

ORIGINAL RESEARCH

Open access

Conversational Artificial Intelligence Assistant for Generating Structured Quality Improvement Reports from Incident Narratives, Safety Event Classifications, Root-Cause Analysis Notes, and Performance Metrics

Ivan Petrov¹, Olga Ivanova^{1*}, Dmitry Smirnov²

Abstract

Quality improvement reports distill incident narratives, safety classifications, root-cause analyses, and performance metrics into actionable learning documents. However, compiling these materials remains a manual, cognitively burdensome task that can delay organizational learning after safety events. Healthcare organizations often hold rich safety data across reporting systems, RCA documents, dashboards, and governance records. Yet these inputs are rarely transformed into standardized QI reports through a single coherent workflow. This article proposes a conversational artificial intelligence assistant that engages quality officers in a structured dialogue, retrieves relevant safety-event evidence, and generates a draft QI report following a pre-specified template. The assistant is conceptualized as a human-supervised system rather than an autonomous decision-maker. The proposed assistant includes an incident narrative NLP module, safety classification aligner, RCA note retriever, performance metric trend summarizer, and template-guided large language model. These components would support structured reporting while preserving human review and organizational accountability. The assistant could shorten the time from incident review to report drafting, improve reproducibility across QI documentation, and reduce administrative burden for patient safety teams. Its value would depend on careful grounding, privacy protection, verification workflows, and user trust. Conversational AI offers a pathway toward AI-augmented safety reporting that supports, rather than replaces, human expertise. The proposed model emphasizes structured synthesis, transparent evidence use, and a learning culture in healthcare quality improvement.

Keywords Large language models, Patient safety, Incident reporting, Conversational artificial intelligence, Quality improvement, Root-cause analysis

*Correspondence:

Olga Ivanova
olga.ivanova@outlook.com

¹ Department of Health Informatics and AI Systems, Faculty of Medicine, Lomonosov Moscow State University, Moscow, Russia

² Department of Intelligent Healthcare Engineering, Faculty of Engineering, Saint Petersburg State University, Saint Petersburg, Russia

Introduction

Healthcare quality improvement teams must often convert fragmented safety information into coherent reports that explain what happened, why it happened, and what corrective actions should follow. Incident narratives, medication event descriptions, and safety reports contain substantial contextual detail, but their review depends on scarce human expertise

and may be slowed by the volume and heterogeneity of submitted reports [1, 2]. Natural language processing studies have shown that incident reports can be classified by event type and severity, suggesting that parts of the review process can be computationally supported [3, 4]. However, the final synthesis into a structured QI report remains largely manual, requiring interpretation, prioritization, and narrative organization.

Existing digital tools can capture safety events, display dashboards, and support committee review, but they often provide limited assistance in transforming raw data into an actionable learning document. Dashboard-based safety systems may improve visibility of patient safety data, yet visual display alone does not generate a coherent explanation of contributing factors, corrective actions, and follow-up responsibilities [5]. Studies of critical incident reporting systems further indicate that NLP can support analysis of event narratives, but these approaches usually focus on classification or extraction rather than full report generation [6, 7]. As a result, safety teams may still face a gap between data availability and report-ready synthesis.

Large language models and conversational agents introduce a new possibility for structured healthcare documentation because they can respond to user instructions, summarize complex text, and generate template-constrained prose. Medical LLM studies demonstrate broad potential for clinical knowledge representation, workflow support, and administrative documentation, while also emphasizing the need for governance, verification, and careful deployment [8–12]. Early clinical documentation studies suggest that generative systems may assist with drafting and summarization tasks when human users retain responsibility for review [13, 14]. These capabilities create a plausible foundation for an assistant that helps quality officers turn safety data into QI reports without presenting itself as an independent adjudicator of harm or causality.

The thesis of this article is that a conversational AI assistant could transform safety-event raw data into structured QI reports through guided dialogue, retrieval-augmented generation, and human-in-the-loop verification. Rather than replacing incident review committees or RCA teams, the assistant would assemble relevant narratives, classifications, RCA findings, and performance metrics into a draft that users can interrogate and revise. Its value would come from combining NLP-based incident understanding with LLM-based report drafting and evidence retrieval. This article therefore presents a conceptual EAI framework for a conversational assistant that supports standardized, auditable, and human-supervised QI documentation.

