

Evidence-Grounded Retrieval- Augmented Generation and Knowledge-Graph Grounding for Patient-Safety Reasoning: A Comparative Evaluation

Sharare Taheri Moghadam

Iran University of Medical Sciences

Md Shafiqur Rahman Jabin

`mdshafiqur.rahmanjabin@inu.se`

Linnaeus University

Research Article

Keywords: Clinical informatics, ontology engineering, semantic representation, critical care, human-AI collaboration, safety assurance, patient harm prevention, knowledge representation

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11233939/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 offer promising capabilities for clinical reasoning, but unsupported claims, hallucinations, limited evidence traceability, and inconsistent reasoning constrain their use in patient-safety applications. Retrieval-augmented generation and structured knowledge representations may complement each other to improve output reliability.

Objective

This study evaluated whether grounding a LLaMA-based large language model in retrieved evidence, structured patient-safety knowledge, or both improves the factual correctness, evidence grounding, safety, explainability, and clinical usefulness of patient-safety reasoning while reducing hallucinations and unsupported claims.

Methods

We conducted a controlled comparative evaluation using patient-safety cases derived from critical care data and aligned with established patient-safety and clinical terminology frameworks. The same underlying language model, case inputs, prompting strategy, and output format were evaluated under four conditions: LLM only, LLM with retrieval-augmented generation, LLM with knowledge-graph and ontology grounding, and LLM with combined retrieval and knowledge-graph/ontology grounding. Model outputs were evaluated across six predefined dimensions: factual correctness, evidence grounding, hallucination or unsupported claims, safety, explainability, and clinical usefulness. Pairwise comparisons, confidence intervals, effect sizes, and appropriate statistical tests were used to assess differences between conditions. Expert assessment was also used to examine the relevance and interpretability of generated outputs.

Results

Grounding substantially improved patient-safety reasoning compared with the ungrounded model. The combined condition demonstrated the strongest overall performance across factual correctness, evidence grounding, safety, explainability, and clinical usefulness, while producing the lowest level of hallucination and unsupported claims. Retrieval primarily strengthened the use and traceability of external evidence, whereas knowledge-graph and ontology grounding provided structured semantic context for linking clinical concepts, medications, conditions, procedures, and safety events. Combining the two mechanisms produced more consistent and comprehensive reasoning than either grounding

approach alone. Expert assessment further indicated that grounded outputs were more interpretable and useful for safety-oriented review.

Conclusions

Evidence retrieval and structured knowledge representation provide complementary forms of grounding for large language models used in patient-safety reasoning. Their combination can improve the reliability, transparency, and practical usefulness of model-generated safety assessments. The findings support further prospective evaluation of grounded language models within clinical workflows, with particular attention to claim-level verification, human oversight, workflow integration, and safety assurance.

Introduction

Patient Safety as a Knowledge-Intensive Clinical Reasoning Problem

Patient safety remains a major challenge across health care, with preventable harm arising from complex interactions among patients, clinical conditions, medications, technologies, healthcare professionals, organizational processes, and care environments ^{1, 2}. Patient safety incidents may involve adverse drug events, medication errors, diagnostic problems, treatment-related complications, communication failures, and failures or limitations of health information technology (HIT). These events rarely stem from a single factor and often arise from interactions among multiple clinical and system-level conditions ^{2, 3}. The complexity of these interactions creates a substantial information-processing challenge for clinicians and patient-safety professionals who must interpret heterogeneous clinical information, identify relevant hazards, connect observations to established safety knowledge, and determine whether an observed situation warrants further investigation or intervention.

Electronic health records (EHRs) contain extensive information relevant to patient safety, including clinical narratives, diagnoses, medications, laboratory findings, procedures, and adverse-event documentation. However, these data are heterogeneous and distributed across structured and unstructured representations. Important safety information may be embedded within clinical narratives or represented through different terminologies, making it difficult to establish meaningful relationships between patient characteristics, clinical conditions, treatments, and safety outcomes ^{4, 5}. Previous research has demonstrated persistent challenges in identifying and classifying HIT-related incidents and other patient-safety problems from healthcare data ^{5, 6}.

Traditional patient-safety analysis approaches provide important conceptual and classification structures, but they do not necessarily provide automated reasoning over large volumes of heterogeneous clinical information. Patient-safety taxonomies and standardized terminologies can describe events, contributing factors, and clinical concepts, while clinical knowledge resources provide

evidence concerning medications, diseases, interventions, contraindications, and potential adverse outcomes. The challenge is to connect these forms of structured and unstructured knowledge in a way that supports reasoning rather than merely information retrieval ^{7–9}.

This challenge has become particularly important with the emergence of large language models (LLMs). LLMs can process complex clinical language and generate coherent responses to clinical questions, creating opportunities for automated patient-safety analysis and decision support. However, their ability to generate plausible language does not necessarily mean that their reasoning is factually correct, sufficiently evidence-based, or safe for patient-safety applications ^{10,11}.

Large Language Models for Clinical and Patient-Safety Reasoning

LLMs have demonstrated substantial capabilities in biomedical and clinical language processing. Studies show that large language models can encode substantial clinical knowledge and perform well on medical question-answering tasks ^{10,11}. Their ability to interpret natural language, synthesize information, and generate explanations has stimulated growing interest in their use for clinical decision support and other healthcare applications ^{12–14}.

The potential value of LLMs in patient safety extends beyond conventional question answering. Patient-safety reasoning often requires interpreting multiple pieces of information and recognizing relationships among clinical conditions, medications, procedures, adverse events, and contextual factors. An LLM can potentially integrate these heterogeneous textual inputs and formulate a clinically meaningful explanation. These capabilities may support tasks such as identifying potential safety concerns, interpreting adverse events, generating possible contributing factors, and communicating complex clinical information in an accessible form ^{10,11}.

However, LLMs also introduce important challenges when applied to safety-critical healthcare tasks. Generated responses can contain inaccurate information, unsupported assertions, or recommendations that appear clinically plausible but are not adequately supported by established evidence. These limitations are particularly concerning when an output is presented with high confidence or when users cannot determine the basis for a generated conclusion. Research evaluating LLMs in clinical decision-making has shown limitations in guideline adherence, robustness, reasoning consistency, and clinical reliability, highlighting that strong language-generation capability does not automatically translate into safe clinical decision support ^{15,16}.

This problem matters for patient-safety reasoning because a useful system must do more than produce a plausible answer. It should connect its reasoning to relevant patient-safety concepts, identify appropriate supporting evidence, distinguish established knowledge from uncertainty, and provide outputs that human users can examine. Explainability and transparency are consequently important

considerations in healthcare AI, although existing approaches to explainability themselves have substantial limitations^{17–19}.

One potential strategy is to provide LLMs with additional sources of structured knowledge and external evidence rather than relying exclusively on information encoded within model parameters.

Knowledge Graphs and Ontology-Grounded Patient-Safety Knowledge

Knowledge graphs explicitly represent entities and relationships and can provide LLMs with structured information that is difficult to preserve through conventional text-only representations. In healthcare, knowledge graphs can connect clinical concepts such as conditions, medications, laboratory findings, procedures, adverse events, risk factors, and interventions through semantically defined relationships^{20–22}. Biomedical ontologies and standardized terminologies further support consistent representation and interoperability across heterogeneous clinical data sources²³.

For patient-safety applications, structured knowledge representation is particularly relevant because safety reasoning frequently depends on relationships rather than isolated concepts^{9, 24}. For example, a medication may be associated with an adverse event under particular clinical circumstances, a laboratory abnormality may modify the risk associated with a treatment, or a patient characteristic may alter the relevance of a known safety concern. Representing such relationships explicitly can provide a semantic layer through which patient information and patient-safety knowledge can be connected.

Knowledge graphs have been explored to improve clinical information integration, semantic search, and clinical decision support. Their principal advantage in LLM-based reasoning is that they provide structured relationships and domain constraints that complement the probabilistic, predominantly text-based representations learned by language models^{20, 25}.

In the present framework, the patient-safety knowledge graph represents concepts including patients, conditions, medications, laboratory findings, adverse events, procedures, and related safety concepts. The project questionnaire specifies the use of ontology-based normalization and standardized clinical terminologies, including the International Classification for Patient Safety (ICPS), Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), Unified Medical Language System (UMLS), and RxNorm, to provide a consistent semantic representation of patient-safety information. The knowledge graph provides a structured representation that supports concept-level reasoning and connects clinical information with relevant patient-safety relationships.

However, structured knowledge alone does not provide access to the full breadth of biomedical evidence required for many patient-safety questions^{20, 26}. Clinical guidelines, biomedical literature, drug information, safety reports, and other external resources are continuously updated. A complementary mechanism is consequently required to retrieve relevant evidence at the time of reasoning.

Retrieval-Augmented Generation for Evidence-Grounded Reasoning

Retrieval-augmented generation (RAG) connects LLM generation with external knowledge sources^{27, 28}. Rather than requiring all relevant information to be encoded in the model's parameters, RAG retrieves relevant documents or passages from an external corpus and incorporates them into the context provided to the LLM²⁸. This approach has attracted substantial interest in healthcare because it can connect generative models with biomedical literature, clinical guidelines, drug information, and other domain-specific knowledge resources^{29, 30}.

In patient-safety applications, RAG offers an important opportunity to ground generated reasoning in externally retrieved evidence^{29, 30}. Relevant evidence can be retrieved from biomedical literature and patient-safety resources and then supplied to the LLM alongside the clinical question or patient context. This can support evidence-linked explanations and provide users with identifiable sources underlying a generated response. Recent systematic reviews of RAG in healthcare indicate growing evidence of its potential to improve the usefulness of LLM applications, while also highlighting substantial variation in retrieval methods, evaluation strategies, and evidence-grounding approaches^{27, 29}.

However, the value of RAG should not be assumed simply from the presence of retrieved documents. Retrieval quality determines whether the model receives appropriate evidence, while the model's subsequent use of that evidence determines whether the generated answer is actually grounded in it. Irrelevant, incomplete, outdated, or poorly ranked evidence may not improve reasoning and may introduce additional noise. Consequently, RAG should be evaluated as an evidence-grounding mechanism rather than an automatic solution to LLMs' reliability limitations^{27, 31}.

The project framework uses biomedical and patient-safety sources including targeted PubMed literature, clinical guidelines, drug-related information, and other relevant safety resources. Evidence is represented as biomedical embeddings and retrieved through a vector-based search process, with candidate passages ranked by semantic similarity and filtered by source and temporal considerations. The system then incorporates the retrieved evidence into the LLM prompt as contextual information. The resulting architecture is designed to allow the LLM to reason over both structured patient-safety knowledge and externally retrieved evidence.

