

ORIGINAL RESEARCH

Open access

Retrieval-Augmented Clinical Operations Assistant for Answering Hospital Policy Questions Using Local Protocols, Staffing Guidelines, Bed Management Rules, Escalation Pathways, and Real-Time Operational Dashboards

Ricardo Alves¹, Bruno Costa^{1*}, Helena Martins², Nuno Faria¹

Abstract

Hospital staff routinely spend substantial cognitive effort locating operational policies, staffing rules, escalation pathways, and dashboard metrics across fragmented repositories. This hidden search burden can slow decision-making during high-pressure clinical operations. Current hospital knowledge environments rarely support natural-language policy questions answered from the institution's own approved documents. Staff may know what they need to ask, but not where the relevant rule, protocol, or dashboard field is stored. This article proposes a retrieval-augmented clinical operations assistant that accepts free-text questions and retrieves relevant passages from local policy repositories and structured operational data sources. The assistant would synthesize a grounded response while exposing the sources used to generate the answer. The proposed assistant includes a document ingestion pipeline, a vector store, a permissioned large language model, a real-time dashboard connector, and a simple chat interface embedded in the hospital intranet. These components would work together to make local protocols, staffing guidelines, bed management rules, and escalation pathways conversationally accessible. The assistant would be expected to reduce staff search burden, improve visibility of current policy, and support more consistent use of institutional operating rules. Its value would depend on strict grounding in authoritative documents, robust version control, and clear boundaries when policies are missing or contradictory. A retrieval-augmented clinical operations assistant represents an early step toward conversational, trustworthy, and continually updated operational decision support. Such a system should complement, rather than replace, human judgment and formal policy governance.

Keywords Hospital operations, Large language models, Retrieval-augmented generation, Clinical policy, Conversational AI, Operational dashboards

*Correspondence:

Bruno Costa
bruno.costa@gmail.com

¹ Department of Digital Healthcare Informatics, Faculty of Medicine, University of Porto, Porto, Portugal

² Department of Clinical AI and Smart Systems, Faculty of Engineering, University of Aveiro, Aveiro, Portugal

Introduction

Hospital operations depend on a large body of local knowledge that is often more procedural than clinical, including staffing matrices, escalation policies, bed management rules, infection-control procedures, and service-specific operating manuals. Conversational agents in healthcare have been reviewed as tools for clinical communication, patient support, and health information delivery [1], but the administrative burden of locating internal operating rules remains less directly

addressed. Broader reviews of healthcare conversational agents show that these systems can support interaction and information access when designed around user workflow [2, 3]. A clinical operations assistant would apply that conversational logic to the institutional knowledge that nurses, bed managers, administrators, and on-call leaders consult during daily operational decisions.

Conventional document search is poorly suited to questions that are semantic, conditional, or operationally contextual. A staff member may ask whether a temporary bed can be opened under a particular staffing pattern, and a keyword search may miss the answer if the relevant policy uses different terminology. Criteria2Query showed that natural-language interfaces can translate user questions into structured clinical database logic [4], while text-to-SQL work on electronic medical records demonstrates that health questions can be routed toward structured data retrieval [5]. Retrieval-augmented text-to-SQL for epidemiological question answering further suggests that retrieval and query translation can be combined when user questions require both text interpretation and database access [6].

Large language models and retrieval-augmented generation offer a plausible path toward on-demand, citation-backed policy question answering because they can combine natural-language understanding with evidence retrieval at runtime. Almanac demonstrates how retrieval augmentation can ground clinical language-model answers in medical sources [7], while retrieval-augmented trial screening shows how generative systems can be constrained by retrieved eligibility or decision criteria [8]. Biomedical RAG work further indicates that domain-specific retrieval can support more grounded biomedical question answering [9], and EHR-focused RAG shows how clinical information can be extracted and summarized from institutional records [10]. For hospital operations, the same pattern could be adapted to local policy repositories, staffing guidance, bed management rules, and escalation pathways.

The thesis of this article is that a tailored RAG assistant could transform hospital operational knowledge from a fragmented document archive into an instantly accessible conversational resource. A liver disease-specific RAG chat interface illustrates how a domain-bound assistant can retrieve specialty-relevant material before responding, and guideline-focused LLM work shows the value of comparing generated answers with authoritative clinical guidance. Nutrition and cardiovascular guidance studies also demonstrate why source adherence matters when users may act on the assistant's answer. In hospital operations, these principles imply that the assistant should answer only from approved local documents and permitted dashboard data, with visible evidence and clear refusal when sources are insufficient.

Background

Hospital operational knowledge and its fragmentation

Hospital operational knowledge is typically distributed across policy documents, staffing grids, bed escalation procedures, infection-control manuals, disaster plans, transfer rules, security pathways, and department-specific instructions. RAG in healthcare has been framed as a way to improve communication and decision-making by addressing the limitations of unsupported large language model output [11], which is directly relevant when hospital rules are scattered across local repositories. Biomedical RAG emphasizes the need to connect generation to curated domain corpora [9], and EHR summarization work shows how institutional text can be transformed into retrievable clinical information [10]. In the operational setting, the assistant would need to treat local policy texts as the authoritative source of truth rather than relying on general model knowledge.