Background

Quality improvement and patient safety reporting

Quality improvement reporting typically begins with the capture of an incident, near miss, adverse event, or performance concern, followed by classification, review, causal analysis, and action planning. Patient safety event reports often contain unstructured narratives that describe contextual details not captured in coded fields, making them valuable but difficult to synthesize at scale [1, 2]. RCA notes and committee review records add interpretive layers that connect event descriptions to contributing factors, system vulnerabilities, and proposed corrective actions [15, 16]. A structured QI report must therefore integrate factual event details, safety taxonomy information, causal interpretation, and measurable follow-up plans into one document.

NLP and machine learning for incident narratives

Prior work on NLP and machine learning for incident narratives has largely focused on automatic classification, severity identification, event-type labeling, and extraction of contributing factors. Supervised models have been used to classify patient safety incident reports by type and severity, showing that narrative text can support structured safety analysis when appropriately encoded [2, 3, 17]. Semantic representations and text-mining methods have also been explored for improving incident categorization and medication-error classification [18, 19]. These studies provide an important

foundation for a report-generation assistant because they demonstrate that unstructured safety narratives can be transformed into structured signals that guide subsequent synthesis.

Large language models and conversational agents in healthcare

Large language models have expanded the scope of healthcare AI from classification and prediction toward interactive documentation, summarization, and workflow support. Reviews of LLMs in medicine emphasize their potential to generate clinical text, answer domain-specific questions, and support administrative tasks, while also warning that output quality depends on context, prompting, oversight, and safety controls [8, 10, 11]. Studies assessing clinical workflow utility and documentation support indicate that conversational AI may be useful when embedded within human review processes rather than deployed as an autonomous source of truth [12, 13]. For QI reporting, this implies that the assistant should be designed as a drafting and synthesis partner for quality officers, not as a final arbiter of patient safety conclusions.

Retrieval-augmented generation for fact-grounded reports

Retrieval-augmented generation is relevant to QI reporting because safety reports must remain grounded in documented events, RCA findings, local templates, and performance metrics. Medical evidence summarization research shows that LLM outputs require careful assessment for factuality, completeness, and faithful representation of source material [20]. A QI assistant could use retrieval to supply the LLM with incident narratives, relevant RCA excerpts, prior QI reports, safety taxonomies, and approved organizational phrasing before generating report sections. This architecture would be expected to reduce unsupported claims and make the drafting process more traceable than unconstrained generation.

Human-AI collaboration in high-stakes safety work

High-stakes patient safety work requires transparent, verifiable, and accountable AI support because misclassification or unsupported narrative synthesis could distort organizational learning. Ethical analyses of generative AI in healthcare emphasize human oversight, governance, transparency, and protection against unsafe automation [21, 22]. Human-AI collaboration studies in safety event classification further suggest that algorithms can assist review workflows when their outputs are presented for expert validation rather than automatic acceptance [23]. Accordingly, a conversational QI assistant should preserve user control, expose evidence links, log AI-generated content, and require human sign-off before any report enters a governance system.

Conversational Assistant Overview

High-level interaction flow

The proposed assistant would begin when a quality officer selects an incident, cluster of related events, or performance topic for reporting. The system would retrieve the relevant incident narrative, classification fields, RCA notes, and associated metric trends, then ask targeted clarifying questions to resolve ambiguity before drafting the report [16, 23]. The interaction would be iterative: the user could confirm the primary contributing factor, request a more concise background section, or ask the assistant to include a recurring theme from prior RCA reviews. This dialogue-based process aligns with emerging views of LLMs as workflow assistants that help users organize and refine complex documentation rather than independently producing final clinical or safety judgments [8, 12].

Figure 1 presents the proposed left-to-right operational workflow for a grounded conversational AI assistant that converts safety-event evidence into a structured, human-approved QI report.