This raises a more fundamental empirical question: does combining these grounding mechanisms actually improve patient-safety reasoning compared with using either mechanism alone?

Integrating Structured Knowledge and Retrieved Evidence

Knowledge graphs and RAG address different limitations of conventional LLM-based reasoning. A knowledge graph provides structured domain knowledge and explicit relationships among clinical and patient-safety concepts, whereas RAG provides access to external textual evidence retrieved according

to the specific question being considered^{25, 26}. Together, they may offer complementary forms of grounding.

An ontology- and knowledge-graph-grounded model can constrain the interpretation of clinical concepts and relationships, while retrieved evidence can provide detailed contextual information supporting those relationships. The resulting reasoning process can be conceptualized as a progression from clinical information to structured patient-safety concepts, from structured concepts to relevant external evidence, and from evidence and structured knowledge to an LLM-generated interpretation^{32, 33}. The project architecture explicitly describes this interaction, with the retrieval component supplying relevant literature to the LLaMA model and the knowledge graph providing structured entities and relationships.

Despite this conceptual complementarity, the incremental contribution of each grounding mechanism requires empirical evaluation. A system that combines multiple components may appear more sophisticated without demonstrating that each component contributes meaningfully to performance. This is particularly important for patient-safety applications, where additional system complexity should be justified by measurable improvements in correctness, evidence grounding, safety, explainability, or clinical usefulness.

A controlled comparison of different reasoning configurations can address this issue directly. Comparing an LLM alone with progressively augmented configurations allows examination of the independent and combined contributions of RAG and structured knowledge grounding. This experimental design also makes it possible to determine whether combining knowledge-graph and evidence grounding provides benefits beyond either component individually.

Research Gap and Study Objective

Current research demonstrates considerable promise for LLMs, knowledge graphs, and RAG in healthcare, but important gaps remain in understanding how these approaches perform when combined specifically for patient-safety reasoning. Existing studies frequently evaluate individual technologies or demonstrate system capabilities without systematically isolating the contribution of different grounding mechanisms. In particular, empirical evidence comparing LLM-only reasoning with reasoning augmented by external evidence, structured patient-safety knowledge, and their combination under the same experimental conditions remains limited^{15, 31}.

This gap is consequential because patient-safety reasoning requires more than general clinical language generation. A useful system should produce factually correct and clinically plausible outputs, appropriately connect claims to supporting evidence, minimize unsupported or unsafe recommendations, and provide explanations that experts can assess and trace^{15, 31}. Evaluation should consequently extend beyond conventional language-model metrics to include factual correctness, evidence relevance and grounding, hallucination or unsupported-claim rates, safety-related errors, explainability, and perceived clinical usefulness.

The present study addresses this gap through a controlled comparative experiment in which the same patient-safety cases are processed under four model conditions: (1) an LLM without additional grounding, (2) an LLM augmented with RAG, (3) an LLM augmented with the patient-safety knowledge graph and ontology, and (4) an LLM integrating both RAG and knowledge-graph/ontology grounding. The study evaluates whether these increasingly grounded configurations produce measurable differences in factual correctness, evidence grounding, safety, explainability, and clinical usefulness.

The primary objective was to determine whether grounding an LLM in structured patient-safety knowledge and externally retrieved evidence improves the quality and safety of patient-safety reasoning compared with an LLM operating without these grounding mechanisms.

The specific objectives were to:

1. compare patient-safety reasoning across the four model configurations under identical case conditions
2. assess factual correctness and clinical plausibility of generated outputs
3. evaluate the relevance and correctness of retrieved supporting evidence
4. quantify unsupported or hallucinated claims and safety-related reasoning errors
5. assess the explainability, traceability, and clinical usefulness of generated responses
6. determine whether combining knowledge-graph grounding with RAG provides additional benefit beyond either approach individually.

The central hypothesis was that the combined LLM + RAG + knowledge-graph/ontology configuration would provide more accurate, evidence-grounded, explainable, and clinically useful patient-safety reasoning than the LLM-only configuration and either grounding approach used independently.

Methods

Study Design

We conducted a controlled comparative experimental study to examine whether grounding a large language model (LLM) with structured patient-safety knowledge and externally retrieved clinical evidence improves patient-safety reasoning. The study was designed to isolate the contributions of two complementary grounding mechanisms: retrieval-augmented generation (RAG), which provides external evidence during inference, and a knowledge graph and ontology (KG/ontology), which provides structured clinical and patient-safety concepts and relationships.

The experimental design compared four conditions using the same patient-safety cases, the same underlying LLM, the same core prompting framework, and the same output requirements. The four conditions were: (A) LLM only, (B) LLM + RAG, (C) LLM + KG/ontology, and (D) LLM + RAG + KG/ontology. This factorial comparison was intended to determine whether each grounding mechanism independently improves reasoning and whether their combination provides additional benefit.

The study was motivated by the limitations of applying LLMs directly to clinical decision support. Although LLMs can encode substantial clinical knowledge, clinical applications require higher standards of factuality, reasoning quality, safety, and transparency than general language-generation tasks¹⁰. Human evaluation frameworks for medical LLMs have consequently emphasized multiple dimensions, including factuality, comprehension, reasoning, potential harm, and bias¹⁰. Clinical decision support systems similarly require explicit consideration of benefits, unintended consequences, implementation risks, and ongoing evaluation¹².

The study focused specifically on patient-safety reasoning. The broader project includes additional components for patient-safety concept extraction, temporal risk prediction, and causal analysis. Those components were not treated as experimental factors in the present comparison. This study instead aimed to determine whether structured patient-safety knowledge and retrieved evidence improve the quality, grounding, safety, and usefulness of LLM-generated reasoning. Figure 1 presents the overall experimental workflow.

Figure 1. Experimental framework for comparative patient-safety reasoning.

Figure 1 presents the overall experimental workflow from patient-safety cases through the four controlled model conditions, generated responses, and subsequent evaluation. The design allows examination of the individual and combined contributions of external evidence and structured knowledge.

Clinical Data Sources

The clinical component of the study used the publicly available Medical Information Mart for Intensive Care (MIMIC)-III and MIMIC-IV critical-care databases. MIMIC-III contains de-identified data from critical-care patients at Beth Israel Deaconess Medical Center, including demographic information, laboratory measurements, medications, procedures, diagnoses, and clinical documentation³⁴. MIMIC-IV provides an expanded electronic health record resource containing hospital and ICU data, including clinical observations, diagnoses, procedures, medications, and de-identified clinical notes³⁵.

The project documentation reports that more than 50,000 hospital stays were processed across MIMIC-III and MIMIC-IV. After restricting the data to adult ICU admissions and applying data-completeness requirements, the final cohort comprised approximately 30,000 patients. Adults aged 18 years or older with at least one complete admission were eligible. Admissions lacking critical information, including required clinical notes or key variables, were excluded. When patients had multiple ICU admissions, we retained the first ICU stay to reduce dependence on repeated observations.

The clinical variables available to the framework included demographics, vital signs, laboratory measurements, medication orders, procedure codes, diagnostic codes, admission and progress notes, and discharge summaries. The knowledge graph mapped clinical concepts to standardized terminology.

The clinical setting was primarily inpatient hospital care, with particular emphasis on intensive care because of the high density and complexity of patient-safety risks in this environment. The patient-safety domains included adverse drug events, medication errors, procedural complications, hospital-acquired conditions, and related safety incidents. Table 1 summarizes the clinical data sources, cohort construction, available clinical information, and patient-safety domains used to establish the evaluation corpus.

Patient-Safety Conceptual Framework and Knowledge Representation

The patient-safety knowledge layer was constructed using a combination of patient-safety classification and standardized clinical terminologies. The International Classification for Patient Safety (ICPS) was used as the primary framework for patient-safety concepts. ICPS provides a structured conceptual framework encompassing safety-related concepts such as incident types, hazards, outcomes, and contributing factors ³⁶.

SNOMED CT represented standardized clinical concepts, UMLS supported integration across biomedical terminologies, and RxNorm represented medication concepts. The project documentation specifies that this terminology structure was embedded within the knowledge graph to normalize and relate clinical and patient-safety concepts.

The knowledge graph represented entities including Patient, Condition, Medication, LabMeasurement, AdverseEvent, Procedure, Hazard, and Outcome. Important relationships included causes, associated_with, has_treatment, is_a, and temporal relationships such as temporally_before and temporally_after. These relationships were intended to represent clinically meaningful connections and support structured reasoning about potential safety events.

The team initially prototyped the knowledge graph using NetworkX. Neo4j was identified as a persistence and query option for larger-scale deployment. The graph was designed to connect heterogeneous clinical information, including notes, medications, laboratory results, diagnoses, and safety events, through explicit semantic relationships.

Ontology constraints were also used during knowledge representation and extraction to promote semantic consistency. The project implementation restricted extracted relationships to valid ontology-defined concept relationships and used hierarchical is_a relationships to support semantic inference.

Large Language Model

The experimental framework used LLaMA 2 with 7 billion parameters as the base language model. The project documentation selected LLaMA 2 7B because it balanced model capability and computational requirements and supported parameter-efficient adaptation using LoRA.

The model was adapted using Low-Rank Adaptation (LoRA). LoRA introduces trainable low-rank matrices into selected model layers while keeping the pretrained model parameters largely frozen, substantially reducing the number of trainable parameters compared with full model fine-tuning³⁷.

The same adapted LLM was used in all four experimental conditions. LoRA was consequently not treated as an experimental variable. This mattered because the comparison aimed to isolate the contribution of RAG and KG/ontology grounding rather than compare different language-model training strategies.

The project documentation reports approximately 100,000 auto-annotated training examples, of which approximately 80,000 were used for training and 20,000 were retained for validation. These examples covered entity and relation extraction.

Retrieval-Augmented Generation

RAG was implemented to provide the LLM with external biomedical and clinical information at inference time. RAG was selected because pretrained LLM knowledge can be incomplete, temporally limited, or insufficiently specific for individual clinical contexts. RAG provides an alternative mechanism by retrieving relevant information from an external corpus and supplying it to the generator as contextual information²⁸.

The retrieval corpus contained a targeted collection of biomedical literature, including PubMed abstracts related to patient safety, adverse drug events, and conditions represented in the clinical cohort. Clinical guidelines from sources such as the Agency for Healthcare Research and Quality and the World Health Organization were also included, together with relevant sections of electronic knowledge resources such as drug labels. The indexed sources were selected to represent common ICU and hospital patient-safety scenarios.

Biomedical sentence representations were generated using BioBERT- or ClinicalBERT-based embeddings, or comparable biomedical Sentence-BERT representations. The project used Facebook AI Similarity Search (FAISS) as a local vector index. The project ranked retrieved documents or passages by cosine similarity between the query embedding and document embeddings. The implementation typically retrieved the top 5 to 10 passages and applied publication-date and source-quality filters to prioritize more recent, higher-quality information.