Retrieval-augmented generation for domain-specific question answering

Retrieval-augmented generation combines a retrieval layer with a generative model so that answers are produced from documents retrieved for a particular query rather than from the model's internal parameters alone. Almanac illustrates this pattern in clinical medicine by using retrieval to support medical question answering [7], while RAG-enabled trial screening applies similar logic to matching clinical questions with trial criteria [8]. COVID-19 fact-checking with retrieval-augmented large language models shows how retrieved evidence can support health-related verification [12], and patient

education work in orthopedic and trauma surgery demonstrates how RAG could support domain-specific explanatory dialogue [13]. A hospital policy assistant would use the same paradigm to retrieve local procedural rules before generating an answer.

Vector stores, embeddings, and document chunking

A hospital policy corpus would need to be transformed into retrievable units before it could support conversational question answering. Scoping work on RAG in medical and nursing domains emphasizes that retrieval quality depends on how source documents are represented and searched [14], and biomedical RAG similarly shows that domain language must be encoded in a way that supports relevant evidence retrieval [9]. RAGAS contributes an evaluation framework for retrieval-augmented generation by separating retrieval, grounding, and generation concerns [15]. For operational policies, chunking would need to preserve context such as department, effective date, role applicability, and exception clauses while still allowing focused retrieval.

LLM capabilities and limitations in highly procedural contexts

Large language models can synthesize complex natural-language answers, rephrase procedural information, and support dialogue across follow-up questions, but they can also generate unsupported statements when grounding is weak. Prostate-specific antigen testing work with retrieval-augmented language models shows how guideline-concordant answering depends on alignment with source guidance [16], while chronic kidney disease nutrition work demonstrates the importance of assessing whether generated advice follows accepted principles [17]. Precision oncology work using expert-guided language models further highlights the need for domain oversight in high-stakes procedural contexts [18]. A clinical operations assistant should therefore be designed to decline unsupported questions rather than improvise answers about staffing, escalation, or bed management.

Conversational AI in healthcare workflow

Conversational AI in healthcare has been explored through systematic reviews, scoping reviews, patient-facing systems, physician perception studies, and acceptability research. Reviews of healthcare conversational agents show that these tools can be useful when they support the user's task and are embedded in realistic workflows [1-3], while rapid review evidence emphasizes that chatbot roles, benefits, and limitations vary substantially across healthcare settings [19]. Expert interviews informing the DISCOVER framework show that design and implementation require attention to context, stakeholders, and evaluation [20]. The proposed operations assistant would extend this literature toward administrative and operational policy support rather than patient-facing advice alone.

Assistant System Overview

High-level architecture

At a high level, the assistant would begin with a user submitting a question through a secure chat interface embedded in the hospital intranet or another approved access point. The query would be embedded and compared with indexed document chunks, as clinical RAG systems such as Almanac demonstrate for evidence-grounded medical answering [7]. If the question requires live operational context, the routing layer could also invoke structured data access, building conceptually on natural-language database interfaces such as Criteria2Query [4] and clinical text-to-SQL approaches [5]. The language model would then synthesize a concise response grounded in retrieved policy passages and permitted dashboard outputs.

Core knowledge sources

The core knowledge sources would include hospital-approved PDFs, Word documents, intranet pages, staffing guidelines, bed management rules, service-line protocols, and escalation pathways. Domain-specific RAG chat interfaces

show how a bounded knowledge corpus can shape an assistant's answers [21], while RAG-based EHR summarization illustrates how institutional documents can be used as source material for generated responses [10]. For questions about current state, retrieval-augmented text-to-SQL demonstrates how natural-language questions can be linked to structured health data sources [6]. The assistant would therefore distinguish between static policy evidence, real-time operational metrics, and hybrid questions that require both.

Design principles

The assistant should follow four design principles: local or tightly controlled deployment, source-grounded answers, a user-friendly conversational interface, and transparent citation of policy sources. RAG in healthcare has been proposed as a way to reduce limitations of general language models by connecting answers to retrievable evidence [11], and medical coding work with retrieval-augmented language models shows why grounding is important when generated output may affect administrative decisions [22]. Studies of physician perceptions of chatbots indicate that clinicians care about trust, accountability, and appropriate use boundaries [23, 24]. These principles would position the assistant as a controlled institutional knowledge interface rather than an open-ended general-purpose chatbot.

Figure 1 presents the proposed end-to-end workflow of the retrieval-augmented clinical operations assistant, showing how local policy sources, dashboard data, grounded generation, citation safeguards, human review, and governance controls connect in hospital operations.