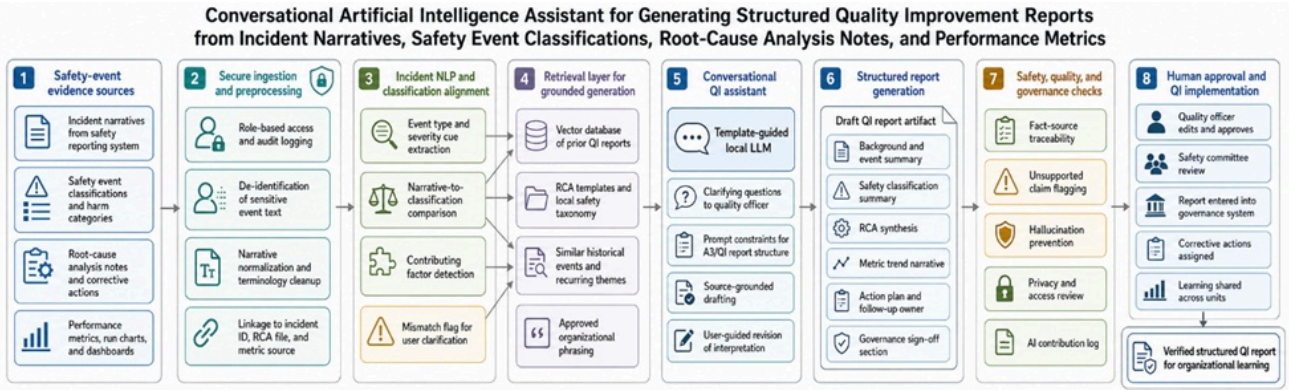


Figure 1. Grounded conversational AI workflow for generating structured healthcare quality improvement reports from safety-event evidence

Core inputs and outputs

The assistant’s inputs would include incident narratives, safety event classification codes, harm or severity categories, RCA notes, performance metric dashboards, and the user’s conversation turns. The output would be a structured QI report in an A3-style format, including background, problem statement, event summary, metric trend summary, root-cause analysis, proposed countermeasures, follow-up plan, and governance notes. Existing NLP studies on incident classification, contributing factor extraction, and critical incident analysis support the feasibility of converting raw narrative fields into structured elements for downstream use [2, 17, 24]. The generated report would remain a draft until the quality officer verifies its claims, revises its interpretation, and approves it for organizational use.

Table 1 summarizes the input-to-output generative workflow through which incident narratives, safety classifications, RCA findings, and performance metrics are transformed into a structured QI report draft.

Table 1. Input-to-output generative workflow for the conversational QI report assistant

Workflow stage	Operational input	AI processing function	Generated or structured output	Human user interaction	Practical value for QI reporting
Safety event selection	Incident ID, safety event title, reporting unit, event date, event status	Identifies the focal event or event cluster for report drafting	Event-specific report workspace	Quality officer selects the event or topic to begin the dialogue	Starts report generation from a traceable operational record rather than a blank document
Incident narrative ingestion	Free-text incident description, near-miss report, adverse event narrative, medication event note	De-identifies, normalizes, segments, and prepares narrative text for interpretation	Cleaned narrative summary with source linkage	User confirms whether the narrative reflects the event under review	Preserves frontline context while reducing manual reading burden

Safety classification alignment	Event type, harm category, severity level, contributing factor code, reporting taxonomy	Compares coded classifications with narrative evidence and flags possible mismatches	Classification summary with uncertainty notes	User resolves mismatches between coded fields and narrative content	Improves consistency between official safety coding and report narrative
RCA knowledge retrieval	RCA notes, causal statements, contributing factors, corrective actions, committee findings	Retrieves relevant RCA excerpts and prior similar themes	Candidate root-cause synthesis and recurring theme suggestions	User accepts, edits, or rejects AI-suggested causal interpretations	Supports institutional memory without allowing the assistant to invent root causes
Performance metric retrieval	Event rates, process compliance measures, run-chart data, SPC views, days since last event	Retrieves relevant metric context and generates cautious descriptive trend language	Metric summary and draft visualization caption	User confirms whether the metric interpretation is appropriate	Connects safety narrative to measurable QI monitoring
Prompt-guided report drafting	Template rules, A3 sections, organizational reporting requirements, user instructions	Uses constrained prompts to generate structured QI report sections	Draft background, problem statement, analysis, action plan, and follow-up plan	User requests revisions, adds context, or clarifies priorities	Converts scattered safety evidence into a report-ready structure
Conversational clarification	User dialogue, unresolved fields, unclear causal links, missing action owners	Asks targeted questions to fill gaps and reduce ambiguity	Revised draft with user-confirmed assumptions	User provides missing information and confirms interpretation	Prevents unsupported model assumptions from entering the report
Safety and quality checks	Source links, incident record, RCA excerpts, metric fields, AI-generated text	Checks traceability, unsupported claims, privacy risks, and template completeness	Review-ready report with flags and evidence markers	User reviews flagged content before approval	Supports auditability, accountability, and safer generative AI use
Final report artifact	Verified draft, human edits, governance sign-off fields	Produces final structured QI report for local workflow entry	Approved report for safety governance and improvement tracking	Quality officer or committee signs off	Moves from AI-assisted drafting to human-owned organizational learning