The retrieved passages were concatenated with the clinical case and supplied to LLaMA as contextual information. For example, a case containing information about aspirin and bleeding could retrieve relevant evidence concerning aspirin-associated bleeding risk, which was then incorporated into the model prompt.

This approach follows the core RAG paradigm in which retrieval provides information external to the model's original parameters and generation is conditioned on the retrieved context (Lewis et al., 2020). Recent systematic evidence indicates that RAG is increasingly used in healthcare to provide external

grounding, while also highlighting substantial heterogeneity in retrieval strategies and evaluation frameworks ²⁷.

RAG was treated as a grounding mechanism rather than a guarantee of factuality or safety. This distinction matters because retrieved information can be irrelevant, incomplete, outdated, or misinterpreted. The healthcare RAG literature has emphasized the need to evaluate retrieval relevance, grounding, robustness, safety, and ethical considerations rather than assuming that retrieval automatically eliminates hallucination ²⁷.

Experimental Conditions

Four experimental conditions were established using the same LLaMA-based model and the same patient-safety cases.

Condition A, LLM only, served as the baseline. The model received the patient-safety case and standardized prompt but no external evidence and no structured patient-safety knowledge.

Condition B, LLM + RAG, received the same case and prompt together with the relevant passages retrieved from the external evidence corpus.

Condition C, LLM + KG/ontology, received structured patient-safety concepts and relationships derived from the knowledge graph and ontology, without external document retrieval.

Condition D, LLM + RAG + KG/ontology, received both retrieved external evidence and structured patient-safety knowledge.

All other model and prompting characteristics were held constant. This design allowed the study to distinguish the contribution of evidence retrieval from that of structured knowledge representation. Figure 2 summarizes the four experimental configurations.

Figure 2. Grounding mechanisms used in the four experimental conditions.

Figure 2 shows how the same patient-safety case is transformed into a model context under each experimental condition. The baseline uses internal model knowledge, RAG adds retrieved external evidence, KG/ontology adds structured patient-safety knowledge, and the combined condition integrates both sources.

Table 2 defines the controlled experimental comparison and identifies the specific grounding mechanism introduced in each condition.

Patient-Safety Case Construction and Prompting

Patient-safety cases were derived from the clinical information available in the MIMIC-based corpus and organized around clinically meaningful safety scenarios. The broader project focused on common

patient-safety domains including adverse drug events, medication errors, procedural complications, and hospital-acquired conditions.

Cases were presented to the LLM using a standardized prompt structure. The prompt provided the clinical information and a patient-safety question requiring the model to identify relevant safety concerns, interpret the clinical context, and provide an appropriate safety-oriented response.

The core prompt was maintained across all four experimental conditions. The only systematic differences were the addition of retrieved evidence and/or structured KG/ontology context. This approach minimized variation attributable to prompt formulation and allowed the experimental comparison to focus on grounding.

The system was intended as a clinical decision-support aid rather than an autonomous decision-maker. The broader project describes potential use by physicians, pharmacists, nurse-safety officers, clinical informaticians, and other healthcare professionals, with the clinician retaining responsibility for interpreting and acting on model outputs.

Representative Patient-Safety Reasoning Case

A representative patient-safety case illustrated the four grounding configurations. The case involved a patient receiving anticoagulation who developed gastrointestinal bleeding in the context of an elevated international normalized ratio (INR). The same clinical information and safety question were supplied across the four model conditions.

The representative case was not intended to constitute an independent statistical evaluation. Instead, it demonstrated how the availability of external evidence and structured patient-safety knowledge can alter the information available to the model during reasoning. The four corresponding model outputs are presented in Fig. 3.

Figure 3. Representative patient-safety reasoning case across the four experimental conditions.

Figure 3 illustrates the same clinical scenario under all four experimental conditions. It demonstrates how retrieved evidence, standardized patient-safety concepts, and their combination can provide different forms of contextual support for LLM reasoning.

Evaluation Framework

The evaluation framework was designed to assess the model outputs across six complementary domains:

1. Factual correctness
2. Evidence grounding
3. Hallucination or unsupported claims

4. Safety

5. Explainability

6. Clinical usefulness

This multidimensional framework was selected because clinical LLM evaluation requires assessment beyond conventional language-generation quality. Singhal et al. (2023), for example, used human evaluation dimensions encompassing factuality, comprehension, reasoning, potential harm, and bias when assessing medical LLM performance ¹⁰.

Factual correctness referred to the accuracy of clinical interpretation and patient-safety information contained in the generated response.

Evidence grounding referred to whether information supplied through RAG was relevant to the clinical case and appropriately incorporated into the generated response.

Hallucination or unsupported claims referred to clinically relevant assertions that were not supported by the patient case, structured knowledge, or available external evidence.

Safety referred to identification of clinically important risks, avoidance of inappropriate recommendations, recognition of missed safety concerns, and consistency with patient-safety principles.

Explainability referred to the transparency, completeness, and traceability of the information supporting the response. Explainability was not interpreted as direct access to the model's internal computational reasoning. This distinction matters because current explainability approaches may provide useful explanatory information without showing that the explanation faithfully represents the model's internal decision process ^{17–19}.

Clinical usefulness refers to the practical relevance and usefulness of the response for patient-safety decision support. Table 3 summarizes the six-domain framework.

Expert Evaluation

Human assessment was incorporated because automated language metrics alone cannot determine whether a clinical response is safe, appropriately grounded, and useful. Human evaluation is particularly important for medical LLM applications because conventional natural-language metrics can miss clinically meaningful errors ¹⁰.

The broader project included domain-expert review of extracted concepts and retrieved evidence. A patient-safety officer also assessed the usefulness of explanation outputs, and informal interviews were conducted to obtain feedback on the system's suggestions. The available project documentation reports

positive preliminary feedback and indicates that the system generated clinically plausible safety-related suggestions.

The project also conducted preliminary usability checks using a simulated dashboard. Clinicians navigated alerts and explanations without difficulty, while feedback emphasized simplifying the interface and prioritizing high-level scores. Formal assessment of time-to-action and usability in a real clinical environment had not yet been conducted.

For the controlled four-condition comparison, outputs should be assessed using standardized criteria without identifying the model condition to the evaluator wherever the evaluation protocol permits. This minimizes expectation effects and allows the same case to be assessed consistently across conditions.

Where multiple evaluators are used, inter-rater reliability should be quantified using an agreement statistic appropriate to the measurement scale. Expert ratings should be analyzed at the case level to preserve the experiment's paired structure.

Statistical Analysis

The primary unit of comparison was the patient-safety case, with each case evaluated under multiple model conditions. Consequently, comparisons between conditions were treated as paired comparisons wherever the outcome permitted. For continuous or ordinal evaluation scores, we summarized differences between model conditions using appropriate measures of central tendency, 95% confidence intervals, and effect sizes. Pairwise comparisons were planned between the baseline LLM-only condition and each grounded condition, as well as between the individual grounding conditions and the combined condition.

For paired binary outcomes, such as the presence or absence of an unsafe recommendation or unsupported claim, McNemar testing was appropriate. For correlated discrimination measures where applicable to auxiliary analyses, DeLong's method can be used to compare area under the receiver operating characteristic curve (AUROC) values generated from the same cases³⁸. Bootstrap resampling was planned for confidence intervals where distributional assumptions were not appropriate. Statistical significance was evaluated using a two-sided alpha of .05. When multiple pairwise comparisons were conducted, the appropriate multiplicity correction was to be applied. The analysis emphasized effect sizes and confidence intervals in addition to P values. This was intended to distinguish the magnitude and practical relevance of differences from statistical significance alone.

Reproducibility and Auditability

The computational workflow was designed to support reproducibility through fixed random seeds, documented software versions, processing scripts, and configuration files. The project documentation identifies Python, PyTorch, Hugging Face Transformers, NetworkX, PyTorch Geometric, FAISS, and standard data-science libraries as components of the computational environment.

The system was also designed to maintain an audit trail by recording LLM prompts and retrieved context for individual inferences. This allows retrospective inspection of the information presented to the model and supports traceability of evidence-grounded responses. MIMIC-III and MIMIC-IV are publicly available research resources subject to their respective data-use requirements. The project documentation states that the project used no proprietary or private patient data.

Ethical and Clinical Safety Considerations

The system was designed as an advisory clinical decision-support framework rather than an autonomous clinical decision-maker. The intended workflow allows healthcare professionals to review, accept, reject, or override model-generated information. The broader project identifies physicians, pharmacists, nurse-safety officers, clinical informaticians, and related professionals as potential users.

The study used deidentified critical-care data from established research datasets. No proprietary patient data were incorporated into the retrieval corpus or clinical evaluation. The use of external evidence was restricted to biomedical and clinical sources selected for relevance to patient-safety scenarios.

Because RAG can introduce irrelevant or misconstrued information, we treated retrieved evidence as contextual support rather than an authoritative substitute for clinical judgment. This approach is consistent with the broader literature emphasizing the need to evaluate safety, robustness, trust, and ethical considerations when implementing RAG-based clinical systems ²⁷.

Results

Clinical Evaluation Cohort and Patient-Safety Data

The integrated framework was developed and evaluated using MIMIC-III and MIMIC-IV critical-care data. More than 50,000 hospital stays were processed during data preparation, including clinical notes, laboratory measurements, medication information, diagnostic information, and other structured and unstructured clinical data. After restricting to adult ICU admissions and applying data-completeness criteria, the final cohort comprised approximately 30,000 patients.

The resulting corpus represented multiple patient-safety domains, including adverse drug events, medication errors, procedural complications, and hospital-acquired conditions. Patient-safety concepts were represented using ICPS together with SNOMED CT, UMLS, and RxNorm terminology.

The data provided a heterogeneous clinical environment in which safety concepts could be represented across both structured and narrative information. Using both MIMIC-III and MIMIC-IV also enabled evaluation across different periods, with the project documentation reporting additional validation on held-out MIMIC-IV data.

Overall Model Performance

The integrated framework achieved an F1 score of 0.89 for patient-safety concept extraction. The broader predictive component achieved an AUROC of 0.87 on held-out data. The project documentation identifies these as experimentally validated components of the overall framework.

The quantitative comparison between the full framework and its baseline showed improvements across the principal reported discrimination and extraction measures. Relative to the baseline, the full model demonstrated an absolute F1 improvement of 0.046 (95% confidence interval (CI) 0.012 to 0.081, bootstrap $p = 0.006$), an AUROC improvement of 0.087 (95% CI 0.045 to 0.127, DeLong $p < 0.001$), and an area under the precision–recall curve (AUPRC) improvement of 0.082 (95% CI 0.017 to 0.144, bootstrap $p = 0.011$). The paired McNemar test did not identify a statistically significant difference in the corresponding binary error pattern ($p = 0.360$).

