

Evidence-Grounded Large Language Model Reasoning for Patient Safety Using Knowledge Graphs and Retrieval-Augmented Generation: Development and Evaluation Study

Sharare Taheri Moghadam

Iran University of Medical Sciences

Md Shafiqur Rahman Jabin

`mdshafiqur.rahman.jabin@inu.se`

Linnaeus University

Research Article

Keywords: adverse event detection, semantic normalization, ontology engineering, clinical natural language processing, temporal graph neural networks, concept extraction, evidence provenance, clinical decision support

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11197970/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Large language models have shown substantial potential for supporting clinical reasoning, but their use in patient safety applications remains constrained by limited access to structured clinical knowledge, variable evidence grounding, and difficulty tracing generated reasoning to supporting sources. Knowledge graphs can provide structured representations of clinical and safety concepts, while retrieval-augmented generation can incorporate relevant biomedical evidence into model inputs.

Objective

This study aimed to develop and evaluate an integrated framework for evidence-grounded patient safety reasoning that combines a large language model, a patient-safety knowledge graph, and retrieval-augmented generation. The study examined how structured knowledge and evidence retrieval contribute to concept extraction and reasoning performance.

Methods

The framework was developed using adult intensive care data from the Medical Information Mart for Intensive Care (MIMIC)-III and MIMIC-IV. Patient safety concepts were normalized using established clinical terminology and patient safety classifications and represented within a structured knowledge graph. An LLaMA-based language model adapted using low-rank adaptation was used for concept extraction and reasoning. Retrieval-augmented generation incorporated relevant biomedical literature, patient safety resources, clinical guidelines, and drug-related knowledge using biomedical sentence embeddings and a FAISS-based vector index. Retrieved evidence was ranked using cosine similarity and incorporated into the model context together with relevant graph information. Performance was evaluated using F1, area under the receiver operating characteristic curve, and area under the precision-recall curve. Component ablation analyses examined the effects of removing the knowledge graph, low-rank adaptation, and retrieval augmentation. Expert assessment also examined the relevance and usefulness of retrieved evidence and generated explanations.

Results

The full framework achieved an F1 score of 0.89 for concept extraction. Removing the knowledge graph resulted in a reduction of 0.047 in F1 (95% Confidence Interval (CI) 0.015–0.079; $P=.004$), 0.084 in area under the receiver operating characteristic curve (95% CI 0.043–0.125; $P<.001$), and 0.110 in area under the precision-recall curve (95% CI 0.044–0.175; $P=.001$). Removing low-rank adaptation produced smaller, nonsignificant changes in F1, area under the receiver operating characteristic curve, and area under the precision-recall curve. Removing retrieval augmentation produced changes of 0.029, 0.034, and 0.031 for these measures, none of which reached statistical significance. Domain experts judged the retrieved documents to be relevant to the corresponding queries and considered the generated explanations clinically useful and plausible. A preliminary usability assessment indicated that users could navigate the system outputs without substantial difficulty, although they recommended simplifying the interface and prioritizing high-level information.

Conclusions

Integrating structured patient safety knowledge with large language model reasoning measurably improved concept extraction and overall system performance, with the knowledge graph contributing most among the evaluated components. Retrieval-augmented generation added a mechanism to incorporate externally retrieved biomedical evidence into model reasoning, although its contribution was less pronounced in the quantitative ablation analysis. The framework demonstrates the feasibility of combining structured semantic representations and externally retrieved evidence to support traceable, evidence-grounded patient safety reasoning. Further evaluation in multicenter datasets and prospective clinical workflows is required before deployment in routine care.

1. Introduction

1.1 Patient Safety Intelligence and the Need for Evidence-Grounded Reasoning

Patient safety surveillance increasingly depends on integrating heterogeneous clinical information, identifying safety-relevant patterns, and translating these patterns into actionable knowledge [1, 2]. Patient-safety information is distributed across incident reports, electronic health records (EHRs), clinical documentation, medication records, diagnostic information, and other health information systems [2, 3]. Conventional reporting and review processes remain important for learning from safety incidents, but they can be labor-intensive and may fail to capture clinically meaningful relationships distributed across sources and documented in different forms [3, 4]. The development of standardized classifications, taxonomies, and ontologies has consequently become an important component of patient-safety information management because structured representations allow safety information to be consistently categorized, linked, and analyzed [2, 5, 6].

Taheri Moghadam, Hooman and Sheikhtaheri (2022, 2023) demonstrated substantial heterogeneity in the methods used to develop and evaluate patient-safety classifications and ontologies, while also identifying recurring requirements such as content validity, comprehensiveness, logical consistency, and semantic coherence [6, 7]. Their findings highlight an important distinction between merely extracting safety-related information and establishing a computational representation that can support subsequent reasoning. More recent work has similarly emphasized structured taxonomy development as a prerequisite for reliable automated analysis of patient-safety information [5, 7]. This distinction is particularly important for artificial intelligence (AI) systems because an extracted concept has limited clinical value if it cannot be connected to its context, contributing factors, supporting evidence, or potential consequences [1, 7].

The need for structured reasoning is also evident in health information technology (HIT)-related safety research. Patient-safety incidents involving health information systems can arise through complex interactions among technology, people, workflows, organizational conditions, and clinical processes [4, 8, 9]. Jabin (2024) argued for more refined classification and reporting mechanisms for HIT-related incidents, reflecting the limitations of approaches that treat safety events as isolated categories rather than as interconnected sociotechnical phenomena [4, 8, 10–12]. Consequently, patient-safety AI requires mechanisms that move beyond detection toward contextual interpretation and evidence-supported reasoning.

1.2 Large Language Models in Clinical Decision Support

Large language models (LLMs) have substantially expanded AI's potential to process unstructured clinical information and generate clinically meaningful language [13]. LLMs have demonstrated capabilities in medical question answering, summarization, information extraction, and clinical reasoning, with studies such as Med-PaLM demonstrating that large models can encode substantial clinical knowledge and perform strongly on medical examination and question-answering tasks [13, 14]. Reviews of LLM applications in medicine have subsequently documented rapid growth across diagnostic support, clinical documentation, medical information retrieval, education, and decision support [15–18].

Despite these capabilities, the direct use of LLMs in high-stakes clinical settings presents important limitations. LLMs can generate plausible but unsupported statements, demonstrate inconsistent responses, and provide information without transparent links to the evidence underlying a conclusion [17–19]. These limitations are particularly consequential for patient safety, where an apparently plausible explanation not adequately supported by clinical evidence may increase rather than reduce risk. Recent systematic evidence indicates that reliability, interpretability, consistency, bias, and factual correctness remain major concerns in healthcare applications of conversational LLMs [15, 18, 19].

Clinical decision support has historically addressed some of these challenges by connecting patient-specific information with explicit clinical knowledge, guidelines, and decision rules [20–22]. However, conventional clinical decision support systems (CDSSs) also have recognized limitations, including excessive alerts, limited contextual sensitivity, workflow disruption, and insufficient transparency [21, 22]. Effective clinical AI requires more than high predictive or generative performance [19, 23]. It requires mechanisms that connect recommendations to patient context, clinical knowledge, evidence, and appropriate human oversight [21, 24].

1.3 Knowledge Graphs and Ontology-Grounded Clinical Reasoning

Knowledge graphs (KGs) offer a complementary approach by explicitly representing clinical entities and their relationships [25, 26]. Rather than encoding knowledge only in model parameters, KGs can represent relationships among medications, clinical conditions, safety events, contributing factors, outcomes, and evidence sources in a structured, machine-readable form [25, 27]. This structure can support semantic integration, graph-based reasoning, knowledge retrieval, and explanation generation [25, 26, 28].

The value of knowledge representation is particularly apparent in medication and patient-safety research [29, 30]. Beierle et al. (2017) demonstrated how ontological modeling can represent co-medication patterns, drug-drug interactions, contraindications, and interaction severity within a clinical decision-support context [29]. Hochheiser et al. (2020) further proposed a minimal information model for potential drug-drug interactions that incorporates contextual information and evidence transparency [31]. Similarly, Kamdar and Musen (2017) demonstrated how mechanism-based pharmacovigilance can combine adverse-event information with biomedical knowledge to support mechanistic reasoning [30].

More recent knowledge-graph approaches have expanded this concept to heterogeneous biomedical evidence. Taneja et al. (2023) developed an ontology-grounded knowledge graph integrating biomedical ontologies, linked data, and scientific literature to support mechanistic investigation of pharmacokinetic natural product-drug interactions [32]. Importantly, their evaluation demonstrated both congruent and contradictory information within the graph, illustrating that knowledge integration alone does not guarantee reliable reasoning and that evidence quality and provenance must remain explicit [26, 32]. Knowledge graphs have consequently become increasingly relevant to explainable AI because they can expose relationships and reasoning pathways that are difficult to obtain from opaque model representations [25, 26].

The patient-safety literature supports the same principle. Patient-safety classifications and ontologies can provide a semantic backbone for integrating heterogeneous safety concepts, while systematic taxonomy-development research emphasizes explicit conceptual definitions, logical validation, and meaningful relationships among concepts [5–7]. These principles provide an important foundation for connecting LLM-generated interpretations with structured safety knowledge.

1.4 Retrieval-Augmented Generation for Evidence-Grounded Clinical AI

Although knowledge graphs provide structured representations of relationships, clinical knowledge is dynamic and continuously expanded through scientific publications, clinical guidelines, safety reports, and other evidence sources [3, 33]. Retrieval-augmented generation (RAG) was introduced to address a related limitation of language models by allowing them to retrieve information from external knowledge sources at inference time rather than relying exclusively on information encoded in model parameters [34]. Subsequent work has positioned RAG as a mechanism to improve factual grounding, reduce dependence on static model knowledge, and incorporate domain-specific or newly available information into LLM outputs [33, 35].

These characteristics are particularly relevant to healthcare. Clinical information changes rapidly, guidelines are updated, and safety knowledge continues to evolve [33]. A system that generates patient-safety reasoning solely from pretrained model parameters may not adequately reflect current evidence [17, 36]. RAG can instead connect clinical inputs with relevant biomedical literature, guidelines, and safety resources before generating a response [34]. In the present framework, the retrieval component was designed to use biomedical literature, patient-safety publications, clinical guidelines, and relevant drug information. Retrieved passages are ranked according to semantic similarity and incorporated into the LLM context to provide external evidence for subsequent extraction and reasoning [33–35].

However, RAG itself does not guarantee evidence-grounded reasoning [36]. Retrieval systems can return irrelevant or incomplete information, while LLMs may fail to use retrieved evidence appropriately or may generate statements that extend beyond the available evidence [36]. Recent research on healthcare RAG has consequently emphasized the need to distinguish answer correctness from retrieval quality, evidence faithfulness, and provenance [36, 37]. Evaluation research has also identified substantial variation in how retrieval, grounding, evidence linkage, and human assessment are currently performed in healthcare RAG systems [36, 37].

This issue is especially important for patient safety. The objective should not simply be to retrieve information that resembles the clinical input. Retrieved evidence must be relevant to the safety problem, semantically connected to the underlying patient-safety concepts, and sufficiently transparent to allow users to understand why a particular conclusion was generated [26, 38]. Combining RAG with structured knowledge representation offers a potentially stronger approach than either mechanism alone.