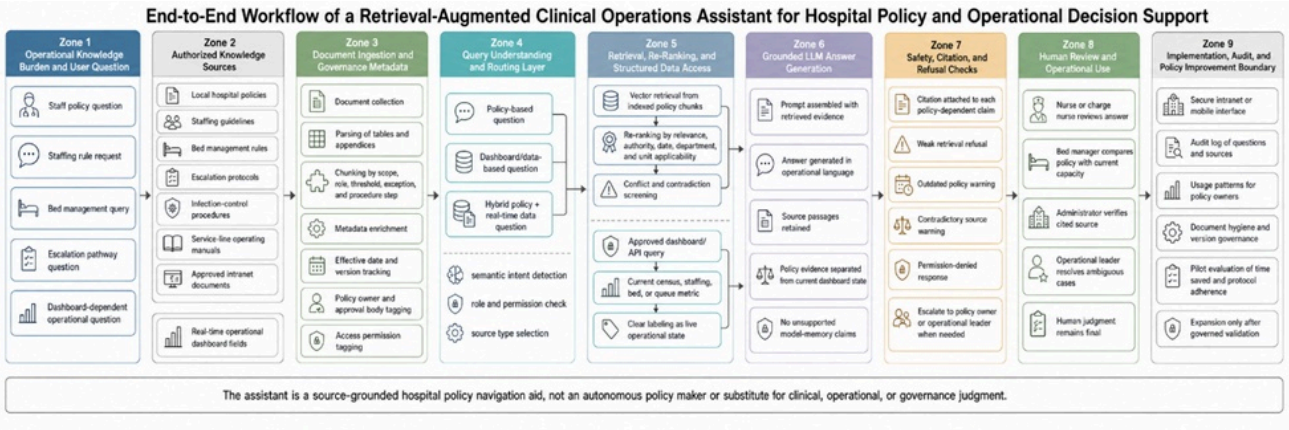


Figure 1. End-to-End Workflow of a Retrieval-Augmented Clinical Operations Assistant for Hospital Policy and Operational Decision Support

Knowledge Base Construction and Document Ingestion

Document collection and preprocessing

Knowledge base construction would begin with collecting approved policy documents from the hospital's repository, intranet, shared drives, and governance systems, using automated ingestion where allowed and controlled upload where automation is not possible. RAG-based EHR summarization shows how clinical text can be processed so that a language model can extract and summarize relevant information [10], while biomedical RAG demonstrates the importance of preparing domain corpora for retrieval [9]. A specialty RAG interface for liver disease further illustrates that source preparation should reflect the boundaries of the intended domain [21]. For hospital operations, preprocessing would need to parse headings, tables, appendices, approval metadata, version histories, and cross-references so that procedural rules remain interpretable after ingestion.

Chunking and indexing strategy

The chunking strategy should divide policies into semantically meaningful units, such as policy purpose, scope, definitions, role responsibilities, escalation levels, staffing thresholds, exception clauses, and procedure steps. RAG scoping work in medical and nursing domains indicates that retrieval effectiveness depends on how source content is segmented and represented [14], and RAGAS frames retrieval quality as a component that should be evaluated separately from final answer quality [15]. Biomedical RAG shows that domain-specific retrieval can support answer generation when the retrieved units capture relevant biomedical meaning [9]. In an operations assistant, chunking should be conservative enough to avoid losing context but specific enough to prevent irrelevant policy sections from being blended into a misleading answer.

Metadata enrichment and freshness tracking

Each retrieved chunk should carry metadata such as document title, department, owner, effective date, revision date, approval body, policy category, unit applicability, and access permissions. Guideline-concordant prostate-specific antigen testing work shows that generated responses should be judged against current guidance rather than against general plausibility [16], and oncology guideline comparison work demonstrates why source authority and guideline versioning matter for language-model outputs [25]. Expert-guided precision oncology systems similarly show that governance and domain review are important when model-supported recommendations depend on specialized source material [18]. Freshness tracking would therefore be a governance function as much as a technical feature, ensuring that policy owners are alerted when documents require review.

Real-time dashboard data integration

For real-time operational questions, the assistant would interpret a restricted class of natural-language requests as calls to approved dashboard APIs rather than as document retrieval tasks. Criteria2Query shows that natural-language questions can be translated into computable clinical database criteria [4], and text-to-SQL generation for electronic medical records demonstrates how health questions can be routed toward structured query logic [5]. Retrieval-augmented text-to-SQL further suggests that retrieved context can support structured question answering when database queries alone are insufficient [6]. In the proposed assistant, dashboard integration would need strict permission checks, audit logs, and clear separation between current metrics and policy interpretation.

Rag Architecture and Real-Time Dashboard Integration

Query understanding and routing

The assistant would first classify each question as policy-based, data-based, or hybrid before deciding whether to retrieve document chunks, query a dashboard API, or combine both sources. RAG-enabled clinical trial screening illustrates how a system can route a user question toward relevant eligibility criteria [8], while Criteria2Query demonstrates how natural-language input can be transformed into structured retrieval logic [4]. Clinical text-to-SQL work shows that structured data access may be necessary when the answer depends on records or operational variables rather than static text [5]. The routing layer would therefore act as an operational triage mechanism for information needs.

Retrieval and re-ranking

After routing, dense retrieval would identify candidate policy passages whose meanings align with the user's question, and a re-ranking step would prioritize the most relevant, authoritative, and non-contradictory evidence. Biomedical RAG demonstrates how domain-specific retrieval can support medical question answering when the relevant source passages are surfaced [9], while RAGAS shows that retrieval quality should be examined as its own component of RAG performance [15]. RAG reviews in medical and nursing domains reinforce that retrieval and generation should not be treated as a single black-box process [14]. In hospital operations, re-ranking should also consider metadata such as policy owner, effective date, and unit applicability.