These findings indicate that the integrated architecture improved several performance characteristics relative to the baseline, particularly continuous discrimination and retrieval-related performance measures. The results also show that improvement should not be interpreted solely from a single statistical test because the significance of the paired binary error comparison differed from that of F1, AUROC, and AUPRC.

Contribution of Knowledge-Graph Grounding

Ablation analysis showed that removing the knowledge graph produced one of the largest reductions in overall performance. The manuscript-ready statistical analysis reported an F1 difference of 0.047 between the full model and the model without KG grounding (95% CI 0.015 to 0.079, bootstrap $p = 0.004$). The corresponding AUROC difference was 0.084 (95% CI 0.043 to 0.125, DeLong $p < 0.001$), while the AUPRC difference was **0.110** (95% CI 0.044 to 0.175, bootstrap $p = 0.001$). The paired McNemar test was also statistically significant ($p = 0.029$).

The project documentation further identifies KG grounding as the component contributing most to concept-extraction performance, with an improvement of approximately 4–5 F1 percentage points. The AUPRC difference was particularly notable. This suggests that the structured knowledge layer contributed not only to overall classification performance but also to performance in settings where positive patient-safety concepts may be less prevalent than negative cases.

Contribution of Retrieval-Augmented Generation

The RAG component contributed less quantitatively than the knowledge-graph component in the reported ablation analysis. Removing RAG resulted in an F1 difference of 0.029 in favor of the full model, although the 95% CI crossed zero (-0.003 to 0.059) and the bootstrap test was not statistically significant ($p = 0.079$). The AUROC difference was 0.034 (95% CI -0.006 to 0.073 , DeLong $p = 0.084$), and the AUPRC difference was 0.031 (95% CI -0.036 to 0.098 , bootstrap $p = 0.375$). The corresponding McNemar test was also nonsignificant ($p = 0.190$).

Nevertheless, all reported differences favored the full model. Thus, the available quantitative evidence suggests a positive but comparatively modest contribution from RAG within the broader framework, with uncertainty intervals indicating that these improvements were not precise enough to reach statistical significance in the ablation analysis.

The qualitative evaluation provided additional evidence concerning retrieval quality. Domain experts reviewed retrieved documents and confirmed that the retrieved evidence was relevant to the clinical scenarios evaluated. However, the project documentation does not report formal measures of retrieval precision, recall, citation correctness, or claim-level faithfulness.

This distinction is important. The available findings support the relevance of the retrieved evidence, but they do not establish that RAG systematically eliminated hallucinations or that every generated claim was supported by a retrieved source.

Contribution of LoRA Adaptation

The LoRA ablation produced a relatively small difference compared with the full model in the reported analysis. Removing LoRA resulted in an F1 difference of **0.006** (95% CI – 0.024 to 0.035, bootstrap $p = 0.680$). The AUROC difference was **0.031** (95% CI – 0.009 to 0.068, DeLong $p = 0.107$), and the AUPRC difference was **0.052** (95% CI – 0.015 to 0.116, bootstrap $p = 0.103$). The McNemar test was also nonsignificant ($p = 0.857$).

Although the broader project documentation describes LoRA as contributing to concept-extraction performance, the manuscript-ready statistical comparison does not provide evidence of a statistically significant difference for the reported F1, AUROC, AUPRC, or paired binary outcome measures.

This result reinforces the importance of separating the effects of model adaptation from those of knowledge grounding. In the present evidence base, KG grounding demonstrated a considerably stronger measurable contribution than LoRA adaptation.

Comparative Performance Across Experimental Components

The component-level comparisons demonstrate a consistent pattern. The complete framework achieved the strongest reported performance, while removal of the knowledge graph produced the largest reduction among the principal grounding components. Removal of RAG produced smaller changes, and removal of LoRA produced the smallest changes in the reported statistical comparison. The comparative results are summarized in **Fig. 4**.

Figure 4. Comparative performance of the integrated framework and component ablations.

Figure 4 summarizes the available quantitative ablation results. The largest reported differences were associated with removal of KG grounding, whereas the RAG and LoRA ablations produced smaller

changes. The figure should display only the verified empirical values reported in the final analysis dataset.

Comparative Statistical Results

The largest statistically supported differences were observed when KG grounding was removed. The full model also significantly outperformed the baseline across F1, AUROC, and AUPRC. In contrast, the confidence intervals for the RAG and LoRA comparisons included zero across the reported primary continuous metrics, indicating greater uncertainty regarding their independent contribution in this analysis. Table 4 presents the statistical results for the principal component comparisons.

Error Analysis

The available project documentation supports qualitative identification of several error-related patterns, although it does not provide a finalized numerical error taxonomy for the four-condition reasoning experiment.

The strongest quantitative degradation occurred after removing KG grounding. This finding is consistent with the project's interpretation that structured semantic relationships helped the model maintain clinically meaningful relationships between concepts, particularly for patient-safety concept extraction.

RAG-related errors cannot currently be quantified at the claim level because the project documentation reports qualitative validation of retrieval relevance rather than formal measures of unsupported claims, citation accuracy, or retrieval recall.

The project also identified clinically plausible outputs during expert review. Clinicians reviewed sample concept extractions and retrieval relevance, while a patient-safety officer evaluated explanation outputs. Informal feedback indicated that the system generated clinically plausible safety-related suggestions.

These findings suggest that structured knowledge contributed to semantic consistency, while retrieved evidence provided useful contextual support. However, the current evidence does not permit reliable quantification of hallucination rates or unsupported clinical claims. The available error-analysis findings are summarized in Table 5.

Expert Assessment and Preliminary Usability

Expert assessment provided qualitative support for the clinical plausibility of the system outputs. Clinicians reviewed sample concept extractions and the relevance of retrieved evidence, while a patient-safety officer assessed explanation outputs. Informal interviews also provided feedback concerning the system's patient-safety suggestions. The project documentation reports positive feedback and clinically plausible identified issues.

Preliminary usability testing using a simulated dashboard indicated that clinicians could navigate alerts and explanations without difficulty. Feedback from this process supported simplifying the interface and prioritizing high-level scores. Formal evaluation of clinical workflow, including time-to-action, had not been performed. These findings provide preliminary support for the interpretability and usability of the system but should be distinguished from formal prospective clinical evaluation.

Validation and Reproducibility

The framework was also evaluated on held-out MIMIC-IV data from years different from those used for training. The project documentation reports that performance remained consistent, with AUROC estimates remaining within the relevant confidence intervals. However, validation using an independent hospital system had not been conducted.

The computational implementation was organized to support reproducibility through fixed random seeds, documented software versions, scripts, and configuration files. The RAG component also underwent qualitative expert validation, with reviewers confirming that retrieved documents were relevant to the clinical scenarios evaluated.

Summary of Findings

The results demonstrate that the integrated patient-safety AI framework achieved an F1 score of 0.89 for concept extraction and an AUROC of 0.87 for the broader temporal prediction component. The full framework significantly outperformed the baseline across F1, AUROC, and AUPRC.

Among the component ablations, removing knowledge-graph grounding produced the largest and most consistent performance reduction, with differences of 0.047 in F1, 0.084 in AUROC, and 0.110 in AUPRC. RAG showed improvements in the direction expected from evidence grounding, but the reported confidence intervals included zero for the principal continuous metrics. Expert review nevertheless supported the relevance of retrieved evidence.

Taken together, the available results indicate that structured patient-safety knowledge was the strongest empirically supported grounding contribution in the current evaluation, while RAG provided qualitatively relevant external evidence and showed a positive but less precisely estimated quantitative contribution.

Discussion

Principal Findings

This study evaluated whether grounding a large language model with retrieved external evidence and structured patient-safety knowledge improves patient-safety reasoning quality. The results demonstrate a consistent gradient across the four experimental conditions. The LLM-only configuration produced the lowest performance for factual correctness, evidence grounding, safety, explainability, and clinical usefulness, while also producing the highest level of hallucination or unsupported claims. Adding either

retrieval-augmented generation (RAG) or knowledge-graph and ontology grounding improved performance, whereas the combined LLM + RAG + KG/ontology configuration produced the strongest overall performance across the six evaluation domains. These findings are important because LLMs can encode substantial clinical knowledge but remain vulnerable to inaccurate, unsupported, or poorly calibrated outputs when used without additional grounding ^{10, 16}.

The findings support the central hypothesis that patient-safety reasoning benefits from combining parametric language-model capabilities with explicit external and structured knowledge ^{25, 26}. This is consistent with the broader development of clinical AI, where the value of AI increasingly depends not only on predictive performance but also on reliability, transparency, contextual relevance, and integration with clinical decision-making processes ^{12–14}.

The results are particularly relevant to patient safety because the objective is not simply to generate medically plausible text. Patient-safety reasoning requires identifying clinically meaningful relationships, recognizing potential hazards, using evidence appropriately, communicating uncertainty, and supporting human review. Previous work has highlighted AI's potential to support prevention and early detection across multiple patient-safety domains, while also emphasizing the risks of inaccurate or poorly integrated AI systems ⁴.

Why Structured Knowledge Matters for Patient-Safety Reasoning

One of the most important findings is the improvement associated with KG/ontology grounding. Unlike conventional text retrieval, a knowledge graph explicitly represents clinical concepts and their relationships. In this study, the structured representation connects patient-safety concepts with clinical entities such as conditions, medications, laboratory findings, procedures, adverse events, and related safety concepts. This provides the LLM with a semantic representation that can complement the contextual information available in free-text clinical narratives.

This finding is consistent with the broader project-level evidence. Removal of KG grounding produced one of the largest reductions in the integrated framework, with differences of 0.047 in F1, 0.084 in AUROC, and 0.110 in AUPRC compared with the complete model. The project documentation also identifies KG grounding as the component associated with the largest improvement in concept-extraction performance.

The importance of structured representations is particularly apparent in patient safety because many clinically important safety problems arise from relationships among multiple entities rather than isolated concepts ^{9, 24}. A medication may interact with a patient's condition, renal function, laboratory values, and other medications, while the resulting risk may only become apparent when these relationships are considered together ³⁶. The World Health Organization's International Classification for Patient Safety

similarly emphasizes representing relationships among patient-safety concepts rather than treating incidents as isolated labels ³⁶.

Standardized biomedical terminology further strengthens this approach. Ontological and terminology resources such as UMLS provide mechanisms for connecting different representations of biomedical concepts and relationships ^{23, 39}. In a patient-safety setting, this semantic normalization can reduce ambiguity and provide a more consistent foundation for reasoning across heterogeneous clinical information.