1.5 Research Gap and Study Objective

The existing literature demonstrates substantial progress in three related areas: automated extraction of clinical and patient-safety information, ontology- and KG-based representation of clinical knowledge, and LLM-based generation supported by external retrieval [6, 7, 25, 36, 39]. However, these components have often been evaluated independently [36]. Patient-safety Natural Language Processing (NLP) systems have primarily focused on classification or concept extraction, while knowledge graphs have emphasized representation and relational reasoning, and RAG-based clinical systems have frequently focused on question answering or general decision support [1, 6, 7, 25, 36]. Integrating these components specifically for evidence-grounded patient-safety reasoning remains comparatively underdeveloped [25, 36].

The project underlying the present study was designed to address this gap by integrating an LLM with a patient-safety knowledge graph and a retrieval layer containing biomedical and clinical evidence. The framework uses structured patient-safety concepts as an intermediate representation and retrieves relevant evidence to support downstream reasoning. Whereas the first study examines whether an LLM can accurately identify patient-safety concepts from clinical narratives, the present study examines whether those concepts can be connected to structured knowledge and external evidence to support more traceable and clinically plausible reasoning. The project documentation explicitly identifies RAG as a mechanism for grounding LLM outputs in external evidence and linking generated explanations to source documents.

Accordingly, this study aimed to develop and evaluate an evidence-grounded patient-safety reasoning framework that integrates an LLM, a patient-safety knowledge graph, and retrieval-augmented biomedical evidence. The study specifically examined whether the integrated approach could retrieve relevant supporting evidence and generate patient-safety reasoning that was clinically plausible, semantically grounded, and traceable to external knowledge sources. By linking structured safety concepts with retrieved evidence, the framework seeks to establish a computational pathway from clinical information to structured safety knowledge, evidence retrieval, and explainable reasoning.

2. Methods

2.1 Study Design and Overall System Architecture

This study describes the development and evaluation of an evidence-grounded patient-safety reasoning framework integrating a patient-safety knowledge graph, retrieval-augmented generation (RAG), and a large language model (LLM). The framework was designed to support interpreting patient-safety information from electronic health records (EHRs) by combining structured clinical knowledge with externally retrieved biomedical evidence. The central design principle was to separate structured semantic representation from external evidence retrieval and natural-language reasoning. The resulting architecture consisted of five principal stages: extracting and normalizing clinical information, constructing a patient-safety knowledge graph, retrieving relevant external evidence, integrating the graph and retrieved evidence into the LLM context, and generating an evidence-grounded response.

The framework was designed for inpatient patient-safety scenarios, with particular relevance to intensive care settings. The system was intended to identify clinically relevant safety concepts, connect those concepts through structured relationships, retrieve supporting biomedical literature or clinical guidance, and use the combined information to generate interpretable reasoning. The project documentation specifies LLaMA-based language modeling with LoRA fine-tuning, knowledge-graph representations, and external evidence retrieval as the principal components of the overall architecture.

Figure 1. Overall framework for evidence-grounded patient-safety reasoning.

Figure 1 illustrates the end-to-end architecture linking clinical information to structured patient-safety concepts, knowledge-graph representation, external evidence retrieval, LLM-based reasoning, and evidence-grounded output.

The system was developed as a modular architecture so the knowledge representation, retrieval, and generation components could be evaluated independently and as an integrated pipeline. This design also allowed the retrieved evidence to remain externally traceable rather than relying exclusively on information encoded within the language model.

2.2 Clinical Data Sources and Study Population

The clinical component of the framework used publicly available, de-identified EHR datasets from MIMIC-III and MIMIC-IV. These databases contain structured and unstructured information from critical care and hospital admissions, including demographic information, clinical notes, laboratory measurements, medications, procedures, and diagnostic information [40, 41]. The project documentation indicates that more than 50,000 patient stays were initially available across the datasets, with a final analytical cohort of approximately 30,000 patients after application of the study-specific filtering criteria.

The study population was restricted to adults aged 18 years or older with at least one complete admission. We excluded admissions without essential clinical information, such as relevant clinical notes or key variables required for the analysis. When multiple ICU admissions were available for the same patient, we selected the first ICU stay to reduce repeated observations from the same individual.

The clinical information extracted from the EHRs included demographic characteristics, vital signs, laboratory results, medication information, procedures, diagnostic codes, and relevant narrative documentation. These data were used to construct structured patient representations and to identify patient-safety concepts that could subsequently be linked to the knowledge graph and retrieval system.

Because the RAG component was designed to provide external biomedical context, the clinical datasets were supplemented by a separate corpus of biomedical literature and patient-safety evidence. The project documentation identifies PubMed abstracts and safety-related documents as sources for the retrieval corpus.

2.3 Patient-Safety Concept Model and Semantic Normalization

Clinical information was transformed into structured patient-safety concepts before construction of the knowledge graph. The semantic representation was designed to capture clinically meaningful entities and relationships relevant to patient safety, including patients, conditions, medications, laboratory findings, adverse events, procedures, and related safety concepts.

The International Classification for Patient Safety (ICPS) was used as the primary patient-safety conceptual framework. Clinical concepts were additionally normalized using established biomedical terminologies, including Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) and the Unified Medical Language System (UMLS), while medications were represented using RxNorm where applicable. This combination provided a structured semantic layer for linking heterogeneous clinical information to patient-safety concepts.

Normalization reduced terminology variation across clinical notes and structured EHR fields. Synonymous or closely related clinical expressions were mapped to standardized concepts where an appropriate mapping was available. The resulting representation allowed clinically related entities to connect regardless of the precise wording used in the original clinical documentation.

The semantic layer was designed to support both patient-specific representations and broader domain-level relationships. Patient-level concepts could be connected to diagnoses, medications, laboratory abnormalities, procedures, and adverse events, while domain-level concepts could represent established relationships between conditions, treatments, hazards, and safety outcomes.

2.4 Patient-Safety Knowledge Graph Construction

A patient-safety knowledge graph (KG) was constructed to provide a structured representation of entities and relationships relevant to clinical safety reasoning. The KG contained nodes representing entities such as Patient, Condition, Medication, LabResult, AdverseEvent, and Procedure, together with typed relationships connecting these entities. Each node was associated with its corresponding semantic class and relevant attributes.

Figure 2. Patient-safety knowledge graph architecture.

Figure 2 illustrates the knowledge graph structure, including clinical entities, patient-safety concepts, semantic relationships, and links to supporting evidence and provenance.

The knowledge graph was initially implemented using NetworkX for prototyping and graph manipulation, with Neo4j identified as a persistence and query environment for subsequent deployment. The project documentation also identifies RDF and SPARQL as alternative representations for semantic implementation. The graph architecture was designed to support structured querying, semantic relationships, and integration with the retrieval and reasoning components.

Relationships within the graph represented clinically meaningful associations, such as medication–adverse-event relationships, condition–risk relationships, laboratory abnormality–safety-event relationships, and patient–clinical-event relationships. Where relationships originated from external evidence, provenance information could be retained to facilitate subsequent evidence tracing.

The graph was also designed to support periodic updating as new evidence, patient information, or ontology releases become available. The project documentation describes the intended ability to incorporate new literature retrieved through RAG and to add new patient-specific information as additional graph nodes or relationships.

2.5 Biomedical Evidence Corpus and Retrieval-Augmented Generation

RAG was incorporated to provide external biomedical evidence and current clinical knowledge during patient-safety reasoning. Rather than relying solely on information encoded during LLM training, the system retrieved relevant documents from a curated external evidence corpus and supplied the retrieved material to the language model as contextual information. The project documentation describes using biomedical literature and clinical guidance to ground model outputs in external evidence.

The retrieval corpus included relevant PubMed abstracts, patient-safety reports, and clinical guidance. The evidence collection was designed to cover patient-safety events, Adverse Drug Events (ADEs), common clinical conditions, and clinically relevant hospital and intensive care scenarios. The corpus could also incorporate authoritative clinical and drug-related information where relevant to the patient-safety query.

Clinical query segments and literature passages were represented using biomedical sentence embeddings. We used BioBERT- or ClinicalBERT-based representations, or comparable biomedical Sentence-BERT embeddings, to capture semantic similarity between clinical information and candidate evidence documents. A local FAISS vector index supported similarity-based document retrieval. The indexed corpus contained embeddings derived from curated biomedical documents, including PubMed abstracts and safety-related evidence.

Figure 3. Retrieval-augmented generation pipeline for patient-safety reasoning.

Figure 3 illustrates the sequence from clinical query representation and evidence retrieval through evidence ranking, context assembly, LLM reasoning, and generation of an evidence-grounded response.

For each patient-safety query, the relevant clinical information was transformed into a retrieval representation and compared with indexed evidence using vector similarity. Candidate passages were ranked according to their similarity to the query. The retrieval configuration used a top-*k* strategy, with the number of retrieved passages determined during system configuration. Retrieved material could also be filtered by source characteristics and publication information before being passed to the generation component.

The highest-ranked evidence passages were incorporated into the LLM prompt as contextual information. The prompt combined the clinical query, relevant patient-safety concepts, knowledge-graph information, retrieved evidence, and task-specific instructions. This allowed the language model to generate a response conditioned on both structured semantic information and externally retrieved evidence.

Importantly, we treated RAG as an evidence-grounding mechanism rather than a guarantee of factual correctness. Retrieval quality and evidence relevance were evaluated separately from the quality of the generated response. This distinction aligns with

contemporary health care RAG evaluation approaches, which emphasize distinguishing retrieval quality, evidence linkage, end-to-end output quality, and safety-related behavior.

2.6 Integration of Knowledge Graph Context and Retrieved Evidence

The KG and RAG components were integrated before generation. The knowledge graph supplied structured information about clinically relevant entities and their relationships, whereas the retrieval system supplied external textual evidence supporting or contextualizing those relationships.

For each reasoning task, the system constructed a unified context containing the relevant patient-specific information, normalized patient-safety concepts, graph-derived relationships, and retrieved evidence passages. Where available, the system retained evidence provenance and source information within the contextual representation.

The combined context was provided to the LLM together with instructions specifying the reasoning task. The model was prompted to use the retrieved evidence and structured graph information when formulating its response and to avoid unsupported conclusions when available evidence was insufficient.

This architecture allowed the KG to function as a structured semantic layer while RAG provided dynamically retrieved external knowledge. The two components were thus complementary rather than interchangeable. The KG represented relationships explicitly, whereas RAG supplied supporting evidence from external sources.

2.7 Large Language Model and LoRA Adaptation

A LLaMA-based LLM served as the primary generation and reasoning model. The project incorporated parameter-efficient fine-tuning using Low-Rank Adaptation (LoRA) to adapt the model to the patient-safety domain while reducing the computational requirements of full model fine-tuning. The project documentation identifies LLaMA fine-tuning with LoRA as a central component of the system architecture.

The model received a structured prompt containing the clinical query, normalized safety concepts, relevant knowledge-graph information, and retrieved biomedical evidence. The generation task was designed to produce clinically interpretable reasoning rather than an isolated classification label. Depending on the task, the output could include identifying a potential patient-safety issue, explaining relevant contributing factors, interpreting supporting evidence, and providing appropriately qualified clinical considerations.