Answer generation with citation

The answer generation layer would synthesize a response only from retrieved evidence and permitted dashboard outputs, presenting the answer in operational language appropriate for the user’s role. Almanac illustrates how clinical language-model answers can be grounded in retrieved sources [7], and oncology guideline comparison work shows why generated answers should remain tied to authoritative guidance [25]. Cardiovascular nutrition guidance work further demonstrates the importance of guideline-adherent responses when users may rely on the answer for practical decisions [26]. The proposed assistant would therefore function less like an oracle and more like a conversational interface to institutional evidence.

Guardrails against hallucination and outdated policy

The assistant should include guardrails that prevent it from answering when retrieval is weak, when documents conflict, when a user lacks permission, or when the relevant policy appears outdated. Medical coding work with retrieval-augmented language models shows why administrative outputs require grounding and review [22], and expert-guided precision oncology work demonstrates the value of human oversight when model-supported decisions are sensitive [18]. Studies of chatbot acceptability and physician perceptions indicate that users may be more willing to adopt conversational systems when limitations are clear and trust is actively managed [23, 24, 27]. These safeguards would be particularly important in hospital operations because policy answers can influence immediate staffing, escalation, and patient-flow decisions.

Table 1 summarizes the input-to-output generative workflow through which institutional policy documents, metadata, dashboard fields, retrieval processes, and grounded language-model responses are transformed into practical operational answers.

Table 1. Input-to-Output Generative Workflow for a Retrieval-Augmented Clinical Operations Assistant

Workflow Layer	Manuscript-Specific Function	Operational Data or Knowledge Used	Technical Mechanism	Practical Output	Safety or Governance Requirement
User-facing question entry	Allows hospital staff to ask natural-language operational questions without knowing where the relevant rule is stored.	Free-text questions about staffing, bed management, escalation, policy scope, dashboard fields, or unit-specific procedures.	Secure chat interface embedded in the hospital intranet or approved internal access point.	Conversational entry point for operational knowledge requests.	Authentication, role-based access, and clear framing as policy navigation support rather than autonomous decision-making.
Institutional source collection	Builds the approved local knowledge base from hospital-controlled sources.	PDFs, Word documents, intranet pages, staffing grids, bed escalation procedures, infection-control manuals, transfer rules, and service-line protocols.	Controlled document ingestion pipeline with parsing of headings, tables, appendices, approval metadata, and cross-references.	Machine-readable operational knowledge corpus.	Only approved and current institutional documents should be ingested; legacy or duplicate content requires cleanup.

Chunking and indexing	Converts long operational documents into retrievable units while preserving procedural context.	Policy purpose, scope, definitions, role responsibilities, staffing thresholds, exception clauses, escalation levels, and procedure steps.	Semantic chunking, embeddings, vector store indexing, and metadata-aware retrieval.	Searchable policy chunks linked to original source sections.	Chunking must avoid separating rule statements from conditions, exceptions, effective dates, or unit applicability.
Metadata enrichment	Makes source authority and applicability visible during retrieval and answer generation.	Document title, department, owner, approval body, effective date, revision date, policy category, unit applicability, and access permissions.	Metadata tagging at ingestion and filtering during retrieval.	Retrieval results that can be prioritized by relevance, authority, currency, and user permission.	Version control and policy ownership must be actively maintained.
Query classification and routing	Determines whether a question should be answered from policy documents, structured dashboards, or both.	Staff query, user role, policy categories, dashboard availability, and permission status.	Intent classification and routing layer that distinguishes policy-based, data-based, and hybrid questions.	Correct pathway selection before retrieval or dashboard access.	The system must not treat live dashboard metrics as policy rules or policy documents as current operational state.
Vector retrieval and re-ranking	Finds the most relevant local evidence for the user's question.	Indexed policy chunks, metadata, departmental applicability, approval status, and effective dates.	Dense retrieval followed by re-ranking for relevance, authority, currency, and non-contradiction.	Candidate evidence passages for answer generation.	Retrieval quality should be evaluated separately from answer fluency.
Structured dashboard access	Supports questions requiring current operational context.	Approved dashboard fields such as bed availability, census, staffing levels, queues, or operational status indicators.	Restricted dashboard/API connector or structured query layer.	Current operational metric returned alongside policy interpretation when permitted.	Strict permission checks, audit logs, and explicit labeling of metrics as real-time or time-sensitive.
Grounded prompt assembly	Constrains the language model to local evidence and permitted operational data.	Retrieved policy passages, metadata, dashboard outputs, and user role context.	Prompt construction with source snippets, role constraints, and refusal instructions.	Evidence package passed to the permissioned LLM.	The prompt must prevent reliance on unsupported general model knowledge.
Generated answer	Produces a concise	Retrieved policy evidence and	Permissioned LLM with retrieval-	Grounded answer explaining the	The answer must cite sources and

