37–41.

5. Almagharbeh, W.T. The impact of AI-based decision support systems on nursing workflows in critical care units. *International Nursing Review* **2024**.
6. Wang, D.; Wang, L.; Zhang, Z.; Wang, D.; Zhu, H.; Gao, Y.; Fan, X.; Tian, F. "Brilliant AI doctor" in rural clinics: Challenges in AI-powered clinical decision support system deployment. In Proceedings of the Proceedings of the 2021 CHI conference on human factors in computing systems, 2021, pp. 1–18.
7. Shortliffe, E.H.; Davis, R.; Axline, S.G.; Buchanan, B.G.; Green, C.; Cohen, S.N. Computer-based consultations in clinical therapeutics: Explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research* **1975**, *8*, 303–320. https://doi.org/https://doi.org/10.1016/0010-4809(75)90009-9.
8. Haug, P.J.; Rocha, B.H.; Evans, R. Decision support in medicine: lessons from the help system. *International Journal of Medical Informatics* **2003**, *69*, 273–284. Working Conference on Health Information Systems, https://doi.org/https://doi.org/10.1016/S1386-5056(02)00110-7.
9. Elhaddad, M.; Hamam, S. AI-Driven clinical decision support systems: an ongoing pursuit of potential. *Cureus* **2024**, *16*.
10. Ferdush, J.; Begum, M.; Hossain, S.T. ChatGPT and clinical decision support: scope, application, and limitations. *Annals of Biomedical Engineering* **2024**, *52*, 1119–1124.
11. Bozyel, S.; Şimşek, E.; Koçyiğit, D.; Güler, A.; Korkmaz, Y.; Şeker, M.; Ertürk, M.; Keser, N. Artificial intelligence-based clinical decision support systems in cardiovascular diseases. *Anatolian Journal of Cardiology* **2024**, *28*, 74.
12. Safdar, S.; Zafar, S.; Zafar, N.; Khan, N.F. Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. *Artificial Intelligence Review* **2018**, *50*, 597–623.
13. Tägil, K.; Bondouy, M.; Chaborel, J.; Djaballah, W.; Franken, P.; Grandpierre, S.; Hesse, B.; Lomsky, M.; Marie, P.; Poisson, T.; et al. A decision support system improves the interpretation of myocardial perfusion imaging. *European journal of nuclear medicine and molecular imaging* **2008**, *35*, 1602–1607.
14. Afrash, M.R.; Erfannia, L.; Amrae, M.; Mehrabi, N.; Jelvay, S.; Nopour, R.; Shanbehzadeh, M. Machine learning-based clinical decision support system for automatic diagnosis of COVID-19 based on clinical data. *Journal of Biostatistics and Epidemiology* **2022**.
15. Susanto, A.P.; Lyell, D.; Widyantoro, B.; Berkovsky, S.; Magrabi, F. Effects of machine learning-based clinical decision support systems on decision-making, care delivery, and patient outcomes: a scoping review. *Journal of the American Medical Informatics Association* **2023**, *30*, 2050–2063.
16. Masood, A.; Naseem, U.; Rashid, J.; Kim, J.; Razzak, I. Review on enhancing clinical decision support system using machine learning. *CAAI Transactions on Intelligence Technology* **2024**.
17. Pinsky, M.R.; Bedoya, A.; Bihorac, A.; Celi, L.; Churpek, M.; Economou-Zavlanos, N.J.; Elbers, P.; Saria, S.; Liu, V.; Lyons, P.G.; et al. Use of artificial intelligence in critical care: opportunities and obstacles. *Critical Care* **2024**, *28*, 113.
18. Wang, L.; Zhang, Z.; Wang, D.; Cao, W.; Zhou, X.; Zhang, P.; Liu, J.; Fan, X.; Tian, F. Human-centered design and evaluation of AI-empowered clinical decision support systems: a systematic review. *Frontiers in Computer Science* **2023**, *5*, 1187299.