The model was not treated as an autonomous clinical decision-maker. The generated response was designed to support interpretation and patient-safety reasoning, with human clinical oversight retained in the proposed workflow.

2.8 Evaluation Framework

Evaluation focused on the principal components of the proposed system. The purpose was to assess retrieval, evidence grounding, response quality, and human interpretability separately rather than relying on a single end-to-end performance measure. Table 1 defines the principal patient-safety entities in the knowledge graph and their corresponding structured attributes.

Table 1
Patient-safety concepts and knowledge graph representation

Patient-safety concept	Knowledge graph representation	Example attributes
Patient	Patient node	Demographics, admission information
Condition	Condition node	Diagnosis, clinical status
Medication	Medication node	Drug, dose, route
Laboratory result	LabResult node	Test, value, reference range
Adverse event	AdverseEvent node	Event type, severity, timing
Procedure	Procedure node	Procedure type, timing
Safety relationship	Typed edge	Causal, temporal, associative, or contextual relationship

Retrieval evaluation focused on whether the evidence returned for a patient-safety query was relevant to the clinical concepts and reasoning task. We assessed retrieved documents for topical relevance, clinical applicability, and their relationship to the patient-safety issue represented in the query. Table 2 summarizes the principal categories of external knowledge sources incorporated into the retrieval corpus.

Table 2
Data sources used for evidence retrieval

Source type	Purpose within the system
PubMed biomedical literature	Retrieval of peer-reviewed biomedical evidence
Patient-safety reports	Retrieval of safety-event and incident-related evidence
Clinical guidelines	Retrieval of evidence-based clinical recommendations
Drug-related knowledge sources	Retrieval of medication-related safety information
Other authoritative clinical sources	Supplementary evidence where relevant

Evidence-grounding evaluation examined whether the generated reasoning was supported by the retrieved evidence. This assessment considered the relationship between claims made in the generated response and the retrieved passages used as supporting context. Retrieval relevance and evidence-groundedness were evaluated separately because relevant retrieved documents do not necessarily mean the generated response accurately reflects them. This distinction is increasingly recognized as important in health care RAG evaluation.

2.9 Knowledge Graph and Retrieval Configuration

The KG schema and retrieval configuration were specified before evaluation to provide consistent system behavior across test cases. Table 3 specifies the principal structural components of the patient-safety knowledge graph.

Table 3
Knowledge graph schema

Component	Description
Entity type	Patient, Condition, Medication, LabResult, AdverseEvent, Procedure, and related concepts
Ontology mapping	ICPS, SNOMED CT, UMLS, and RxNorm where applicable
Relationship type	Clinical, safety-related, temporal, or contextual association
Provenance	Source or evidence associated with a relationship where available
Patient context	Patient-specific entities and relationships derived from EHR information

Table 4 summarizes the principal retrieval and generation configuration used by the framework.

Table 4
Retrieval and generation configuration

Component	Configuration
Query representation	Biomedical semantic embedding
Embedding approach	BioBERT/ClinicalBERT-based or comparable biomedical embeddings
Vector index	FAISS
Retrieval strategy	Similarity-based retrieval
Retrieved context	Top-ranked relevant passages
Generation model	LLaMA-based LLM with LoRA
Generation input	Clinical context + KG context + retrieved evidence + task instructions
Evidence attribution	Source information retained with retrieved evidence

2.10 Comparative and Ablation Evaluation

Where comparative evaluation was performed, the complete system was compared with appropriately defined reduced configurations to determine the contribution of individual system components. These configurations could include removing knowledge-graph context, removing retrieval augmentation, and using a non-adapted language model.

The comparison was designed to isolate the contribution of individual architectural components while keeping the underlying clinical task and evaluation data consistent. Performance measures were selected according to the specific task being evaluated. For concept extraction, classification-oriented measures such as precision, recall, and F1 score could be used. For predictive components, discrimination measures such as AUROC or area under the precision-recall curve could be used where applicable.

The numerical outcomes of these comparisons are reserved for the Results section and are not presented here.

2.11 Human Expert Evaluation

Human evaluation assessed the clinical relevance and usefulness of system-generated reasoning and retrieved evidence. Domain experts reviewed representative system outputs, including the extracted patient-safety concepts, retrieved evidence, and generated explanations. The project documentation indicates that a patient-safety officer evaluated explanation usefulness and that experts assessed the relevance of retrieved documents.

The human assessment was intended to complement automated or computational evaluation by examining whether retrieved evidence was clinically relevant and whether generated explanations helped users understand potential patient-safety issues. The current evaluation did not include a formal prospective clinical user study unless separately documented. The project documentation explicitly distinguishes preliminary usability checks from a future formal user study.

2.12 Error Analysis

Error analysis examined cases in which the system produced incorrect, insufficiently supported, or clinically inappropriate output. Errors were categorized according to the component in which they arose, including concept representation, knowledge-graph relationships, evidence retrieval, evidence grounding, and final response generation.

Particular attention was given to cases in which relevant evidence was not retrieved, retrieved evidence was relevant but insufficient to support the generated claim, or the generated response introduced information not supported by the available context. This component-level analysis was designed to distinguish retrieval failures from generation failures rather than attributing all errors to the LLM. The results of this analysis, including the frequency and characteristics of observed error categories, belong in the Results section.

2.13 Proposed Clinical Deployment Workflow

The framework was additionally organized into a proposed clinical deployment workflow in which patient data are transformed into structured safety information, relevant evidence is retrieved, and evidence-grounded reasoning is presented to an end user.

Figure 4. Proposed workflow for clinical deployment of a knowledge graph-enhanced RAG system for patient safety.

Figure 4 illustrates the proposed transition from clinical data and external knowledge sources through knowledge-graph construction, evidence retrieval, LLM reasoning, clinical interpretation, and feedback for system refinement.

The proposed workflow incorporates continuous monitoring and feedback so that newly available evidence, changes in clinical knowledge, and identified system limitations can inform subsequent updates. This aligns with the project design, which anticipates periodically incorporating new literature, patient information, and ontology updates into the knowledge graph.

The framework is intended to function as a clinical decision-support or patient-safety support system rather than an autonomous decision-making tool. Clinical users remain responsible for interpreting system outputs within the relevant clinical context.

2.14 Reproducibility and Computational Implementation

The computational pipeline was implemented as a modular workflow encompassing clinical data preprocessing, semantic normalization, knowledge-graph construction, evidence indexing, retrieval, context assembly, LLM inference, and evaluation. The project documentation indicates that fixed random seeds, software-version records, scripts, and configuration files were maintained to support reproducibility.

The retrieval corpus, embedding process, vector index configuration, knowledge-graph schema, model configuration, and evaluation procedures were treated as explicit components of the experimental configuration. This structure was intended to allow inspection and reproduction of individual components independent of the final generation model.

2.15 Summary of the Methodological Framework

The methodological framework combined structured patient-safety knowledge representation with external evidence retrieval and LLM-based reasoning. Clinical information from MIMIC-III and MIMIC-IV was normalized into patient-safety concepts and represented within a knowledge graph. External biomedical evidence was indexed using biomedical semantic embeddings and retrieved using a FAISS-based vector search system. Retrieved evidence was combined with knowledge-graph context and clinical information before being supplied to a LoRA-adapted LLaMA-based LLM. Evaluation was structured across retrieval relevance, evidence grounding, generated response quality, comparative component evaluation, and expert assessment. Table 5 defines the evaluation dimensions and the corresponding measures used to assess the framework.

Table 5
Evaluation criteria and metrics

Evaluation component	Evaluation focus	Potential measures
Concept representation	Correct identification of patient-safety concepts	Precision, recall, F1
Evidence retrieval	Relevance of retrieved evidence	Precision@k, Recall@k, relevance assessment
Evidence grounding	Support for generated claims	Expert assessment, evidence-support assessment
Response quality	Clinical plausibility and usefulness	Expert assessment, task-specific metrics
Comparative evaluation	Contribution of individual components	Difference in task-specific performance
Human evaluation	Clinical usefulness and interpretability	Structured expert assessment
Error analysis	Characterization of system failures	Component-level error categories

3. Results

3.1 Study Cohort and System Implementation

The initial MIMIC-III and MIMIC-IV datasets contained more than 50,000 hospital stays. After applying the predefined eligibility criteria, including restricting to adults with eligible ICU admissions and sufficient clinical information, the final analytical cohort comprised approximately 30,000 patient stays. This cohort was used to evaluate the patient-safety concept extraction and integrated reasoning framework.

The implemented framework combined ontology-guided patient-safety concept extraction, knowledge-graph representation, retrieval of external biomedical evidence, and LLM-based reasoning. The RAG component used biomedical sentence embeddings and a FAISS-based local retrieval index containing curated medical literature and patient-safety evidence.

The resulting system was evaluated through model performance comparisons, component ablation analyses, qualitative assessment of retrieved evidence, and expert assessment of generated explanations.

3.2 Patient-Safety Concept Extraction Performance

The complete model achieved an overall F1 score of 0.89 on the annotated concept-extraction evaluation set. This result indicates strong agreement between the system's extracted patient-safety concepts and the reference annotations.

The ablation analyses showed that the knowledge graph contributed substantially. Removing knowledge-graph grounding reduced F1 by 0.047, with a 95% confidence interval of 0.015 to 0.079 and a bootstrap P value of .004. The corresponding McNemar test was also statistically significant ($P = .0293$).

Removal of LoRA fine-tuning produced a substantially smaller difference in F1 of 0.006, with a 95% confidence interval of -0.024 to 0.035 and a bootstrap P value of .680. The McNemar test was not significant ($P = .8565$).

Removal of RAG reduced F1 by 0.029, with a 95% confidence interval of -0.003 to 0.059 and a bootstrap P value of .079. The corresponding McNemar test was not statistically significant ($P = .1901$). These findings indicate that knowledge-graph grounding produced the largest measurable contribution to the evaluated concept-extraction performance.

3.3 Comparative Performance of Model Configurations

The complete model outperformed the corresponding knowledge-graph-ablated configuration across all three reported performance measures. Removing the knowledge graph reduced F1 by 0.047, Area Under the Receiver Operating Characteristic Curve (AUROC) by 0.084, and Area Under the Precision–Recall Curve (AUPRC) by 0.110, with statistically significant differences on all three measures.

The effects of removing LoRA were comparatively small and did not reach statistical significance for F1, AUROC, or AUPRC. Removing RAG reduced F1 by 0.029, AUROC by 0.034, and AUPRC by 0.031. None of these differences reached conventional statistical significance. Overall, the ablation analysis provides the strongest quantitative evidence for the knowledge-graph component's contribution. Table 6 presents the comparative performance of the complete model and the component-ablated configurations.

Table 6
Comparative performance of evaluated model configurations.