Design principles

The assistant should be template-driven, user-customizable, grounded in retrieved evidence, conversationally efficient, and aligned with recognized safety frameworks and local reporting norms. Template enforcement is important because QI reports must be reproducible across departments, while user customization is necessary because safety programs vary in

terminology, governance pathways, and documentation expectations [5, 25]. Evidence grounding is equally important because LLMs can produce fluent but unsupported text if they are not constrained by source material [20, 21]. Therefore, the assistant should combine structured reporting rules with retrieval, explicit evidence references, and mandatory human review.

Data Sources and Incident Narrative Processing

Extracting and preprocessing incident narratives

Incident narratives would be extracted from the organization's safety-reporting system and processed through de-identification, normalization, section detection, and terminology harmonization before being presented to the LLM. NLP work on patient safety narratives has shown that free-text reports contain useful signals for classifying event type, severity, and contributing context, but these signals require careful preprocessing because reports may be incomplete, inconsistent, or locally idiosyncratic [1, 4]. De-identification would be especially important because safety narratives may contain patient identifiers, staff names, locations, and sensitive contextual details. The assistant should therefore treat narrative preprocessing as a safety-critical step that supports both privacy protection and accurate downstream synthesis.

Safety event classification and severity encoding

Safety event classification and severity encoding would translate manually assigned or NLP-predicted labels into a structured taxonomy that the assistant can use during report generation. Prior multiclass and neural network approaches indicate that incident reports can be categorized by type and severity from narrative content, although such outputs should be interpreted as decision support rather than definitive classification [2, 3]. Semantic representation methods may further improve alignment between narrative descriptions and standardized safety concepts [18]. In the proposed assistant, classification fields would help frame the report, but discrepancies between coded labels and narrative evidence should be highlighted for user clarification rather than silently resolved by the model.

Structuring root-cause analysis notes

Root-cause analysis notes would be parsed into causal statements, contributing factors, human factors, system conditions, and proposed corrective actions to support retrieval during report drafting. NLP-based approaches to contributing factor categorization and diagnostic safety review suggest that safety learning documents can be transformed into structured information that supports thematic analysis [16, 25]. The assistant could retrieve similar RCA themes from prior cases and present them as candidate language for the user to accept, revise, or reject. This would help preserve institutional memory while preventing the system from inventing causal explanations that are not supported by reviewed evidence.

Performance metrics and trend data retrieval

Performance metrics would be retrieved from QI dashboards, event-rate systems, run-chart repositories, statistical process control views, or manually maintained safety metric tables. Prior work on patient safety dashboards indicates that visual displays can organize safety data for monitoring, but they do not necessarily convert trends into narrative interpretation or action-oriented reporting [5]. The assistant could generate descriptive captions for metric trends, identify the relevant measure name, and distinguish between observed change, possible signal, and user-confirmed interpretation. Because metric interpretation can be context-dependent, the assistant should avoid claiming improvement, deterioration, or causality unless the user confirms that the underlying analytic criteria support such statements.

Conversational AI Architecture and Report Generation

Large language model core and prompt engineering

The assistant would use a locally governed LLM core with prompts that enforce the structure, tone, and required sections of an A3-style QI report. Prompt engineering would specify that the model must distinguish documented facts, user interpretations, proposed actions, and unresolved uncertainties, thereby reducing the risk that fluent prose is mistaken for verified safety analysis [11, 21]. Medical LLM research supports the idea that domain-specific prompting and workflow constraints are necessary for safe healthcare use, especially when generated text may influence clinical or organizational decisions [8, 9]. In this framework, the LLM would be a constrained drafting engine whose output remains subject to quality officer verification.