	operational response in language appropriate to the user's role.	approved dashboard data.	augmented generation.	relevant policy, operational metric, or next step.	preserve boundaries when evidence is missing, outdated, or contradictory.
Citation and traceability	Allows staff to verify the answer against the source document.	Source title, section, paragraph, table, effective date, and original document link.	Citation mechanism attached to answer claims.	Clickable or visible source evidence for user inspection.	Each policy-dependent claim should be traceable to an authoritative local source.
Refusal and escalation	Prevents unsupported or unsafe answers.	Weak retrieval signals, contradictory documents, outdated policy metadata, permission restrictions, and missing sources.	Guardrail logic, contradiction detection, and escalation messaging.	Refusal, uncertainty statement, or referral to policy owner or operational leader.	The assistant should not improvise staffing, escalation, or bed-management rules.
User review and action	Supports staff interpretation and next-step planning.	Generated answer, cited policy text, dashboard values, and local operational context.	Human-in-the-loop review through the chat interface.	Staff member verifies the source and applies the rule within existing governance structures.	Human judgment, professional responsibility, and formal policy governance remain final.
Audit and improvement	Converts usage into governance intelligence without creating uncontrolled self-learning.	Query logs, retrieval confidence, source use patterns, flagged answers, and frequently misunderstood policies.	Audit dashboards and policy-owner review reports.	Identification of document hygiene problems, outdated rules, and policy clarification needs.	Any knowledge-base update should be governed by policy owners rather than automatic model self-correction.

Answer Grounding, Citation, and Safety

Citation mechanism and traceability

Every answer produced by the assistant should make its evidentiary basis visible, with policy-dependent claims linked to the exact source paragraph, section, or table from which they were derived. Almanac demonstrates the broader principle that clinical language-model answers should be grounded in retrievable evidence rather than unsupported model memory [7], while RAG in healthcare has been framed as a way to improve decision support by reducing the limitations of ungrounded generative output [11]. In the proposed hospital operations assistant, a response about staffing, escalation, or bed management would therefore cite the relevant local policy section in the sentence where the rule is interpreted. Traceability would also allow a user to open the original document, inspect the context, and verify that the assistant's synthesis remains faithful to institutional policy.

Handling ambiguity, gaps, and contradictions

When policies conflict, are silent, or contain ambiguous wording, the assistant should identify the uncertainty rather than resolving the issue through unsupported judgment. Guideline comparison work with GPT-style systems shows why generated answers must be checked against authoritative sources when procedural recommendations are involved [25], and guideline-concordant testing studies similarly illustrate that answer quality depends on alignment with the relevant source guidance [16]. In a hospital policy setting, the assistant could say that two documents appear to provide different escalation instructions and then direct the user to the policy owner or responsible operational leader. This behavior would be safer than presenting a single synthesized answer when the underlying source material does not clearly support one.

Feedback loop and policy improvement

A clinical operations assistant should include a feedback channel through which users can flag answers that seem incomplete, unclear, outdated, or inconsistent with practice. RAGAS shows that retrieval-augmented systems should be assessed in terms of retrieval quality, grounding, and generated answer faithfulness [15], and expert interviews on healthcare conversational agents emphasize that evaluation should account for implementation context and stakeholder needs [20]. Aggregated feedback could help policy owners identify documents that are frequently misunderstood, poorly retrieved, or in need of revision. In this way, the assistant would not only answer policy questions but also reveal weaknesses in the hospital's operational knowledge base.

User Interaction and Conversational Flow

Intuitive dialogue and follow-up questions

The assistant should support ordinary conversational patterns, allowing users to ask follow-up questions such as whether the same rule applies on night shift, on a different unit, or for a different patient population. Reviews of healthcare conversational agents show that dialogue quality and task fit are central to whether users perceive such systems as helpful [1, 2], while physician-facing chatbot studies indicate that professional users value systems that respond to realistic information-seeking behavior [23]. In a hospital operations setting, multi-turn dialogue would allow a nurse, administrator, or bed manager to refine an initial policy question without restating the entire context. The system would still need to retrieve fresh evidence for each materially different follow-up rather than assuming that the previous answer remains applicable.

Multi-lingual and role-based responses

The assistant could adapt the detail, vocabulary, and presentation of a response according to the user's role, while preserving the same underlying policy evidence. Rapid review work on healthcare chatbots shows that chatbot roles and user groups vary widely across health settings [19], and acceptability research indicates that staff and patients may judge conversational systems partly by whether they fit the user's needs and expectations [27]. A bed manager might receive a concise operational interpretation, while a new staff nurse might receive a more explanatory answer with definitions and links to the source policy. Multilingual output could improve accessibility, but any translated answer should remain tied to the authoritative source and should avoid introducing new policy meaning.

Voice interface and mobile accessibility

Optional speech input and audio output could make the assistant easier to use for staff who are moving between units, managing simultaneous tasks, or unable to type during operational work. Healthcare conversational-agent reviews show that interaction modality is part of system usability and workflow integration [3], while studies of chatbot acceptability highlight the importance of convenience and perceived usefulness in adoption [27]. A mobile or voice-enabled interface could support quick policy checks during bed huddles, escalation calls, or unit coordination, provided that authentication and privacy protections remain intact. The assistant should also present source citations visually when possible, because spoken answers alone may not provide enough traceability for policy-dependent decisions.

Integration Into Hospital Operations Workflow

Embedding in existing communication channels

The assistant would be most useful if it appeared where staff already seek operational information, such as the hospital intranet, a secure mobile application, or approved internal communication platforms. Conversational AI reviews emphasize that workflow fit is a major determinant of usefulness in healthcare environments [1, 19], and physician perception studies show that professional acceptance depends on whether chatbot use feels reliable, appropriate, and relevant to daily work [24]. Embedding the assistant in existing channels would reduce the need for users to remember another tool or navigate a separate policy repository. However, integration should preserve access controls so that users only receive policy and dashboard information they are authorized to view.