Model configuration	$\Delta F1$	95% CI for $\Delta F1$	Bootstrap P	$\Delta AUROC$	95% CI for $\Delta AUROC$	DeLong P	$\Delta AUPRC$	95% CI for $\Delta AUPRC$	Bootstrap P
Full model	Reference	—	—	Reference	—	—	Reference	—	—
Without knowledge graph	0.047	0.015 to 0.079	.004	0.084	0.043 to 0.125	< .001	0.110	0.044 to 0.175	.001
Without LoRA	0.006	−0.024 to 0.035	.680	0.031	−0.009 to 0.068	.107	0.052	−0.015 to 0.116	.103
Without RAG	0.029	−0.003 to 0.059	.079	0.034	−0.006 to 0.073	.084	0.031	−0.036 to 0.098	.375

3.4 Retrieval-Augmented Evidence Evaluation

The retrieval component was evaluated qualitatively to determine whether the external evidence supplied to the LLM was relevant to the corresponding patient-safety concepts and clinical queries. The system used biomedical semantic embeddings to match clinical text with relevant biomedical literature and other curated sources.

Domain experts reviewed retrieved documents and confirmed their relevance to the evaluated patient-safety questions. The project documentation specifically reports that the RAG retrieval component was qualitatively validated through expert assessment of retrieved documents.

The retrieval evaluation consequently demonstrated qualitative evidence relevance, but did not establish quantitative retrieval precision, recall, citation accuracy, or claim-level faithfulness of the generated responses. These measures were not available from the reported evaluation and are not inferred here.

3.5 Evidence-Grounded Patient-Safety Reasoning

The integrated system used retrieved biomedical evidence together with structured patient-safety information to support LLM-based reasoning. RAG was incorporated specifically to provide external evidence and up-to-date clinical knowledge during concept extraction and reasoning.

The resulting outputs linked patient-safety concepts with relevant clinical context and supporting evidence. The system was designed to make the reasoning process more transparent by retaining information about the retrieved context and prompts associated with individual inferences. Figure 5 presents an example output generated by the RAG-based patient-safety reasoning system.

Figure 5. Example output of the RAG system for a patient-safety query.

The example illustrates how the system incorporates retrieved biomedical evidence into the generated response to provide evidence-grounded patient-safety reasoning.

The example demonstrates the system's intended output format, in which the generated reasoning is accompanied by supporting evidence rather than presented solely as an unsupported LLM response.

3.6 Expert Assessment and Preliminary Usability

A patient-safety officer evaluated the generated explanations for their usefulness in interpreting the system's suggestions. Informal interviews also gathered feedback on the system's outputs. The reported feedback was positive, with the expert assessment indicating that the system identified clinically plausible patient-safety issues.

A preliminary usability assessment was conducted using a simulated dashboard. Clinicians navigated the alerts and explanations without difficulty. Feedback from this assessment resulted in recommendations to simplify the interface and prioritize high-level scores. These findings provide preliminary evidence of perceived usefulness and usability. A formal clinical workflow assessment, including quantitative evaluation of time-to-action, had not yet been conducted.

3.7 System Output and Clinical Interpretation

The system could present patient-safety information alongside explanatory information intended to support clinician interpretation. The project implementation included a basic FastAPI demonstration that connected extraction and prediction components to a mock dashboard.

The system was intended for advisory use. Clinicians could review, ignore, or override system-generated alerts and categorizations, with the clinical team making the final decision. Figure 6 presents the resulting system output and its proposed presentation within the clinical decision-support context.

Figure 6. Proposed clinical presentation and use of evidence-grounded patient-safety system outputs.

The figure illustrates how evidence-grounded patient-safety information can be presented to the end user together with explanatory information to support clinical interpretation.

The output is intended to support, not replace, clinical judgment. The current evaluation establishes technical feasibility and preliminary expert acceptability, but does not demonstrate improvement in clinical outcomes.

3.8 Error Analysis

The error analysis identified differences in system behavior after removing individual components. The most pronounced deterioration occurred after removing the knowledge graph, consistent with the statistically significant reductions in F1, AUROC, and AUPRC.

RAG removal produced smaller reductions across the evaluated measures. Although these reductions did not reach statistical significance, the differences consistently favored the complete model. LoRA removal produced the smallest changes in the reported extraction and predictive performance measures. Table 7 summarizes the principal error domains identified during the evaluation.

Table 7
Error analysis of evidence-grounded patient-safety reasoning.

Error domain	Evaluation finding	Interpretation
Patient-safety concept extraction	Full model achieved F1 = 0.89	Strong overall extraction performance
Knowledge-graph grounding	Removing KG reduced F1 by 0.047 and also reduced AUROC and AUPRC	Knowledge-graph grounding made the strongest measurable contribution
LoRA adaptation	Removing LoRA produced small, nonsignificant performance differences	Limited measurable contribution within the evaluated configurations
RAG augmentation	Removing RAG reduced F1 by 0.029, AUROC by 0.034, and AUPRC by 0.031	RAG showed a positive directional contribution, although differences were not statistically significant
Evidence retrieval	Experts judged retrieved documents relevant	Qualitative evidence supported retrieval relevance
Generated reasoning	Patient-safety expert judged explanations useful	Preliminary evidence of interpretability and usefulness
Unsupported generation	No quantitative claim-level assessment available	Generation faithfulness requires dedicated evaluation
Usability	Positive preliminary simulated-dashboard feedback	Formal clinical workflow evaluation remains necessary

The error analysis indicates that the system's principal performance sensitivity was associated with removing structured knowledge-graph grounding. The results also show that retrieval augmentation contributed positively across the reported performance measures, although the observed differences were not statistically significant.

3.9 Overall Findings

The complete framework achieved an F1 score of **0.89** for patient-safety concept extraction. Component ablation demonstrated that knowledge-graph grounding had the largest measurable effect, with statistically significant improvements in F1, AUROC, and AUPRC relative to the corresponding configuration without the knowledge graph.

LoRA removal had small, nonsignificant effects, while removing RAG produced modest reductions across F1, AUROC, and AUPRC that did not reach statistical significance.

The retrieval component was qualitatively validated by expert assessment of evidence relevance, while the generated explanations received positive preliminary assessment from a patient-safety expert.

Taken together, the results demonstrate that the integrated system was technically feasible and that knowledge-graph grounding was the strongest quantitatively supported contributor to system performance. The findings also provide preliminary support for using retrieval augmentation to supply relevant external evidence for patient-safety reasoning, while further quantitative evaluation of retrieval and generation faithfulness remains necessary.

4. Discussion

4.1 Principal Findings

This study developed and evaluated an evidence-grounded patient-safety reasoning framework that integrates a patient-safety knowledge graph, retrieval-augmented generation, and a large language model to support extraction and interpretation of patient-safety concepts from clinical information. The complete system achieved an F1 score of 0.89 for patient-safety concept extraction. The most important quantitative finding was the contribution of knowledge-graph grounding. Removing the knowledge graph reduced F1 by 0.047, AUROC by 0.084, and AUPRC by 0.110, with statistically significant differences across all three measures. In contrast, removing LoRA produced only small, nonsignificant differences, while removing RAG produced modest reductions in F1, AUROC, and AUPRC that did not reach statistical significance. These findings suggest that structured patient-safety knowledge was a more important measurable contributor to system performance than either parameter-efficient adaptation or retrieval augmentation within the evaluated configuration. This observation is consistent with the broader literature emphasizing the importance of incorporating structured clinical knowledge into artificial intelligence systems rather than relying exclusively on statistical representations learned from unstructured data [1, 8, 22, 25, 27, 28, 42].

The findings are also relevant to the growing literature on large language models in healthcare [13, 14, 19]. LLMs have demonstrated strong performance across a wide range of biomedical and clinical language tasks, including medical question answering, clinical knowledge retrieval, and medical reasoning [13, 14, 19, 43]. However, high performance on benchmark tasks does not necessarily translate into reliable clinical decision-making [17, 19, 23]. Recent evaluations have demonstrated weaknesses in guideline adherence, information integration, robustness, and instruction following when LLMs are evaluated under more realistic clinical conditions [23]. The present findings support a more structured approach in which the LLM is embedded within a domain-specific knowledge and evidence environment rather than treated as an autonomous clinical reasoning system.

4.2 Knowledge-Graph Grounding and Patient-Safety Reasoning

The strongest quantitative finding was the effect of removing the knowledge graph. The 0.047 reduction in F1 was statistically significant, and the corresponding reductions in AUROC and AUPRC were also statistically significant. This finding indicates that explicit semantic representation of patient-safety concepts contributed materially to system performance [6, 7].

Knowledge graphs provide a mechanism for representing entities and relationships that may be difficult to preserve in conventional text-based representations [25, 26, 28]. In healthcare, clinical knowledge is inherently relational [25, 44]. Medications are associated with indications and adverse effects, laboratory abnormalities may indicate physiological deterioration, and clinical

conditions can modify the significance of treatments and other safety events [5, 29, 31, 44]. Ontologies and knowledge graphs provide formal mechanisms to represent these relationships and support consistent semantic interpretation [6, 25–27, 45].

This is particularly relevant to patient safety because safety events rarely exist in isolation [7, 10, 46]. An adverse drug event, for example, may involve a medication, patient condition, laboratory abnormality, temporal sequence, and clinical response [7, 31, 32]. A knowledge graph can represent these relationships explicitly and provide a structured semantic layer through which an LLM can interpret the clinical narrative. Similar arguments have been made for knowledge-driven clinical decision support, in which structured representations can improve interoperability, semantic consistency, and reasoning across heterogeneous clinical information [1, 22, 25, 27, 44].

The present results extend this argument by providing an empirical ablation result within an LLM-based patient-safety framework. Rather than simply proposing that a knowledge graph improves interpretability, the analysis showed that removing the graph led to measurable deterioration across multiple evaluation measures. This strengthens the interpretation that knowledge-graph grounding contributed to the observed performance rather than serving merely as an auxiliary visualization or data-management layer.

This finding also aligns with the broader movement toward combining symbolic or structured knowledge with neural language models [47]. Purely neural models can capture complex statistical relationships, but their representations are not necessarily constrained by explicit clinical ontologies. Knowledge-grounded approaches provide an opportunity to combine the flexibility of neural language understanding with explicit domain constraints and relationships [26, 28, 30].

4.3 RAG as an Evidence-Grounding Mechanism

The RAG findings require a more nuanced interpretation. Removing RAG reduced F1 by 0.029, AUROC by 0.034, and AUPRC by 0.031. Although all three differences favored the complete system, none reached statistical significance. RAG should therefore not be described as demonstrating a statistically significant improvement in overall model performance in this study.

Its value may instead lie primarily in the evidence-grounding function that it provides [33, 34, 36]. RAG was intended to supply the LLM with external biomedical evidence and current clinical information that may not be reliably available from the model's learned parameters alone [33–36]. This rationale aligns with the original RAG framework proposed by Lewis et al. (2020) and subsequent work showing the potential of retrieval augmentation to improve factual grounding and access to external knowledge [33–36].

This approach is particularly relevant in healthcare because medical knowledge changes continuously [33, 36]. Clinical guidelines, drug safety information, evidence syntheses, and biomedical literature are updated over time [3, 33]. A static language model cannot contain all current clinical evidence [17, 33, 36]. Retrieval augmentation provides a mechanism to incorporate external information at inference time, rather than requiring model retraining whenever new evidence becomes available [34, 35].