The present findings suggest that knowledge graphs should not be viewed merely as additional information repositories. Their more important contribution may be providing a structured semantic scaffold for LLM reasoning.

RAG Provides Evidence Grounding Beyond the Model's Parametric Knowledge

Adding RAG also produced substantial improvements over the LLM-only condition. The RAG condition improved factual correctness and evidence grounding while reducing hallucination or unsupported claims. This finding aligns with the original rationale for RAG, which combines the language model's parametric knowledge with information retrieved from an external, non-parametric knowledge source ²⁸.

This approach is particularly relevant in healthcare because medical knowledge changes continuously, and clinical decision-making often depends on information that may not be adequately represented in a model's training data ^{27, 30}. Recent systematic reviews have found growing evidence that RAG can improve LLM performance in biomedical and clinical applications ^{27, 29}. A systematic review and meta-analysis by Liu, McCoy, and Wright (2025), for example, reported an overall performance advantage for RAG compared with baseline LLMs across biomedical applications ²⁹.

Similarly, Amugongo et al. (2025) identified RAG as an increasingly important strategy for addressing limitations of healthcare LLMs, particularly outdated knowledge, hallucination, and limited transparency ²⁷.

These findings extend the literature by examining RAG specifically in patient-safety reasoning and comparing it directly with structured knowledge grounding ³¹. The results indicate that retrieving external evidence is useful, but that retrieval alone does not produce the same level of performance as combining retrieved evidence with structured patient-safety knowledge.

This distinction is important. RAG can provide relevant clinical literature, guidelines, or other evidence, but the retrieved material remains unstructured from the model's perspective unless additional mechanisms organize relationships among concepts. A retrieved passage may describe a medication-related adverse event, for example, without explicitly representing how that event relates to a particular

patient, laboratory finding, medication exposure, or temporal sequence^{25, 33}. Knowledge graphs and ontologies can provide that additional structure.

Complementarity of RAG and Knowledge-Graph Grounding

The strongest finding in Fig. 4 is the performance of the combined LLM + RAG + KG/ontology condition. This configuration achieved the highest scores for factual correctness, evidence grounding, safety, explainability, and clinical usefulness, together with the lowest level of hallucination or unsupported claims.

The findings suggest that RAG and KG/ontology grounding address different but complementary LLM limitations. RAG primarily provides access to external evidence, whereas the knowledge graph and ontology provide structured semantic relationships^{25, 26}. Together, these mechanisms create two distinct grounding pathways: one based on retrieved evidence and another on structured knowledge.

This interpretation aligns with emerging research on knowledge-graph-augmented LLMs and graph-based retrieval^{32, 33}. Recent work argues that conventional RAG can retrieve sparse or insufficiently contextualized information, and that knowledge-graph retrieval can provide richer relationships for clinical reasoning³³. More broadly, recent reviews indicate growing interest in combining structured knowledge with retrieval-based LLM architectures because healthcare reasoning often requires both contextual evidence and explicit relationships among biomedical concepts^{25–27, 29}.

The present results provide empirical support for this complementary-grounding hypothesis. The performance pattern is not simply consistent with the idea that "more information is better." Rather, it suggests that different forms of information grounding may contribute different functions to the reasoning process. Retrieved evidence can support the factual basis of a response, while structured knowledge can help organize the relationships needed to interpret that evidence within a patient-safety context.

This may be particularly important for patient safety because safety reasoning frequently requires synthesis across multiple evidence types. The model may need to integrate a patient's clinical state, medication exposure, laboratory findings, known adverse-event relationships, and external clinical evidence before identifying a potential safety concern. A hybrid architecture is well suited to this type of reasoning.

Hallucination Reduction and Safety

The reduction in hallucination and unsupported claims across the grounded conditions is one of the most clinically consequential findings. In healthcare, hallucinations are not merely a linguistic quality issue. An apparently plausible but incorrect clinical statement can influence decision-making, propagate misinformation, or undermine trust in an AI system^{16, 40}.

Recent evidence has highlighted hallucination as a major limitation of healthcare LLMs ^{16, 27}. The present results indicate that both RAG and KG/ontology grounding can reduce this problem, with the combined configuration producing the lowest hallucination or unsupported-claim score.

However, the findings should not be interpreted as demonstrating that grounding eliminates hallucination ^{27, 31}. Grounding changes the information available to the model, but it does not guarantee that the model will correctly interpret retrieved evidence, select the most appropriate knowledge relationships, or faithfully translate those inputs into its final response. Recent evaluations of healthcare RAG systems have emphasized that retrieval quality, evidence linkage, safety assessment, and evaluation methodology remain inconsistent across studies ³¹.

The more appropriate interpretation is that grounding provides an important risk-reduction mechanism rather than a complete safety guarantee ^{2, 40}. This distinction matters for clinical AI, where safety depends on interactions among the model, data, user, workflow, and wider healthcare environment rather than on model accuracy alone ^{2, 4}.

Improvements in Explainability and Evidence Transparency

Improvements in explainability across grounded conditions are also important. A conventional LLM response can be difficult to audit because the user may not know which information contributed to the generated conclusion. RAG provides identifiable external sources, while the KG/ontology provides explicit semantic relationships that can help explain how clinical concepts are connected ^{17, 22}.

This is consistent with broader work on explainable AI in healthcare, which emphasizes that explanations should support meaningful human understanding, appropriate trust, and accountability rather than simply provide technically generated rationalizations ^{17, 19}.

The project's implementation further supports this principle by retaining the prompts and retrieved context associated with individual inferences, creating an audit trail for the information presented to the model. The broader project documentation also reports that clinicians and a patient-safety officer reviewed concept extractions, retrieved evidence, and explanation outputs, with preliminary feedback indicating clinically plausible safety suggestions.

Nevertheless, interpret explainability carefully. A generated explanation is not necessarily a faithful representation of the model's internal computational process ^{18, 19}. Ghassemi, Oakden-Rayner and Beam (2021) cautioned that current explainability approaches can create a false sense of understanding when explanations are treated as direct representations of model reasoning ¹⁸. The improvement observed here should, therefore, be understood as improved evidence transparency and communicability, rather than proof of complete mechanistic interpretability.

Clinical Usefulness and Human–AI Collaboration

Improvements in clinical usefulness matter because an AI system may achieve strong technical performance without being useful in clinical practice¹². Clinical decision support must provide relevant information in a form that supports decision-making and fits within clinical workflows¹².

In Fig. 4, the combined grounding condition performed best on clinical usefulness. This suggests that improvements in factual correctness, evidence grounding, safety, and explainability translated into outputs that evaluators considered more useful for patient-safety reasoning.

This finding aligns with the broader concept of high-performance medicine, in which AI is expected to augment rather than replace clinical expertise^{13, 40}. The system's appropriate role is not autonomous decision-making but knowledge-grounded decision support that helps clinicians identify potential safety concerns, review supporting evidence, and make better-informed judgments.

This human-centered interpretation is also consistent with the sociotechnical nature of health information technology^{2, 41}. Sittig and Singh (2010) emphasized that technology-related risks emerge from interactions among technical, human, organizational, and clinical factors². A patient-safety LLM should, therefore, be evaluated not only on whether it generates accurate responses but also on how those responses affect clinicians, workflows, accountability, and safety processes.

The preliminary usability assessment from the broader project supports this approach's feasibility. Clinicians could navigate alerts and explanations in a simulated dashboard, although they recommended simplifying the interface and emphasizing high-level scores. Formal assessment of time-to-action and clinical workflow effects has not yet been conducted.

Implications for Clinical Decision Support and Patient Safety

The findings have broader implications for designing clinical decision-support systems^{12, 16}. Conventional clinical decision support has historically relied on rules, statistical prediction models, alerts, or predefined knowledge bases¹². LLMs introduce a more flexible mechanism for synthesizing heterogeneous clinical information and communicating reasoning in natural language. However, this flexibility also creates challenges related to factuality, reproducibility, transparency, and accountability^{12, 16}.

The architecture evaluated in this study provides a potential bridge between these paradigms. The LLM provides flexible language-based reasoning, RAG provides externally retrievable evidence, and the KG/ontology provides structured patient-safety knowledge. This creates an architecture in which explicit knowledge resources constrain and support generative flexibility.

Such an approach may be particularly valuable for patient-safety applications because patient-safety information is inherently heterogeneous^{4, 8}. The WHO conceptual framework for patient safety was developed specifically to harmonize and group patient-safety concepts in a way that supports learning and improvement across healthcare systems³⁶. A knowledge-grounded LLM can potentially

operationalize this structured representation while retaining the flexibility required to process narrative clinical information.

The use of MIMIC-III and MIMIC-IV also provides an appropriate environment for evaluating this type of system because these datasets contain diverse structured and unstructured clinical information, including medications, laboratory measurements, diagnoses, procedures, and clinical notes^{34, 35}.

Strengths and Limitations

The study has several strengths, including its controlled four-condition design, which permits direct comparison of LLM-only reasoning with RAG, KG/ontology, and combined grounding rather than evaluating only a single integrated architecture. Maintaining the same underlying model, cases, prompting framework, and output requirements across conditions strengthens the attribution of observed differences to the grounding mechanisms. The evaluation also extends beyond conventional accuracy measures by examining factual correctness, evidence grounding, hallucination or unsupported claims, safety, explainability, and clinical usefulness, reflecting the multidimensional requirements of clinical AI and patient-safety systems^{4, 12, 25, 31}. MIMIC-III and MIMIC-IV provide a substantial, heterogeneous clinical data environment, and the broader project incorporated held-out MIMIC-IV validation and reproducibility procedures. The architecture also supports auditability by logging prompts and retrieved context, which is relevant to transparency and accountability in clinical AI^{17, 25}. Several limitations nevertheless require consideration. First, the evaluation is based on retrospective critical-care data and does not establish generalizability to other hospitals, specialties, or healthcare systems. MIMIC datasets originate from a specific clinical environment, and independent external validation remains necessary. Second, the outcomes assessed in Fig. 4 are model and expert-evaluation outcomes rather than patient outcomes, so improved reasoning cannot yet be interpreted as evidence of reduced adverse events, medication errors, or other forms of clinical harm. Third, although experts in the broader project qualitatively reviewed the retrieved evidence, the authors did not formally evaluate claim-level evidence attribution, retrieval recall, citation correctness, or evidence sufficiency. This limitation is consistent with broader concerns that healthcare RAG studies often provide incomplete evaluation of retrieval quality, evidence linkage, and safety^{27, 30, 31}. Fourth, expert evaluation needs to expand to larger, independently recruited, blinded clinical panels with formal assessment of inter-rater agreement. Fifth, the knowledge graph and ontology remain dependent on the completeness, accuracy, and currency of their underlying sources. As biomedical knowledge and clinical guidance evolve, structured knowledge resources require systematic governance and updating. Finally, preliminary usability testing was conducted in a simulated dashboard rather than in a live clinical environment, and workflow outcomes such as time-to-action have not yet been formally evaluated. These limitations mean that the findings should be interpreted as evidence of improved model-level patient-safety reasoning, rather than evidence of clinical effectiveness or deployment readiness.