Policy governance and continuous updating

A policy-governance dashboard could show policy owners which questions are asked frequently, which documents produce low-confidence retrieval, and which policies may need clarification or updating. Domain-specific RAG systems show that source boundaries and corpus maintenance shape assistant behavior [21], while biomedical RAG and medical-domain RAG reviews emphasize that the quality of retrieved evidence depends on the quality and currency of the underlying knowledge base [9, 14]. Governance workflows could therefore treat user questions as signals about operational confusion, not merely as isolated help requests. Continuous updating would be essential because hospital staffing rules, escalation procedures, and bed management pathways can change as local practice evolves.

Evaluation Strategy

Retrieval and answer quality

The assistant should be evaluated by examining whether it retrieves the correct policy passages, whether its answer faithfully reflects those passages, and whether it clearly distinguishes policy evidence from dashboard-derived current state information. RAGAS provides a useful conceptual structure for assessing retrieval-augmented generation through retrieval, grounding, and response quality dimensions [15], while biomedical RAG work shows that domain-specific retrieval is central to answer usefulness [9]. Medical coding research using retrieval-augmented language models also illustrates why administrative outputs should be judged against source-grounded expectations rather than surface fluency [22]. In this article's conceptual framing, such evaluation should be described qualitatively through expert review rather than through fabricated performance figures.

User experience and time saved

User experience evaluation should examine whether staff can ask natural questions, understand the response, inspect the cited source, and decide what to do next without excessive cognitive effort. Systematic reviews of healthcare conversational agents show that usability, acceptance, and workflow alignment are recurring evaluation concerns [1, 3], and DISCOVER framework interviews emphasize that conversational-agent evaluation should reflect the real context in which the system is used [22]. Task-based comparisons could ask staff to locate a policy through current practice and then through the assistant, but any reported findings would need to come from an actual deployment rather than from conceptual claims. Qualitative feedback from nurses, bed managers, administrators, and policy owners would be especially important for understanding whether the assistant supports real operational work.

Operational impact

Operational impact should be evaluated by examining whether the assistant helps staff find current policies, reduces unnecessary escalation for routine questions, and improves confidence in interpreting local rules. Clinical trial screening with RAG shows how retrieval-augmented systems can support operationally meaningful decision processes by matching

questions to relevant criteria [8], while natural-language database interfaces such as Criteria2Query show that conversational access to structured institutional information can support practical query workflows [4]. Real-time dashboard integration should also be evaluated for whether users understand the distinction between a policy rule and a live operational metric. Any claims about reductions in help-desk tickets, protocol violations, or staff time burden should be treated as future evaluation targets rather than assumed outcomes.

Table 2 outlines the safety, quality-control, human-review, and governance requirements needed to deploy a hospital policy RAG assistant without encouraging unsupported, outdated, or over-trusted operational recommendations.

Table 2. Safety, Quality Control, Human Review, and Governance Framework for a Hospital Policy RAG Assistant

Risk or Implementation Challenge	How It Could Affect Hospital Operations	Manuscript-Specific Control Mechanism	Evaluation Approach	Responsible Stakeholders	Practical Implementation Action
Unsupported answer generation	Staff may act on a fluent but unsupported answer about staffing, bed placement, escalation, or policy exceptions.	Restrict generation to retrieved policy passages and permitted dashboard outputs; require refusal when evidence is insufficient.	Expert review of answer faithfulness, citation accuracy, and refusal behavior.	Informatics team, operational leaders, policy owners, frontline users.	Pilot with narrow use cases and test whether answers are grounded before broad deployment.
Outdated policy retrieval	The assistant may retrieve superseded policies or obsolete staffing rules.	Metadata tracking for effective date, revision date, approval body, and policy owner.	Audit retrieved sources against current policy registry.	Policy governance office, department leaders, compliance team.	Establish document freshness checks before indexing and recurring policy-owner review.
Conflicting policy sources	Different documents may give inconsistent instructions for escalation, unit coverage, staffing thresholds, or bed use.	Contradiction warning and escalation pathway to responsible operational leader or policy owner.	Scenario testing using known conflicting or overlapping policies.	Policy owners, hospital administration, clinical operations leaders.	Require the assistant to show uncertainty rather than synthesize a single unsupported answer.
Weak retrieval quality	The assistant may retrieve irrelevant or incomplete passages because of poor chunking, unclear wording, scanned documents, or local jargon.	Conservative chunking, metadata enrichment, re-ranking, and separate retrieval-quality evaluation.	Retrieval recall and relevance review by domain experts.	Health informatics, knowledge management, departmental policy leads.	Clean legacy documents, standardize headings, and refine chunking around scope, conditions, exceptions, and procedure steps.