Retrieval-augmented generation for QI knowledge

The retrieval layer would connect the LLM to prior QI reports, RCA templates, safety taxonomies, incident records, approved terminology, and relevant institutional guidance. Evidence summarization studies show that LLM-generated summaries must be checked for accuracy, source fidelity, and omission of important details, which makes retrieval and verification central to any safety-reporting application [20]. A vector database could retrieve analogous cases, recurring causal themes, and standard language for action plans, while the generation model would adapt those materials to the current incident. This design would be expected to improve factual grounding and consistency, but retrieved content should always be displayed to the user for inspection.

Dialogue manager and clarification loop

A dialogue manager would maintain conversational context, track unresolved report sections, and ask focused questions when the retrieved evidence is incomplete or ambiguous. For example, the assistant could ask which contributing factor should be emphasized, whether the event classification reflects the final committee decision, or whether a proposed action has already been implemented. Clinical workflow studies of LLMs suggest that conversational systems may be most useful when they reduce documentation friction while allowing users to guide and correct the output [12, 13]. The clarification loop is therefore a core safety feature because it prevents the assistant from filling gaps with unsupported assumptions.

Structured output generation

The final output would be rendered as a structured QI report with sections such as background, event summary, analysis, metric context, action plan, accountability, follow-up, and governance sign-off. The assistant would include source links to incident records, RCA excerpts, metric fields, and user-confirmed statements so that each factual claim can be inspected before approval. Work on ambient clinical documentation and clinical note generation demonstrates the broader feasibility of generating structured healthcare documentation, but also reinforces the importance of benchmarking, usability review, and careful domain adaptation [14, 26]. For QI reporting, structured output generation should therefore emphasize auditability, local template compliance, and human ownership of the final document.

Incorporating Safety Classification and RCA Knowledge

Aligning the incident narrative with event classification

The assistant would compare the free-text incident narrative with the official safety classification, including event type, harm category, contributing domain, and severity level. Prior work on automated classification of patient safety reports suggests that narrative features can support classification, but also that model-derived labels should be treated as review aids rather than final determinations [2, 3, 18]. If the narrative describes medication delay, diagnostic delay, communication breakdown, or workflow failure in ways that diverge from the assigned code, the assistant would flag the discrepancy and ask the user whether the classification should be revised or explained. This alignment function would help the report preserve fidelity to the official safety record while making classification uncertainty visible.

Extracting root causes and identifying recurring themes

The assistant could use NLP-derived representations of RCA notes to surface recurring contributing factors, such as communication gaps, handoff ambiguity, staffing constraints, equipment availability, documentation inconsistency, or unclear escalation pathways. Studies of contributing factor categorization and diagnostic safety learning systems show that narrative safety records can be structured into themes that support organizational learning [16, 25]. Recent feasibility work using generative AI and NLP for critical incident systems also suggests that language models may help identify and organize incident themes, although such outputs require careful review by safety experts [24, 27]. In the proposed assistant, recurring themes would be offered as candidate interpretations rather than imposed as definitive root causes.

Integrating performance metrics and run charts

The assistant would integrate performance metrics by retrieving relevant measures, summarizing trends in cautious language, and generating captions for run charts or statistical process control visuals that the user may attach to the report. Patient safety dashboards can make performance signals visible, but QI reports require narrative interpretation that connects the metric to event context, operational constraints, and proposed countermeasures [5]. The assistant should therefore distinguish between descriptive metric statements, such as an observed increase in events, and interpretive claims, such as a process deterioration requiring confirmed analytic review. This approach would allow metric evidence to support report clarity without overstating causality or substituting automated interpretation for QI expertise.