Recent healthcare studies have increasingly applied RAG to medical question answering, clinical information retrieval, summarization, and decision support [33, 36, 37]. However, systematic reviews have also highlighted substantial variation in retrieval architectures and evaluation methods, with no universally accepted framework for evaluating the quality of healthcare RAG systems [35–37]. This matters for interpreting the present study because the qualitative expert assessment showed that retrieved documents were relevant, but it did not establish quantitative retrieval recall, precision, citation correctness, or claim-level faithfulness [36].

The present findings consequently support RAG as a promising evidence-grounding mechanism, but not as a demonstrated solution to hallucination or factual error [36, 37]. Retrieval can fail when the evidence corpus is incomplete, when the query representation is inadequate, or when irrelevant evidence is retrieved [36, 37]. Even when appropriate evidence is retrieved, the LLM may fail to use it correctly or may generate claims that extend beyond the retrieved material [36, 37]. Recent healthcare RAG reviews and evaluations have highlighted these limitations [35, 37].

4.4 Structured Knowledge and External Evidence Are Complementary

An important conceptual implication of this study is that knowledge-graph grounding and retrieval augmentation are not interchangeable forms of grounding. They support clinical reasoning in different ways [47].

The knowledge graph provides structured semantic grounding [25, 26]. It establishes the identity of clinical concepts, their ontology classes, and relationships among patient-safety entities. RAG provides external evidential grounding by retrieving relevant biomedical literature and clinical guidance [33, 34, 36]. These two mechanisms can consequently operate at different stages of reasoning.

This distinction helps explain the observed ablation results. The knowledge graph produced the largest and statistically significant performance contribution because it directly affected semantic representation and concept extraction. RAG produced smaller improvements because its principal role was to provide contextual evidence during reasoning rather than to define the underlying patient-safety ontology.

The complementary relationship between structured knowledge and retrieved evidence is increasingly recognized in clinical AI research [47]. Knowledge graphs can provide structured context and relationships, while retrieval mechanisms can provide access to unstructured and continuously updated evidence [25, 33]. Integrating these approaches may be particularly useful for clinical applications where both semantic consistency and current evidence matter [21, 28, 31, 47, 48].

This distinction is especially important for patient safety. A system may correctly identify that a medication and adverse event are related based on structured knowledge, while RAG can provide the supporting literature describing the circumstances, risk factors, or clinical management associated with that relationship. Together, they can provide a more transparent evidential pathway than either approach alone.

4.5 Interpreting the Limited Effect of LoRA

The relatively small effect observed after removal of LoRA is another important finding. The reduction in F1 was only 0.006 and was not statistically significant. Similar nonsignificant differences were observed for AUROC and AUPRC. These findings do not demonstrate that LoRA is ineffective. Rather, they suggest that its incremental contribution was limited within the specific integrated configuration evaluated in this study.

Parameter-efficient adaptation such as LoRA is attractive for clinical applications because it can adapt large language models to specialized domains without requiring full model fine-tuning [49]. In the present framework, however, LoRA operated alongside ontology-driven annotation, knowledge-graph constraints, and retrieval augmentation. The additional domain information provided by these components may have reduced the marginal contribution attributable specifically to parameter adaptation.

This finding shows why component-level ablation matters when evaluating complex clinical AI systems. A component may be theoretically valuable but provide limited incremental benefit when combined with several other grounding mechanisms. Reporting only the complete system's performance would not reveal this distinction.

4.6 Evidence-Grounded Explanations and Clinical Interpretability

The system was designed to produce reasoning outputs that could be traced to structured patient-safety concepts and retrieved biomedical evidence. The preliminary expert assessment was favorable. A patient-safety officer considered the explanations useful, and informal expert feedback indicated that the system generated clinically plausible patient-safety suggestions. The simulated dashboard assessment also suggested that users could navigate alerts and explanations without difficulty.

These findings are encouraging but should be interpreted cautiously. Explainability in clinical AI is not synonymous with clinical validity [38, 50]. An explanation may be understandable without accurately representing the causal or computational basis of a model output [50, 51]. This distinction has been emphasized extensively in the explainable AI literature [38, 50, 51].

The present framework therefore contributes evidence traceability and structured reasoning support rather than definitive mechanistic interpretability [38, 50]. The system records retrieved context and uses explicit knowledge relationships, making the evidential pathway more inspectable. This is particularly valuable for patient-safety applications because clinicians and safety professionals may need to understand why a potential safety issue was identified before deciding whether further review is warranted.

4.7 Implications for Patient-Safety Practice

Patient-safety surveillance remains challenging because clinically relevant safety signals can be distributed across fragmented documentation, medication records, laboratory results, and other clinical information [1, 3, 52]. Conventional incident reporting systems also depend on recognition and reporting of events, which can result in incomplete capture of safety problems [3, 4, 52]. Automated clinical NLP has consequently been proposed as a mechanism for identifying safety-related information at scale [1, 4].

The present framework offers a potential approach for augmenting such surveillance by connecting unstructured clinical information with structured patient-safety concepts and external evidence. The system does not need to replace existing patient-safety reporting or review mechanisms. Instead, it could serve as an additional surveillance layer that identifies potentially relevant patterns for human review [1, 22].

This distinction matters because current evidence does not support autonomous clinical decision-making by LLM-based systems [17, 19, 23]. Recent evaluations show that LLMs can produce clinically inappropriate recommendations, fail to follow guidelines consistently, and be sensitive to changes in information presentation [23]. The present framework is consequently better positioned as an evidence-grounded patient-safety decision-support system under human oversight rather than an autonomous clinical decision-maker.

This human-supervised approach is consistent with international recommendations emphasizing safety, transparency, accountability, and appropriate governance for AI in healthcare [53, 54]. It also aligns with the principle that appropriately qualified healthcare professionals should retain clinical responsibility when AI systems provide advisory information.

4.8 Responsible AI, Safety, and Governance

Deploying generative AI in patient safety requires consideration beyond predictive accuracy [3, 19, 23]. Healthcare AI systems must address transparency, accountability, privacy, bias, robustness, and appropriate human oversight [53–56].

The framework incorporates several features relevant to these requirements, including explicit semantic representations, evidence retrieval, logging of prompts and retrieved context, and an advisory rather than autonomous role. These features may facilitate auditability and human review [38, 54]. However, they do not by themselves establish that the system is safe for clinical deployment [19, 23, 42, 57].

A particularly important issue is the quality of the evidence corpus. If the retrieval collection contains incomplete, outdated, or inappropriate sources, the system may provide apparently well-supported but clinically inappropriate reasoning [36, 37]. This creates a different risk from conventional LLM hallucination because the generated response may appear authoritative precisely because it is accompanied by retrieved evidence. Recent RAG research has consequently emphasized the need to evaluate source quality, retrieval relevance, and generation faithfulness separately [35–37].

4.9 Strengths and Limitations

The principal strengths of this study are the integration of structured patient-safety knowledge, external biomedical evidence, and LLM-based reasoning within a single framework, together with component-level ablation analysis that allows the contribution of the knowledge graph, LoRA, and RAG to be examined separately. The study also combines quantitative evaluation with qualitative expert assessment of retrieved evidence and generated explanations, providing complementary perspectives on technical performance and clinical relevance. Using MIMIC-III and MIMIC-IV provides access to large-scale clinical data, while established clinical terminologies and patient-safety concepts support semantic consistency. However, several limitations constrain interpretation. The concept-extraction evaluation used a relatively small annotated test set, and the cohort came from adult ICU data, limiting generalizability to pediatric, outpatient, community, and other hospital settings. The RAG evaluation was qualitative and did not quantify retrieval precision, recall, source quality, citation correctness, or claim-level faithfulness [36]. The preliminary usability assessment was conducted in a simulated rather than real clinical workflow, and no prospective study has established effects on clinical decision-making, patient safety outcomes, or clinician workload [19, 23]. In addition, the integrated framework's complexity and computational requirements may pose implementation challenges in resource-constrained healthcare environments. These limitations align with broader concerns about generalizability, workflow integration, and prospective validation of clinical AI systems [23, 35, 37].

4.10 Future Research

The next stage of research should prioritize external and prospective validation. Multi-center evaluation is particularly important because the current system was developed using MIMIC data from adult ICU populations [19, 23, 40, 41]. Differences in documentation practices, patient populations, clinical workflows, terminology, and institutional safety processes may affect both concept extraction and retrieval performance [19, 23].

A second priority should be a dedicated evaluation of the RAG component. Future work should assess retrieval precision and recall, evidence quality, temporal validity of retrieved sources, citation correctness, and claim-level entailment between retrieved evidence and generated statements [36, 48]. Recent systematic reviews have identified the absence of standardized evaluation frameworks as a major limitation of healthcare RAG research [35, 36, 48].

Third, prospective clinician studies should evaluate whether evidence-grounded outputs improve patient-safety surveillance or decision-making. Relevant outcomes should include diagnostic or safety-event recognition, time to review, alert burden, clinician trust, appropriate reliance, and the rate of accepted versus overridden recommendations. Human-centered evaluation matters because clinical AI performance cannot be inferred from algorithmic metrics alone [19, 21, 22].

Fourth, future versions should investigate dynamic knowledge updating. The knowledge graph and evidence corpus could be updated as new guidelines, safety evidence, drug information, and ontology releases become available. Such mechanisms may reduce dependence on static model knowledge and make the system more responsive to changes in clinical evidence [33–35].

Finally, future development should address fairness, governance, privacy, and regulatory requirements alongside technical performance. The WHO has emphasized that safe and equitable AI deployment in health requires governance structures that address accountability, transparency, human oversight, and potential harms [53].

5. Conclusions

This study demonstrates the feasibility of combining a patient-safety knowledge graph, retrieval-augmented generation, and an LLM to support evidence-grounded patient-safety reasoning. The complete system achieved an F1 score of 0.89, and the ablation analysis identified knowledge-graph grounding as the strongest quantitatively supported contributor to performance. Removing the knowledge graph produced statistically significant reductions in F1, AUROC, and AUPRC, whereas removing LoRA produced minimal, nonsignificant changes, and removing RAG produced smaller, statistically nonsignificant reductions.

The findings support a model of clinical AI in which structured domain knowledge and external biomedical evidence complement, rather than replace, LLM-based reasoning. The qualitative relevance of retrieved evidence and positive preliminary expert assessment further support the potential usefulness of this approach for patient-safety applications. At the same time, the results should not be interpreted as evidence that RAG eliminates hallucination or that the system is ready for autonomous clinical use. Prospective, multi-center, human-centered evaluation with dedicated assessment of retrieval quality and claim-level evidence faithfulness is required before establishing clinical effectiveness and safety.

Abbreviations

ADE — Adverse Drug Event

AI — Artificial Intelligence

EHR — Electronic Health Record

F1 — F1 Score

ICPS — International Classification for Patient Safety

KG — Knowledge Graph

LLM — Large Language Model

LoRA — Low-Rank Adaptation

MIMIC — Medical Information Mart for Intensive Care

NLP — Natural Language Processing

SNOMED CT — Systematized Nomenclature of Medicine Clinical Terms

AUROC—Area Under the Receiver Operating Characteristic Curve

AUPRC— Area Under the Precision–Recall Curve

CDSS— Clinical Decision Support System

CI— Confidence Interval

HIT— Health Information Technology

RAG— Retrieval-Augmented Generation

UMLS— Unified Medical Language System

Declarations

Acknowledgments

The authors thank the clinicians and patient-safety experts who contributed to reviewing and assessing the patient-safety concept extraction framework. The authors also acknowledge the developers and maintainers of the publicly available datasets, clinical terminologies, and open-source software resources used in this study.