Hallucination despite retrieval	The model may add unsupported interpretation beyond the cited evidence.	Prompt constraints, answer-grounding checks, citation-by-claim requirement, and refusal rules.	Compare generated answers with source passages for faithfulness and unsupported additions.	AI governance committee, informatics team, operational subject-matter experts.	Use answer templates that separate “policy says,” “current dashboard shows,” and “consult leader if unclear.”
Permission and access-control failure	Users may receive restricted staffing, capacity, incident, or operational information beyond their role.	Role-based access, permission-aware retrieval, dashboard authorization, and audit logging.	Access-control testing across simulated user roles.	Information security, compliance, operations leadership, IT.	Integrate with hospital identity management and restrict both source retrieval and dashboard outputs by role.
Misinterpretation of dashboard data as policy	Staff may confuse current operational metrics with institutional rules.	Separate policy evidence from live operational state in the generated answer.	Usability testing to assess whether users understand the distinction.	Bed management, nursing leadership, informatics, dashboard owners.	Display dashboard values in a clearly labeled section such as “Current operational state,” distinct from “Policy source.”
Over-reliance by staff	Users may trust the assistant without reading citations or involving leaders in complex cases.	Visible citations, limitations statement, uncertainty flags, and escalation guidance.	User interviews, observation, and review of high-stakes query logs.	Clinical leaders, operations managers, patient safety team.	Train users that the assistant supports policy navigation but does not replace judgment or governance.
Poor workflow fit	Staff may not use the assistant if it is difficult to access during real operational work.	Embed the assistant in the intranet, secure mobile app, or approved internal communication platform.	Task-based usability testing and time-to-answer comparisons.	Frontline staff, UX team, hospital operations, IT.	Start with high-frequency workflows such as staffing policy lookup, bed escalation, and protocol clarification.
Incomplete policy corpus	The assistant may fail because key policies are missing, scanned, duplicated, or inconsistently formatted.	Document inventory, source approval workflow, ingestion validation, and missing-source alerts.	Coverage analysis comparing common operational questions against indexed documents.	Policy governance office, departments, health information management.	Conduct document hygiene work before deployment and maintain a controlled ingestion pathway.
Ambiguous multilingual or	Role-based or translated responses may	Preserve the same source evidence across role-	Review translated and role-adapted	Language services, policy owners, frontline	Treat translations and simplified explanations as

role-adapted output	unintentionally change policy meaning.	specific or multilingual responses; avoid adding new interpretation.	answers against the original source.	representatives, compliance.	derivative views of the cited policy, not new policy statements.
Insufficient implementation evidence	Conceptual claims about time saved, fewer escalations, or improved adherence may be overstated without deployment data.	Frame operational impact as a future pilot outcome rather than an assumed result.	Pilot evaluation of search time, user confidence, help-desk burden, protocol adherence, and escalation patterns.	Research team, hospital operations, quality improvement, implementation science experts.	Begin with measurable pilots in policy-dense areas and avoid claiming impact before evaluation.
Audit and accountability gaps	It may be unclear who is responsible when the assistant gives an incomplete or contested answer.	Audit trails for user queries, retrieved sources, generated answers, and dashboard access.	Periodic governance review of logs, flagged answers, and high-risk queries.	AI governance board, compliance, policy owners, IT security.	Define accountability before launch, including ownership of documents, system behavior, and user escalation pathways.

Limitations

Document hygiene and unstructured legacy content

The assistant would be limited by the quality, consistency, and machine-readability of the hospital documents it ingests. RAG-based summarization of clinical records shows that source text must be processed carefully before it can support reliable generated answers [10], while RAG scoping work in medical and nursing domains indicates that document representation affects retrieval quality [14]. Scanned policies, outdated attachments, inconsistent table formatting, duplicated protocols, and local jargon could all reduce the assistant's ability to retrieve the right evidence. This limitation means that technical deployment should be paired with document cleanup, version governance, and policy ownership.

Trust and over-reliance

Even with citations, users may over-trust a fluent answer or fail to inspect the underlying source document when the situation is complex. Studies on physician perceptions of chatbots show that trust, responsibility, and perceived reliability influence professional acceptance [23, 24], and acceptability research indicates that users may vary in how much confidence they place in AI-led services [27]. The assistant should therefore display clear disclaimers, refuse unsupported questions, and encourage consultation with responsible leaders when policies are ambiguous or high-stakes. Regular auditing would be needed to ensure that the assistant remains a policy-navigation aid rather than an informal substitute for governance or professional judgment.

Conclusion

A retrieval-augmented clinical operations assistant could provide a conversational interface to the policies, staffing guidelines, bed management rules, escalation pathways, and operational dashboards that hospital staff consult every

day. Its central function would be to translate natural-language questions into grounded answers drawn from the institution's own approved sources.

The key strengths of such an assistant would be continuous availability, rapid access to institutional knowledge, transparent citation of source documents, and controlled integration with live operational dashboards. By combining document retrieval with structured data access, the assistant could help staff distinguish between what the policy says and what the current operational state shows.

Important challenges would remain, especially around legacy document quality, inconsistent policy formatting, staff trust, cultural adoption, and the need for continuous curation. The assistant would require governance by policy owners, operational leaders, informatics teams, and frontline users to ensure that its knowledge base remains accurate and useful.