19. Liu, Y.K.; Tsai, Y.C. Explainable AI for Trustworthy Clinical Decision Support: A Case-Based Reasoning System for Nursing Assistants. In Proceedings of the 2024 IEEE International Conference on Big Data (BigData). IEEE, 2024, pp. 6502–6509.
20. Arslan, M.; Ghanem, H.; Munawar, S.; Cruz, C. A Survey on RAG with LLMs. *Procedia Computer Science* **2024**, *246*, 3781–3790. 28th International Conference on Knowledge Based and Intelligent information and Engineering Systems (KES 2024), https://doi.org/https://doi.org/10.1016/j.procs.2024.09.178.
21. Piñeiro-Martín, A.; García-Mateo, C.; Docío-Fernández, L.; López-Pérez, M.d.C. Ethical Challenges in the Development of Virtual Assistants Powered by Large Language Models. *Electronics* **2023**, *12*. https://doi.org/10.3390/electronics12143170.
22. Piñeiro-Martín, A.; García-Mateo, C.; Docío-Fernández, L.; del Carmen López-Pérez, M. Ethics Guidelines for the Development of Virtual Assistants for e-Health. In Proceedings of the IberSPEECH 2022, 2022, pp. 121–125. https://doi.org/10.21437/IberSPEECH.2022-25.
23. Piñeiro-Martín, A.; Santos-Criado, F.J.; García-Mateo, C.; Docío-Fernández, L.; López-Pérez, M.d.C. Context Is King: Large Language Models' Interpretability in Divergent Knowledge Scenarios. *Applied Sciences* **2025**, *15*. https://doi.org/10.3390/app15031192.
24. Mehandru, N.; Miao, B.Y.; Almaraz, E.R.; Sushil, M.; Butte, A.J.; Alaa, A. Evaluating large language models as agents in the clinic. *NPJ digital medicine* **2024**, *7*, 84.
25. Ferber, D.; El Nahhas, O.S.; Wölflein, G.; Wiest, I.C.; Clusmann, J.; Leßman, M.E.; Foersch, S.; Lammert, J.; Tschochohei, M.; Jäger, D.; et al. Autonomous artificial intelligence agents for clinical decision making in oncology. *arXiv preprint arXiv:2404.04667* **2024**.
26. Kahn, S.R.; Kerstin de Wit, M. Pulmonary Embolism. *N Engl J Med* **2022**, *387*, 45–57.
27. Konstantinides, S.V.; Meyer, G.; Becattini, C.; Bueno, H.; Geersing, G.J.; Harjola, V.P.; Huisman, M.V.; Humbert, M.; Jennings, C.S.; Jiménez, D.; et al. 2019 ESC Guidelines for the diagnosis and management of acute pulmonary embolism developed in collaboration with the European Respiratory Society (ERS). *European heart journal* **2020**, *41*, 543–603.
28. Chase, L.H. *LangGraph: A graph-based extension for LangChain*.

29. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**, *35*, 24824–24837.
30. Hu, X.; Ru, D.; Qiu, L.; Guo, Q.; Zhang, T.; Xu, Y.; Luo, Y.; Liu, P.; Zhang, Y.; Zhang, Z. RefChecker: Reference-based Fine-grained Hallucination Checker and Benchmark for Large Language Models. *arXiv preprint arXiv:2405.14486* **2024**.
31. Chase, L.H. *LangChain: Framework for developing applications powered by language models*.
32. Team, D.S. Docling Technical Report. Technical report, IBM, 2024, [2408.09869]. https://doi.org/10.48550/arXiv.2408.09869.
33. Huber, J.; Troynikov, A. *ChromaDB: AI-native open-source embeddings store*.
34. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217* **2023**.
35. OpenAI. *OpenAI API*.
36. OpenAI. Hello, GPT-4o, 2024.
37. OpenAI. New Embedding Models and API Updates, 2024.
38. OpenAI. Introducing GPT-4.1 in the API, 2025.