Ethical Considerations

This study used de-identified, publicly available electronic health record data from the Medical Information Mart for Intensive Care (MIMIC) databases. The data were accessed and used in accordance with the applicable data-use requirements and institutional policies governing the MIMIC databases. No directly identifiable patient information was accessed by the authors. Because the study involved secondary analysis of de-identified data and no direct interaction with human participants, additional institutional ethical approval was not required.

Funding

No specific funding was received for this study. The authors conducted the study independently and report no external financial support for developing, analyzing, or preparing this manuscript.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability

The datasets analyzed in this study are available through the PhysioNet repository, subject to the applicable data access, credentialing, and data-use requirements. The authors did not create or redistribute the underlying patient-level data. Code, model configurations, and other reproducibility materials will be made available in an appropriate public repository upon publication, where permitted by the applicable data-use restrictions.

Authors' Contributions

Sharare Taheri Moghadam: Conceptualization, methodology, data curation, software, formal analysis, investigation, model development, validation, visualization, and manuscript drafting.

Md Shafiqur Rahman Jabin: Conceptualization, methodology, supervision, investigation, interpretation of results, critical revision of the manuscript, and project administration.

Both authors contributed to the study's development and interpretation of the findings, critically reviewed the manuscript, approved the final version, and agreed to be accountable for all aspects of the work.

AI Disclosure

Generative AI tools were used for language editing and figure generation. They were not used to generate or analyze scientific data or to determine, interpret, or draw the study conclusions. The authors reviewed and approved all content and remain responsible for the manuscript.

References

1. Bates DW et al (2021) The potential of artificial intelligence to improve patient safety: a scoping review. *npj Digit Med* 4(1):54
2. Organization WH (2021) Global patient safety action plan 2021–2030: towards eliminating avoidable harm in health care. World Health Organization
3. Organization WH (2024) Global patient safety report 2024. ed. W.H. Organization
4. Jabin MSR (2024) *The need for a refined classification system and national incident reporting system for health information technology-related incidents*. Volume 6–2024
5. Syryllä T et al (2024) Taxonomy development methods regarding patient safety in health sciences – A systematic review. *Int J Med Informatics* 187:105438
6. Taheri Moghadam S, Hooman N, Sheikhtaheri A (2022) Patient safety classifications, taxonomies and ontologies: A systematic review on development and evaluation methodologies. *J Biomed Inform* 133:104150
7. Taheri Moghadam S, Sheikhtaheri A, Hooman N (2023) Patient safety classifications, taxonomies and ontologies, part 2: A systematic review on content coverage. *J Biomed Inform* 148:104549
8. Sittig DF, Singh H (2010) A new sociotechnical model for studying health information technology. *complex Adapt Healthc Syst* 19(Suppl 3):i68–i74
9. Jabin MSR et al (2019) *Identifying and Characterizing the 18 Steps of Medical Imaging Process Workflow as a Basis for Targeting Improvements in Clinical Practice*. in. *IEEE International Conference on Imaging Systems and Techniques (IST)*. 2019. IEEE Xplore
10. Jabin MSR (2026) Understanding the Limits of Patient Safety Classification Systems for Health Information Technology–Related Incidents. *JMIR Hum Factors* 13:e91783
11. Jabin MSR (2026) Why health information technology safety problems remain invisible. *Front Digit Health*, Volume 8–2026
12. Jabin MSR (2026) Why digital health fails silently: a sociotechnical theory of health information technology–related risk. *Front Digit Health*, Volume 8–2026
13. Singhal K et al (2023) Large Lang models encode Clin Knowl 620(7972):172–180
14. Singhal K et al (2025) *Toward expert-level medical question answering with large language models*. 31(3): pp. 943–950
15. Busch F et al (2025) Current applications and challenges in large language models for patient care: a systematic review. *Commun Med* 5(1):26
16. Meng X et al (2024) The application of large language models in medicine: A scoping review. *iScience*, 27(5)
17. Omiye JA et al (2024) *Large language models in medicine: the potentials and pitfalls: a narrative review*. 177(2): pp. 210–220
18. Wang L et al (2024) *Applications and concerns of ChatGPT and other conversational large language models in health care: systematic review*. 26: p. e22769
19. Bedi S et al (2025) *Testing and evaluation of health care applications of large language models: a systematic review*. 333(4): pp. 319–328

20. Cho I, Bates DW (2018) Behavioral Economics Interventions in Clinical Decision Support Systems. *Yearb Med Inf* 27(01):114–121
21. Greenes RA et al (2018) Clinical decision support models and frameworks: Seeking to address research issues underlying implementation successes and failures. *J Biomed Inform* 78:134–143
22. Sutton RT et al (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit Med* 3(1):17
23. Hager P et al (2024) Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nat Med* 30(9):2613–2622
24. Juluru K et al (2021) Integrating AI Algorithms into the Clinical Workflow. *Radiol Artif Intell* 3(6):e210013
25. Cui H et al (2025) A review on knowledge graphs for healthcare: Resources, applications, and promises. *J Biomed Inform* 169:104861
26. Rajabi E (2022) Knowledge Graphs and Explainable AI in Healthcare. *Information* 13(10):459
27. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32(Database issue):D267–D270
28. Rajabi E, Etminani K et al (2021) *Towards a Knowledge Graph-Based Explainable Decision Support System in Healthcare*, in *Public Health and Informatics*, J. Mantas, Editors. SAGE Publications
29. Beierle C et al (2017) *On the Ontological Modelling of Co-medication and Drug Interactions in Medical Cancer Therapy Regimens for a Clinical Decision Support System*. in. *IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*. 2017
30. Kamdar MR, Musen MA (2017) *Mechanism-based Pharmacovigilance over the Life Sciences Linked Open Data Cloud*. *AMIA Annu Symp Proc*, 2017: pp. 1014–1023
31. Hochheiser H et al (2021) *A Minimal Information Model for Potential Drug-Drug Interactions*. Volume 11–2020
32. Taneja SB et al (2023) Developing a Knowledge Graph for Pharmacokinetic Natural Product-Drug Interactions. *J Biomed Inform* 140:104341
33. Yang R et al (2025) Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Syst* 2(1):2
34. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... Kiela, D., *Retrieval-augmented generation for knowledge-intensive nlp tasks*. *Advances in neural information processing systems*, 2020(33): pp. 9459–9474
35. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... Wang, H., *Retrieval-augmented generation for large language models: A survey*. *arXiv preprint arXiv:2312.10997*, 2023
36. Amugongo LM et al (2025) Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digit Health* 4(6):e0000877
37. Liu S, McCoy AB (2025) and A.J.J.o.t.A.M.I.A. Wright, *Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines*. 32(4): pp. 605–615
38. Amann J et al (2020) Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf Decis Mak* 20(1):310
39. Young IJB, Luz S (2019), and N.J.I.j.o.m.i. Lone. *Syst Rev Nat Lang Process Classif tasks field incident Report adverse event Anal* 132:103971
40. Johnson AEW et al (2023) MIMIC-IV, a freely accessible electronic health record dataset. *Sci Data* 10(1):1
41. Johnson AEW et al (2016) MIMIC-III, a freely accessible critical care database. *Sci Data* 3(1):160035
42. Jabin MSR et al (2025) Ethical and Quality of Care–Related Challenges of Digital Health Twins in Care Settings for Older Adults: Scoping Review. *JMIR Aging* 8:e73925
43. Nori H et al (2023) *Capabilities of gpt-4 on medical challenge problems*.
44. Bodenreider O, Cornet R, Vreeman DJ (2018) Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. *Yearb Med Inf* 27(1):129–139
45. Yu H et al (2019) ODAE: Ontology-based systematic representation and analysis of drug adverse events and its usage in study of adverse events given different patient age and disease conditions. *BMC Bioinformatics* 20(7):199

46. Jabin MSR (2026) *A qualitative, multi-framework methodology for analysing health information technology–related patient safety incidents*. Volume 8–2026
47. Youbi L et al (2026) *Exploring Synergies Between Large Language Models and Knowledge Models in Healthcare: A Scoping Review*. : pp. 703–707
48. Liu L et al (2025) Human level information extraction from clinical reports with finetuned language models. *Sci Rep* 15(1):45239
49. HU E, J. and P. WALLIS. *Lora: Low-rank adaptation of large language models [J]*. in *10th International Conference on Learning Representations, OpenReview. net*. (2022)
50. Ghassemi M, Oakden-Rayner L (2021) L.J.T.I.d.h. Beam. false hope Curr approaches explainable Artif Intell health care 3(11):e745–e750
51. Holzinger A et al (2019) Causability explainability *Artif Intell Med* 9(4):e1312
52. Jabin MSR (2024) *Operational disruption in healthcare associated with software functionality issue due to software security patching: a case report*. Volume 6–2024
53. Organization WH (2024) Ethics and governance of artificial intelligence for health: large multi-modal models. WHO guidance. World Health Organization
54. Organization WH (2024) *Artificial Intelligence for Health*, in *World Health Organization*
55. Morley J et al (2021) *The ethics of AI in health care: a mapping review*. : pp. 313–346
56. Topol EJJNm (2019) *High-performance medicine: the convergence of human and artificial intelligence*. 25(1): pp. 44–56
57. Jabin MSR et al (2024) Ethical and Quality of Care-Related Challenges of Digital Health Twins in Older Care Settings: Protocol for a Scoping Review. *JMIR Res Protoc* 13:e51153