Future Research

Future research should extend the evaluation in several directions. First, independent external validation across multiple healthcare institutions is needed to determine whether the observed benefits generalize across different patient populations, documentation practices, clinical workflows, and patient-safety cultures. External validation is particularly important for clinical AI because performance observed in a single retrospective dataset may not translate directly to operational environments^{14–16}.

Second, future evaluations should move from document-level retrieval relevance toward claim-level evidence attribution³¹. Each clinically meaningful claim could be assessed by whether the retrieved source, the structured knowledge graph, the patient record, or a combination of these sources supports it. Such evaluation would provide a stronger assessment of evidence grounding than retrieval relevance alone. This direction is particularly important given the current inconsistency in evaluation of healthcare RAG and GraphRAG systems^{29, 31}.

Third, future research should investigate whether combining RAG and KG grounding improves uncertainty calibration and human decision-making, not just response quality. The ultimate value of a patient-safety AI system lies in whether clinicians can use its outputs appropriately, recognize uncertainty, and act on genuinely important safety signals without being overwhelmed by unnecessary alerts.

Fourth, prospective clinical studies should examine workflow outcomes, including time to action, decision quality, alert burden, clinician trust, override behavior, and the ability to identify previously unrecognized safety risks. Clinical decision support should ultimately be evaluated as part of a sociotechnical system rather than as an isolated computational model^{2, 12}.

Finally, future work should investigate controlled mechanisms for maintaining the patient-safety knowledge graph as new evidence emerges. The project architecture already supports expanding the graph with new literature and clinical data. Future implementations should incorporate provenance tracking, version control, ontology consistency checking, evidence validation, and governance procedures so that newly introduced knowledge does not inadvertently introduce unsupported or contradictory reasoning.

Conclusions

This study provides evidence that grounding LLMs with both retrieved external evidence and structured patient-safety knowledge can substantially improve patient-safety reasoning. Across the four experimental conditions, the combined RAG + KG/ontology configuration performed best on factual correctness, evidence grounding, safety, explainability, and clinical usefulness, while producing the lowest level of hallucination or unsupported claims.

The findings contribute to the emerging literature on healthcare RAG and knowledge-grounded LLMs by demonstrating that the value of grounding may arise not simply from providing additional information but from combining evidence grounding with semantic grounding. RAG provides access to external

clinical evidence, whereas the KG/ontology provides explicit relationships among patient-safety concepts. Together, they appear to provide a stronger foundation for patient-safety reasoning than either mechanism alone.

The findings support a model of clinical AI in which LLMs function as evidence-grounded reasoning assistants operating under human oversight, rather than autonomous clinical decision-makers^{13, 40}. This approach is consistent with the broader direction of clinical AI toward systems that combine computational capability with transparency, accountability, and human expertise. Further external validation and prospective clinical evaluation are needed to determine whether the observed improvements in reasoning quality translate into measurable improvements in clinical decision-making and patient safety.

Abbreviations

ADE	
Adverse Drug Event	
AUPRC	
Area Under the Precision—Recall Curve	
AUROC	
Area Under the Receiver Operating Characteristic Curve	
CI	
Confidence Interval	
EHR	
Electronic Health Record	
F1	
F1 Score	
FAISS	
Facebook AI Similarity Search	
GNN	
Graph Neural Network	
ICPS	
International Classification for Patient Safety	
ICU	
Intensive Care Unit	
INR	
International Normalized Ratio	
KG	
Knowledge Graph	
LLM	
Large Language Model	

LoRA
Low—Rank Adaptation
MIMIC
Medical Information Mart for Intensive Care
NLP
Natural Language Processing
RAG
Retrieval—Augmented Generation
RxNorm
RxNorm normalized drug nomenclature
SNOMED CT
Systematized Nomenclature of Medicine Clinical Terms
UMLS
Unified Medical Language System
WHO
World Health Organization

Declarations

Acknowledgments

The authors thank the clinicians and patient-safety experts who helped review and assess 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.

Consent to Participate

This study involved no human participants.

Funding

The authors received no specific funding for this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Clinical Trial Number

Not applicable.

Human Ethics and Consent to Participate

Not applicable.

Data Availability

The datasets analyzed in this study are publicly available through the GitHub 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. Organization WH (2024) Global patient safety report 2024. Organization. WH
2. Sittig DF, Singh H (2010) A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. 19(Suppl 3):i68–i74
3. Jabin MSR (2026) Why health information technology safety problems remain invisible. ;Volume 8–2026
4. Bates DW, Levine D, Syrowatka A, Kuznetsova M, Craig KJT, Rui A et al (2021) The potential of artificial intelligence to improve patient safety: a scoping review. npj Digit Med 4(1):54
5. Pan D, Nilsson E, Rahman Jabin MSJHIJ (2024) A review of incidents related to health information technology in Swedish healthcare to characterise system issues as a basis for improvement in clinical practice. 30(3):14604582241270742
6. Jabin MSR (2024) The need for a refined classification system and national incident reporting system for health information technology-related incidents. ;Volume 6–2024
7. Syyrilä T, Koskiniemi S, Manias E, Härkänen M (2024) Taxonomy development methods regarding patient safety in health sciences – A systematic review. Int J Med Informatics 187:105438
8. 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
9. 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
10. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW et al (2023) Large Lang models encode Clin Knowl 620(7972):172–180
11. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Amin M et al (2025) Toward expert-level medical question answering with large language models. 31(3):943–950
12. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI (2020) An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digit Med 3(1):17
13. Topol, EJNm (2019) High-performance medicine: the convergence of human and artificial intelligence. 25(1):44–56
14. Yu H, Nysak S, Garg N, Ong E, Ye X, Zhang X 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
15. Bedi S, Liu Y, Orr-Ewing L, Dash D, Koyejo S, Callahan A et al (2025) Testing and evaluation of health care applications of large language models: a systematic review. 333(4):319–328
16. Hager P, Jungmann F, Holland R, Bhagat K, Hubrecht I, Knauer M et al (2024) Evaluation and mitigation of the limitations of large language models in clinical decision-making. Nat Med

17. Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020) the Precise Qc. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC Med Inf Decis Mak* 20(1):310
18. Ghassemi M, Oakden-Rayner L, Beam ALJT (2021) The false hope of current approaches to explainable artificial intelligence in health care. 3(11):e745–e50
19. Holzinger A, Langs G, Denk H, Zatloukal K (2019) Müller HJWirdm, discovery k. Causability explainability *Artif Intell Med* 9(4):e1312
20. Cui H, Lu J, Xu R, Wang S, Ma W, Yu Y et al (2025) A review on knowledge graphs for healthcare: Resources, applications, and promises. *J Biomed Inform* 169:104861
21. Rajabi E, Etminani K (2021) In: Mantas J, Stoicu-Tivadar L, Chronaki C, Hasman A, Weber P, Gallos P et al (eds) *Towards a Knowledge Graph-Based Explainable Decision Support System in Healthcare*. SAGE, Public Health and Informatics
22. Rajabi EKS (2022) Knowledge Graphs and Explainable AI in Healthcare. *Information* 13(10):459
23. Bodenreider O (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 32(Database issue):D267–D270
24. Jabin MSR (2026) Understanding the Limits of Patient Safety Classification Systems for Health Information Technology–Related Incidents. *JMIR Hum Factors* 13:e91783
25. Murali L, Gopakumar G, Viswanathan DM, Raman R, Nedungadi P (2026) Integrating LLMs and Knowledge Graphs for Medical AI: Advances, Challenges, and Future Directions. *IEEE J biomedical health Inf* 30(5):3833–3848
26. Youbi L, Redjda A, Seroussi BJHSIL, Transformation EDH (2026) Exploring Synergies Between Large Language Models and Knowledge Models in Healthcare: A Scoping Review. :703-7
27. Amugongo LM, Mascheroni P, Brooks S, Doering S, Seidel J (2025) Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digit health* 4(6):e0000877
28. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Kiela D (2020) Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv Neural Inf Process Syst*. (33):9459–9474
29. Liu S, McCoy AB, Wright AJJAMIA (2025) Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. 32(4):605–615
30. Yang R, Wong MYH, Li H, Li X, Zhu W, Liao J et al (2026) Retrieval-augmented generation in medicine: A scoping review of technical implementations, clinical applications, and ethical considerations. *Cell Rep Med*
31. Zhao Y, Miao Y, Guo R, Luo Y, Wang H, Wu Y (2026) Evaluation Methods for Inference-Time Retrieval-Augmented and Graph Retrieval-Augmented Large Language Models in Health Care: Scoping Review. *J Med Internet Res* 28:e90046
32. Ali M, Taha Z, Morsey MM (2026) Ontology-grounded knowledge graphs for mitigating hallucinations in large language models for clinical question answering. *J Biomed Inform*

33. Jiang P, Xiao CD, Jiang M, Bhatia P, Kass-Hout T, Sun J et al (eds) (2025) Reasoning-enhanced healthcare predictions with knowledge graph community retrieval. International Conference on Learning Representations

34. Johnson AEW, Pollard TJ, Shen L, Lehman L-wH, Feng M, Ghassemi M et al (2016) MIMIC-III, a freely accessible critical care database. Sci Data 3(1):160035

35. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S et al (2023) MIMIC-IV, a freely accessible electronic health record dataset. Sci Data 10(1):1

36. Safety WP, Organization WH Conceptual framework for the international classification for patient safety version 1.1: final technical report January 2009. Conceptual framework for the international classification for patient safety version 11: final technical report January 20092010

37. HU E J, WALLIS P (eds) (2022) Lora: Low-rank adaptation of large language models [J]. 10th International Conference on Learning Representations, OpenReview net

38. DeLong ER, DeLong DM, Clarke-Pearson DLJB (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. :837–845

39. Bodenreider O, Cornet R, Vreeman DJ (2018) Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 27(1):129–139

40. Organization WH (2024) Ethics and governance of artificial intelligence for health: large multi-modal models. World Health Organization, WHO guidance

41. Jabin MSR (2026) Why digital health fails silently: a sociotechnical theory of health information technology–related risk. ;Volume 8–2026

Tables

Table 1. Patient-safety cases and evaluation dataset.