Pilot deployments in large, policy-dense hospitals should examine whether this kind of assistant improves staff efficiency, supports protocol adherence, and reduces the friction of everyday operational decision-making. Such pilots should begin with narrow, well-governed use cases before expanding to broader hospital operations.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 23 Nov 2024

Revised: 15 Dec 2024

Accepted: 29 Jan 2025

Published online: 25 February 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc.* 2018;25(9):1248-58.

Tudor Car L, Dhinakaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res.* 2020;22(8):e17158.

Milne-Ives M, De Cock C, Lim E, Shehadeh MH, De Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. *J Med Internet Res.* 2020;22(10):e20346.

Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc.* 2019;26(4):294-305.

Wang P, Shi T, Reddy CK. Text-to-SQL generation for question answering on electronic medical records. In: *Proc Web Conf.* 2020. p. 350-61.

Ziletti A, D'Ambrosi L. Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records. In: *Proc Clin Nat Lang Process Workshop.* 2024. p. 47-53.

Zakka C, Shad R, Chaurasia A, Dalal AR, Kim JL, Moor M, et al. Almanac—retrieval-augmented language models for clinical medicine. *NEJM AI.* 2024;1(2):Aloa2300068.

Unlu O, Shin J, Mailly CJ, Oates MF, Tucci MR, Varugheese M, et al. Retrieval-augmented generation-enabled GPT-4 for clinical trial screening. *NEJM AI.* 2024;1(7):Aloa2400181.

Li M, Kilicoglu H, Xu H, Zhang R. BiomedRAG: a retrieval augmented large language model for biomedicine. *J Biomed Inform.* 2025;162:104769.

Alkhalaf M, Yu P, Yin M, Deng C. Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. *J Biomed Inform.* 2024;156:104662.

Ng KK, Matsuba I, Zhang PC. RAG in health care: a novel framework for improving communication and decision-making by addressing LLM limitations. *NEJM AI.* 2025;2(1):Alra2400380.

Li H, Huang J, Ji M, Yang Y, An R. Use of retrieval-augmented large language model for COVID-19 fact-checking: development and usability study. *J Med Internet Res.* 2025;27:e66098.

Baur D, Ansorg J, Heyde CE, Voelker A. Development and evaluation of a retrieval-augmented generation chatbot for orthopedic and trauma surgery patient education: mixed-methods study. *JMIR AI.* 2025;4:e75262.

Miao Y, Zhao Y, Luo Y, Wang H, Wu Y. Improving large language model applications in the medical and nursing domains with retrieval-augmented generation: scoping review. *J Med Internet Res.* 2025;27:e80557.

Es S, James J, Anke LE, Schockaert S. RAGAS: automated evaluation of retrieval augmented generation. In: *Proc Eur Chapter Assoc Comput Linguist Syst Demonstr.* 2024. p. 150-8.

Tung JY, Le Q, Yao J, Huang Y, Lim DY, Sng GG, et al. Performance of retrieval-augmented generation large language models in guideline-concordant prostate-specific antigen testing: comparative study with junior clinicians. *J Med Internet Res.* 2025;27:e78393.

Bingöl FG, Ağagündüz D, Bingöl MC. Accuracy of current large language models and the retrieval-augmented generation model in determining dietary principles in chronic kidney disease. *J Ren Nutr.* 2025;35(3):401-9.

Lammert J, Dreyer T, Mathes S, Kuligin L, Borm KJ, Schatz UA, et al. Expert-guided large language models for clinical decision support in precision oncology. *JCO Precis Oncol.* 2024;8:e2400478.

Laymouna M, Ma Y, Lessard D, Schuster T, Engler K, Lebouché B. Roles, users, benefits, and limitations of chatbots in health care: rapid review. *J Med Internet Res.* 2024;26:e56930.

Martinengo L, Lin X, Jabir AI, Kowatsch T, Atun R, Car J, et al. Conversational agents in health care: expert interviews to inform the definition, classification, and conceptual framework. *J Med Internet Res.* 2023;25:e50767.

Ge J, Sun S, Owens J, Galvez V, Gologorskaya O, Lai JC, et al. Development of a liver disease-specific large language model chat interface using retrieval-augmented generation. *Hepatology.* 2024;80(5):1158-68.

Klang E, Tessler I, Apakama DU, Abbott E, Glicksberg BS, Arnold M, et al. Assessing retrieval-augmented large language models for medical coding. *NEJM AI.* 2025;2(10):AIcs2401161.

Koman J, Fauvelle K, Schuck S, Texier N, Mebarki A. Physicians' perceptions of the use of a chatbot for information seeking: qualitative study. *J Med Internet Res.* 2020;22(11):e15185.

Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. *J Med Internet Res.* 2019;21(4):e12887.

Ferber D, Wiest IC, Wölflein G, Ebert MP, Beutel G, Eckardt JN, et al. GPT-4 for information retrieval and comparison of medical oncology guidelines. *NEJM AI.* 2024;1(6):AIcs2300235.

Parameswaran V, Bernard J, Bernard A, Deo N, Tsung S, Lyytinen K, et al. Evaluating large language models and retrieval-augmented generation enhancement for delivering guideline-adherent nutrition information for cardiovascular disease prevention: cross-sectional study. *J Med Internet Res.* 2025;27:e78625.

Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: a mixed-methods study. *Digit Health.* 2019;5:2055207619871808.