Figures

Figure 1. Overall framework for evidence-grounded patient safety reasoning

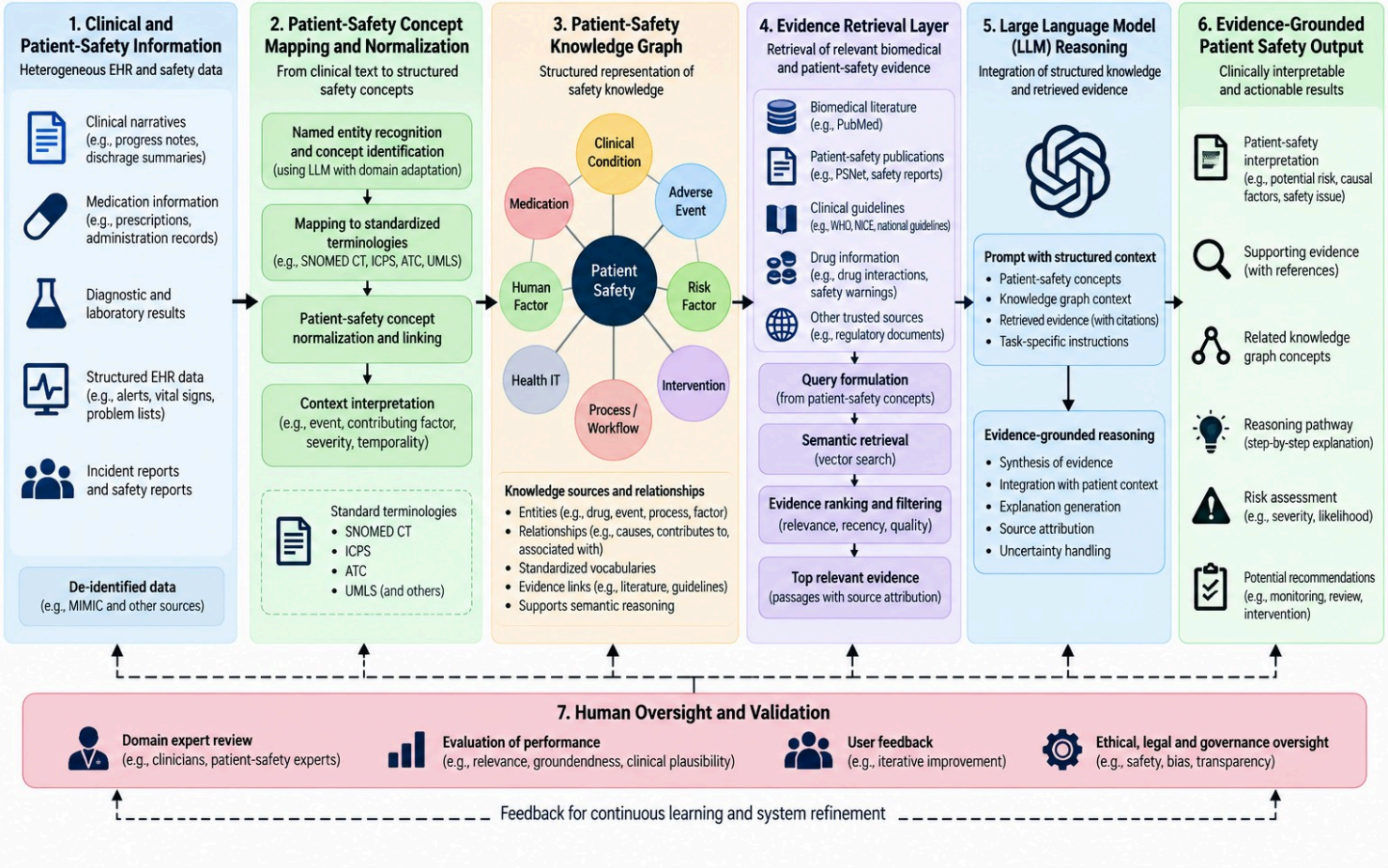


Figure 1
Overall framework for evidence-grounded patient-safety reasoning.

Figure 2. Patient-safety knowledge graph architecture

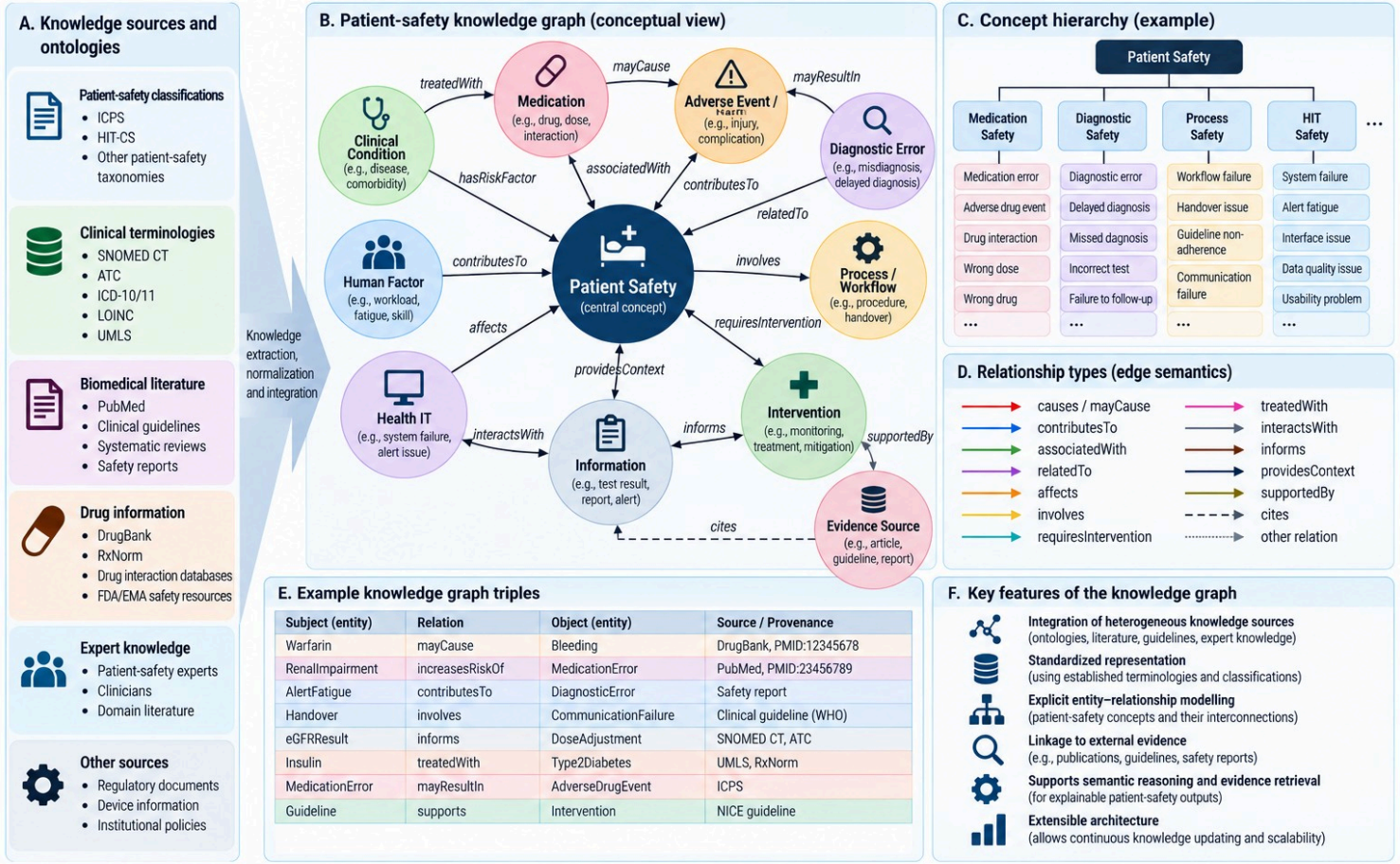


Figure 2

Patient-safety knowledge graph architecture.

Figure 3. Retrieval-augmented generation pipeline for patient-safety reasoning

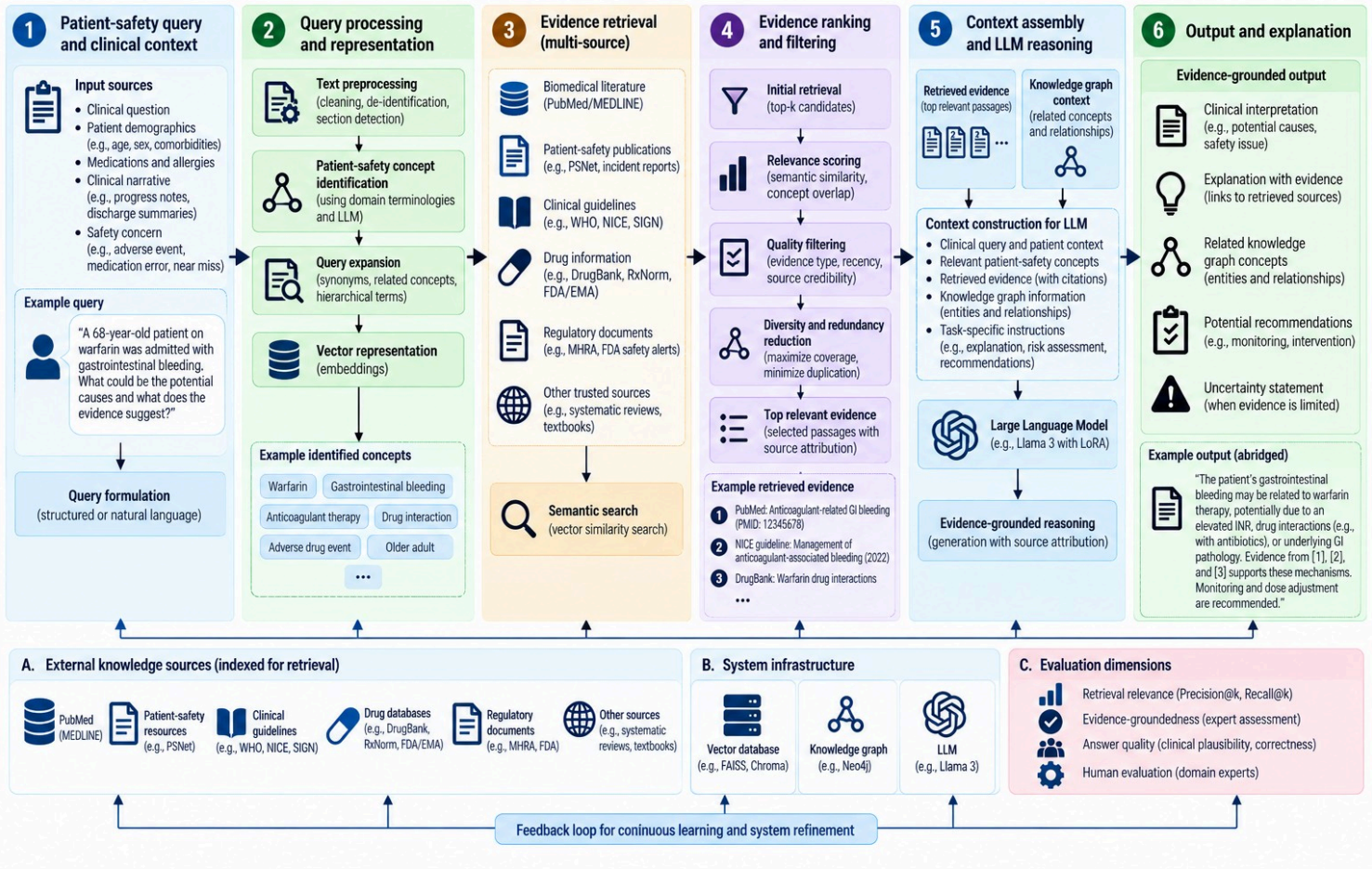


Figure 3

Retrieval-augmented generation pipeline for patient-safety reasoning.