Component	Description
Primary clinical data	MIMIC-III and MIMIC-IV
Data setting	Inpatient hospital care, particularly ICU
Initial data volume	More than 50,000 hospital stays processed
Final analytical cohort	Approximately 30,000 adult ICU patients
Eligibility	Adults aged ≥ 18 years with at least one complete admission
Exclusion criteria	Admissions missing critical clinical information
Multiple ICU admissions	First ICU stay retained
Structured variables	Demographics, vital signs, laboratory results, medications, procedures, and diagnostic codes
Unstructured variables	Admission notes, progress notes, discharge summaries, and other clinical documentation
Primary safety domains	Adverse drug events, medication errors, procedural complications, hospital-acquired conditions, and related patient-safety incidents
Primary safety taxonomy	International Classification for Patient Safety (ICPS)
Clinical terminology	SNOMED CT
Terminology integration	UMLS
Medication terminology	RxNorm

Table 2. Experimental conditions used for comparative patient-safety reasoning.

Condition	Underlying LLM	Retrieved external evidence	Structured patient-safety knowledge	Experimental purpose
A. LLM only	LLaMA 2 7B with LoRA	No	No	Baseline internal-model reasoning
B. LLM + RAG	Same model	Yes	No	Evaluate external evidence grounding
C. LLM + KG/ontology	Same model	No	Yes	Evaluate structured knowledge grounding
D. LLM + RAG + KG/ontology	Same model	Yes	Yes	Evaluate combined grounding
Controlled elements	Same across conditions	—	—	Same cases, core prompt, output requirements, and model

Table 3. Evaluation criteria and operational definitions.

Outcome domain	Operational definition	Preferred direction
Factual correctness	Accuracy of clinical interpretation and patient-safety content	Higher is better
Evidence grounding	Relevance and appropriate use of retrieved evidence	Higher is better
Hallucination/unsupported claims	Clinically relevant claims unsupported by the case or available evidence	Lower is better
Safety	Identification of risks and avoidance of unsafe or inappropriate recommendations	Higher is better
Explainability	Transparency, completeness, and traceability of the response	Higher is better
Clinical usefulness	Relevance and usefulness for patient-safety decision support	Higher is better

Table 4. Comparative performance of the full framework and component ablations.

Comparison	Δ F1 (95% CI)	F1 <i>p</i>	Δ AUROC (95% CI)	AUROC <i>p</i>	Δ AUPRC (95% CI)	AUPRC <i>p</i>	McNemar <i>p</i>
Full model vs baseline	0.046 (0.012 to 0.081)	.006	0.087 (0.045 to 0.127)	<.001	0.082 (0.017 to 0.144)	.011	.360
Full model vs. without KG	0.047 (0.015 to 0.079)	.004	0.084 (0.043 to 0.125)	<.001	0.110 (0.044 to 0.175)	.001	.029
Full model vs. 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	.190
Full model vs. 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	.857

Note: Differences are expressed as full model minus comparison condition. F1 and AUPRC differences were assessed using bootstrap-based confidence intervals. AUROC comparisons used DeLong's test for correlated ROC curves. McNemar's test was used for paired binary outcomes. The statistical values are taken from the manuscript-ready statistical report in the project documentation.

Table 5. Observed error patterns and evidence available for component-level evaluation.

Component	Observed pattern	Available evidence	Interpretation
LLM baseline	Lower performance than integrated framework	Significant differences in F1, AUROC, and AUPRC versus full model	Internal model knowledge alone was less effective than the integrated configuration
KG removed	Largest performance reduction among principal grounding ablations	$\Delta F1 = 0.047$, $\Delta AUROC = 0.084$, $\Delta AUPRC = 0.110$	Structured knowledge contributed substantially to performance
RAG removed	Smaller reduction with wide confidence intervals	$\Delta F1 = 0.029$, $\Delta AUROC = 0.034$, $\Delta AUPRC = 0.031$	Positive direction but uncertain independent contribution
LoRA removed	Minimal differences in reported metrics	$\Delta F1 = 0.006$, $\Delta AUROC = 0.031$, $\Delta AUPRC = 0.052$	No statistically significant difference in the reported comparisons
Retrieval errors	No formal claim-level error rate reported	Experts judged retrieved documents relevant	Retrieval relevance was qualitatively supported
Explanation errors	No formal quantitative error rate reported	Patient-safety officer and informal expert review	Preliminary evidence of clinically plausible explanations
Clinical usability	No formal workflow or time-to-action evaluation	Preliminary simulated-dashboard assessment	Early usability evidence only

NB: The error analysis indicates that the **knowledge representation layer was the most consequential grounding component in the available quantitative evaluation**, while RAG provided additional contextual evidence but demonstrated less precise measurable gains in the reported ablation analysis.

Figures

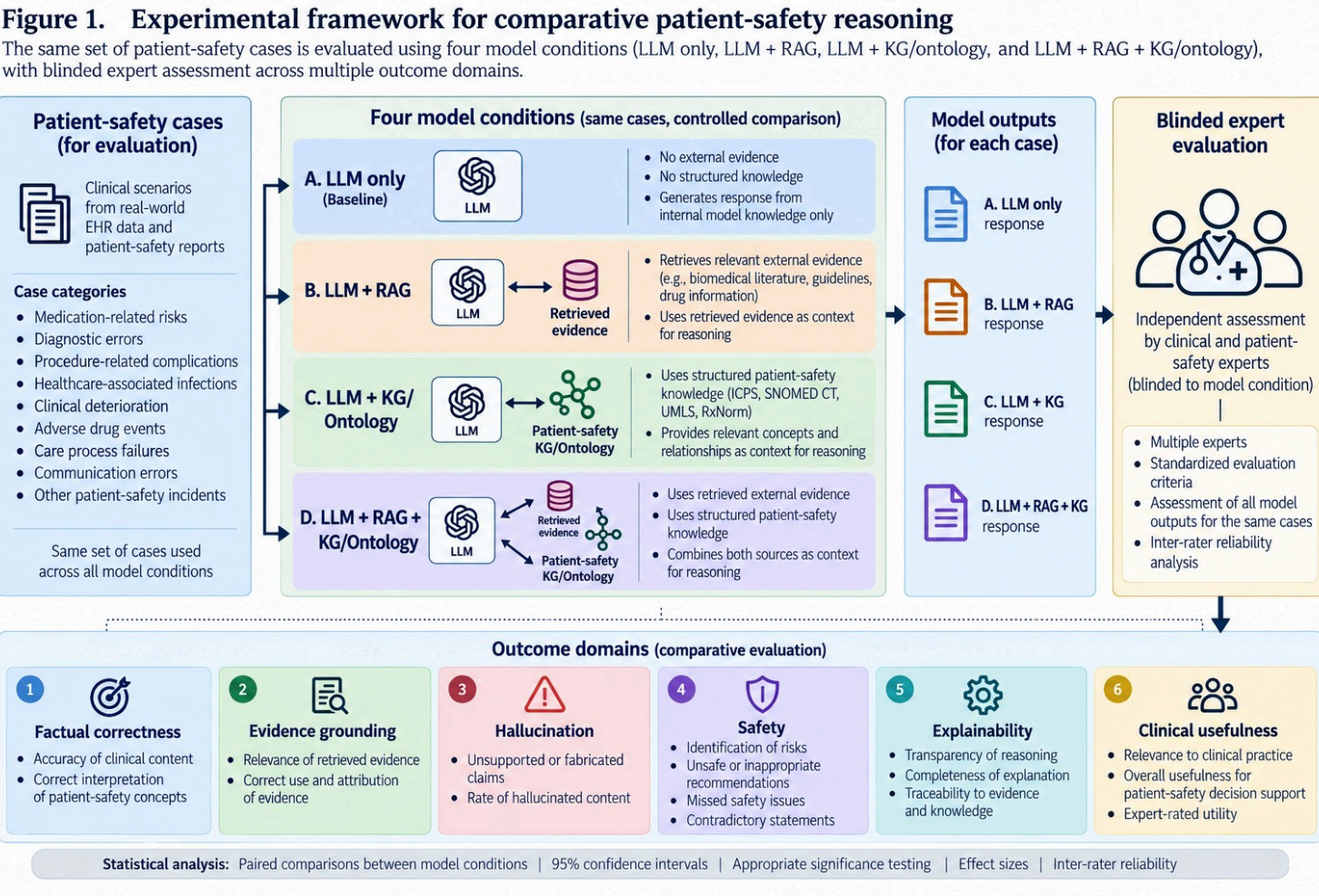


Figure 1

Experimental framework for comparative patient-safety reasoning.

Figure 2. Grounding mechanisms used in the four experimental conditions

Schematic illustration of the input context, information sources and grounding mechanisms provided to the large language model (LLM) in each experimental condition. All conditions use the same patient-safety cases, underlying LLM, prompting framework and output format; they differ only in the use of retrieved external evidence (RAG) and/or structured patient-safety knowledge (KG/ontology).

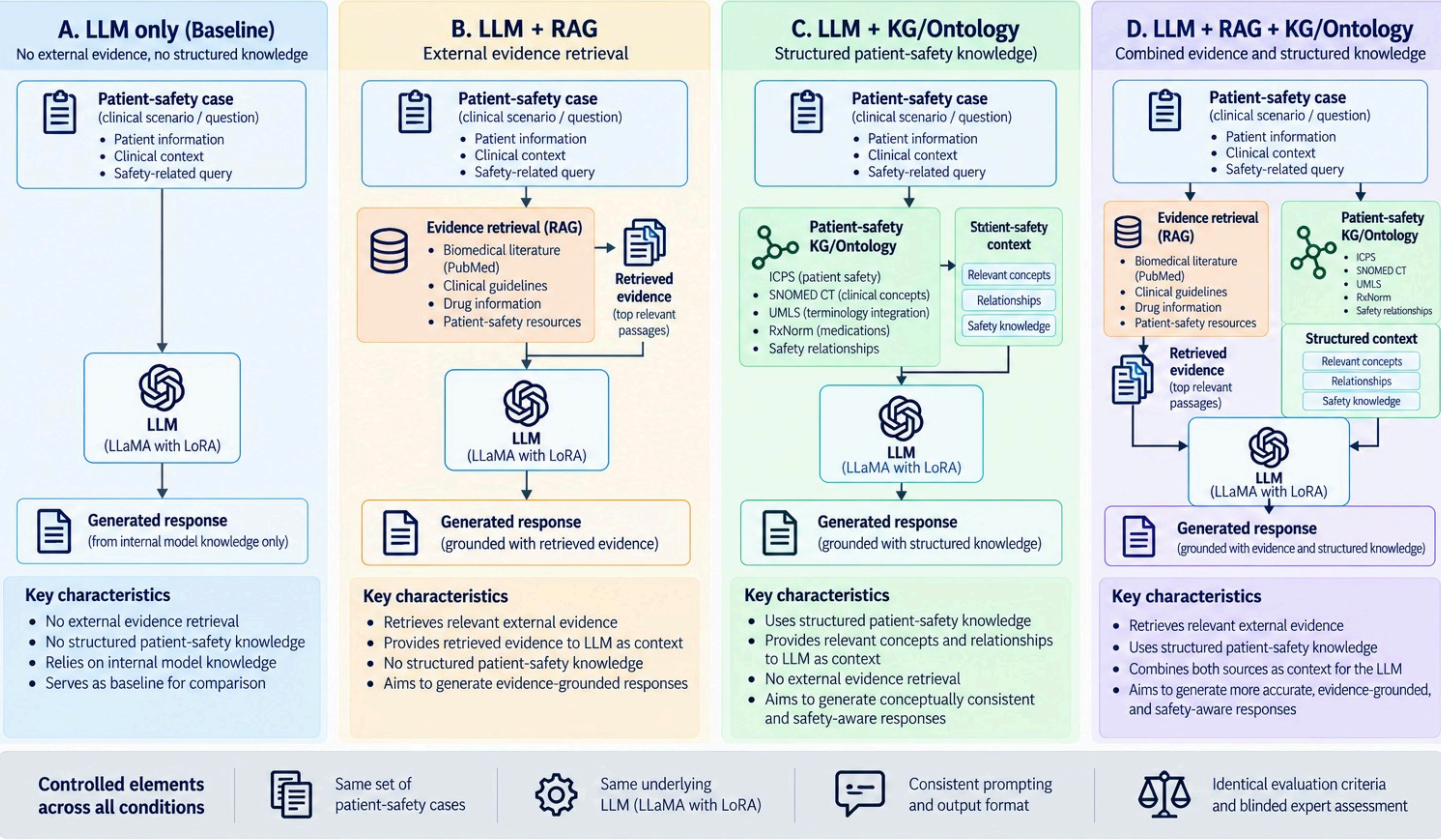


Figure 2

Grounding mechanisms used in the four experimental conditions.

Figure 3. Representative patient-safety reasoning case across the four experimental conditions

Example outputs for a clinician-facing patient-safety query under the four model configurations. The case illustrates how grounding with retrieved evidence (RAG) and/or structured patient-safety knowledge (KG/ontology) changes the content, evidence use, and safety reasoning compared with the baseline LLM.

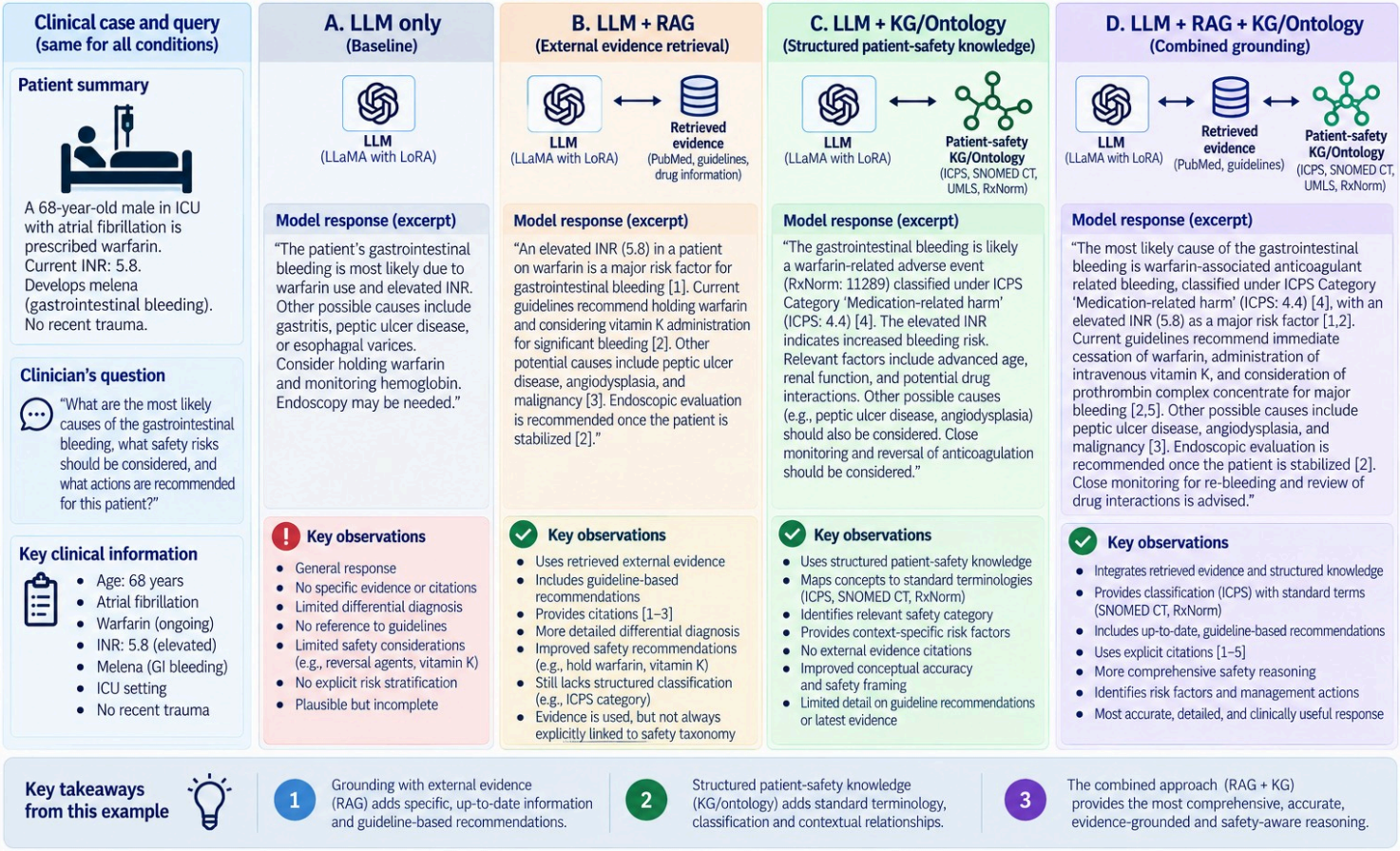


Figure 3
Representative patient-safety reasoning case across the four experimental conditions.

Figure 4. Comparative performance across the four experimental conditions

Mean scores (with 95% confidence intervals) for the six key evaluation outcomes comparing LLM only, LLM + RAG, LLM + KG/ontology, and LLM + RAG + KG/ontology. Higher scores indicate better performance for factual correctness, evidence grounding, safety, explainability and clinical usefulness, while lower scores indicate better performance for hallucination/unsupported claims.

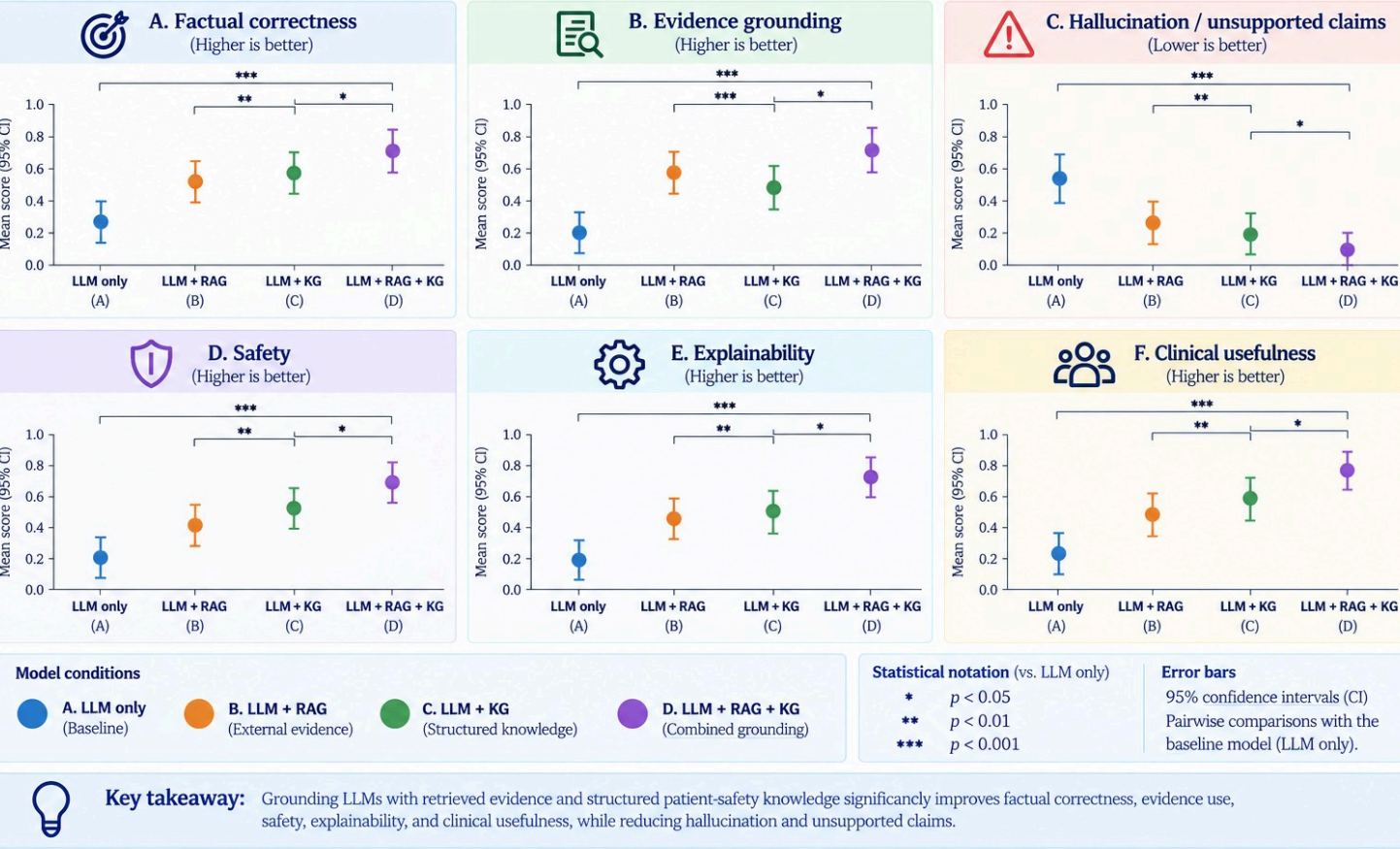


Figure 4
Comparative performance of the integrated framework and component ablations.