Figure 4. Example of knowledge graph-supported evidence-grounded patient-safety reasoning

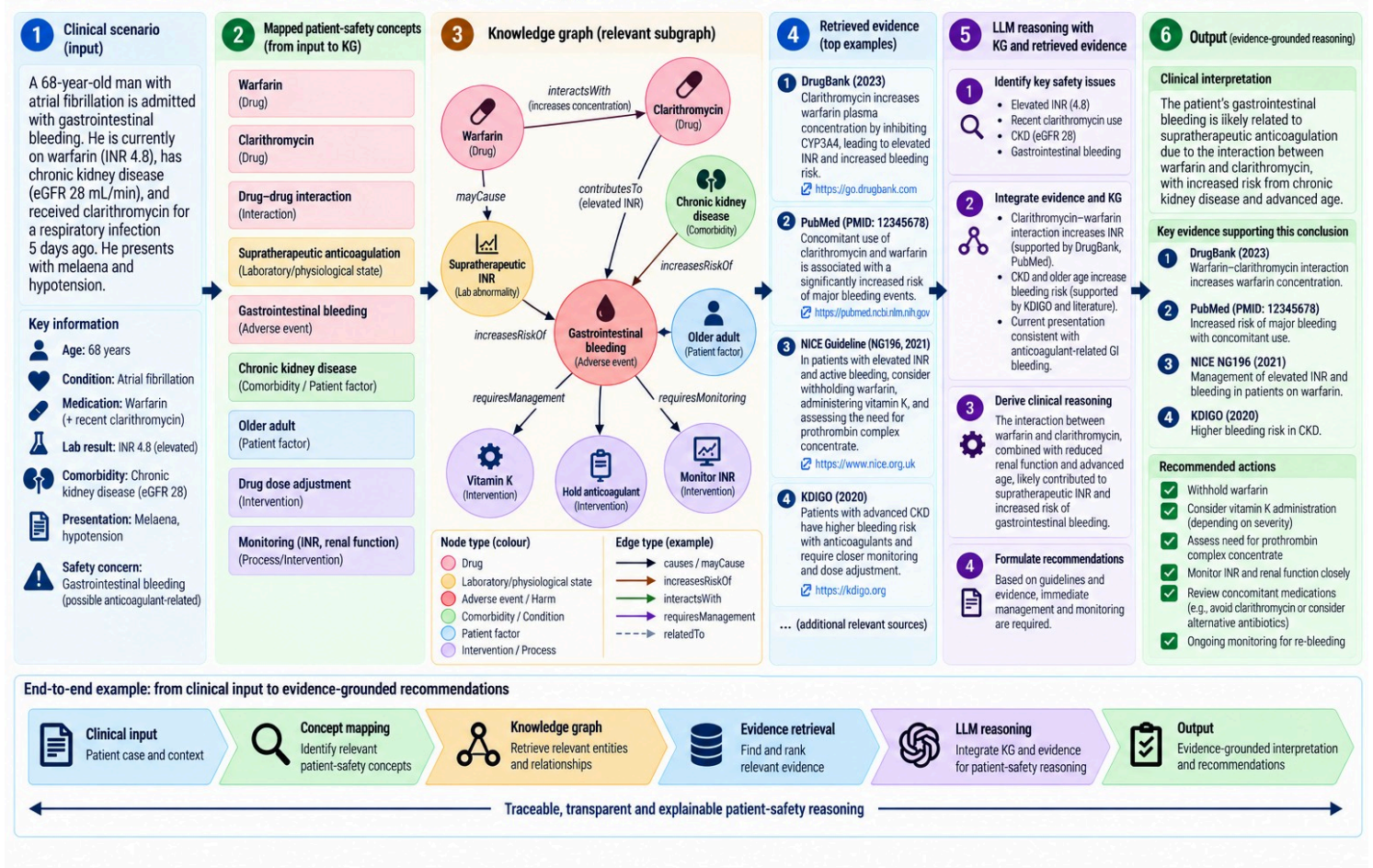


Figure 5. Example output of the RAG system for a patient-safety query

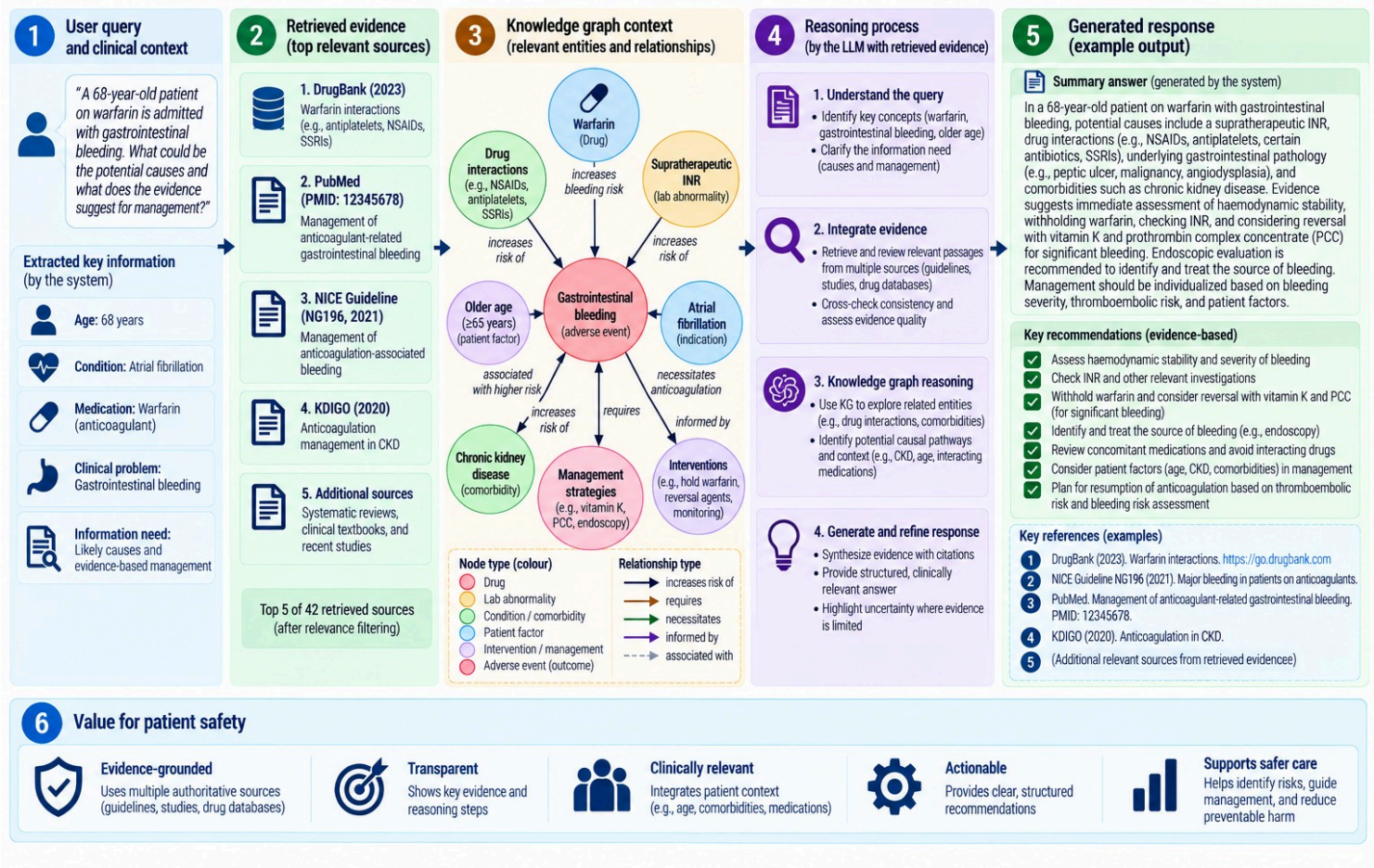


Figure 5

Example output of the RAG system for a patient-safety query.

Figure 6. Proposed workflow for clinical deployment of a knowledge graph–enhanced RAG system for patient safety

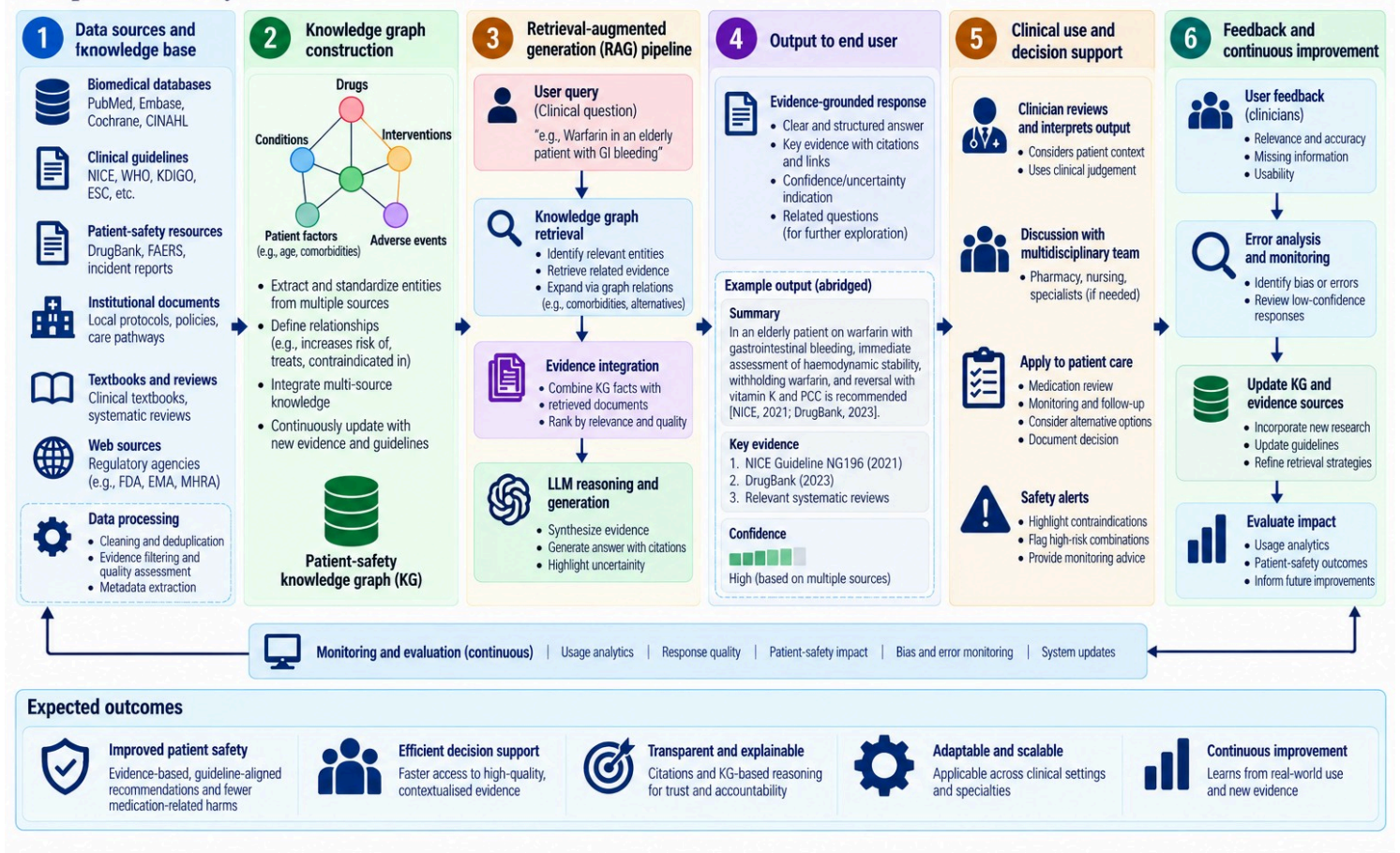


Figure 6

Proposed clinical presentation and use of evidence-grounded patient-safety system outputs.