Report Structuring, Quality, and Safety Guardrails

Template enforcement and A3 framework alignment

The assistant would enforce the selected QI template by ensuring that the report contains the required sections, maintains concise problem framing, and connects root causes to corrective actions and follow-up plans. Generative AI reviews in healthcare emphasize that unconstrained LLM output may be fluent but inconsistent, making structure and governance essential for reliable use in administrative documentation [11, 28]. Template enforcement would also help reduce variation across departments by guiding users back to required elements when the conversation drifts into unsupported narrative or excessive speculation. In this framework, the A3 format would function as both a writing scaffold and a safety control.

Fact-verification and hallucination prevention

Every factual claim in the draft report should be traceable to a source, such as an incident record, RCA excerpt, metric field, governance decision, or explicit user statement. Medical LLM research has repeatedly emphasized that generated outputs require verification because models can omit details, distort source meaning, or produce unsupported assertions even when the prose appears credible [8, 20, 21]. The assistant should therefore mark unsupported statements, separate evidence from interpretation, and invite the user to confirm uncertain causal language before finalization. This guardrail is central to preventing hallucinated explanations from entering the organizational safety record.

Human-in-the-loop approval and sign-off

The completed report would be presented to the quality officer or safety committee for editing, approval, and formal sign-off before it is entered into the safety governance system. Human-AI collaboration in safety event classification shows that AI support is most appropriate when expert users can review, correct, and contextualize model outputs [23]. Clinical documentation studies similarly suggest that generative systems should support drafting efficiency while preserving professional accountability for final content [13, 26]. For QI reporting, all AI-generated text, retrieved source fragments, user edits, and approval actions should be logged to support transparency, accountability, and later audit.

Table 2 outlines the safety, quality-control, hallucination-prevention, human-review, and governance safeguards required for responsible deployment of the conversational QI report assistant.

Table 2. Safety, quality control, hallucination prevention, human review, and governance safeguards for the conversational QI assistant

Risk or governance domain	Why it matters in QI reporting	Required safeguard	Human review responsibility	Implementation mechanism	Expected practical benefit
Patient and staff privacy	Incident narratives and RCA notes may contain protected health information, staff identifiers, rare event details, or sensitive organizational content	Local or private deployment, role-based access, de-identification, audit logging, and restricted retrieval permissions	Privacy officer and QI lead confirm that access rules match institutional policy	Access-control layer, encrypted storage, automatic de-identification, usage logs	Reduces risk of inappropriate exposure of sensitive safety information
Hallucinated causal explanations	Unsupported causal statements could distort safety learning and corrective action planning	Require every causal claim to be linked to an RCA excerpt, user statement, or documented committee finding	RCA facilitator confirms whether each causal claim is supported	Source-linked generation, unsupported-claim flags, mandatory evidence display	Prevents invented root causes from entering governance documents
Misaligned safety classification	Narrative evidence may not match the assigned event type, harm category, or severity code	Cross-check narrative content against coded classifications and flag discrepancies	Quality officer resolves whether the code, narrative, or report wording needs correction	Classification alignment module and discrepancy panel	Improves consistency between coded safety data and final QI report
Overstated metric interpretation	Metric trends may be descriptive and should not be framed as causal improvement or deterioration without review	Use cautious metric language and require user confirmation for causal or improvement claims	QI analyst verifies trend interpretation and chart caption	Metric retrieval layer with controlled language templates	Keeps performance reporting analytically defensible
Template drift	LLMs may generate fluent but off-template or overly broad content	Enforce required A3/QI sections, section length limits, and local reporting language	Quality officer confirms completeness and relevance	Prompt constraints, report schema, section-level validation	Produces consistent reports across departments and events
Loss of source traceability	Safety governance requires the ability	Attach every factual statement to incident ID,	Reviewer checks evidence links before sign-off	Source citation panel, evidence map, audit trail	Makes the report reviewable and defensible

	to verify where each claim originated	RCA line, metric field, or user-confirmed input			
Automation bias	Users may over-trust polished AI-generated prose	Display AI uncertainty, unresolved fields, and review warnings before approval	User must actively approve or revise each high-impact section	Human-in-the-loop approval workflow and sign-off checklist	Reinforces that the assistant drafts but does not decide
Organizational variation	Hospitals differ in taxonomies, RCA templates, QI governance, and reporting language	Support local configuration of templates, classification labels, and approval steps	Local QI governance team validates configuration before deployment	Configurable template library and taxonomy mapping	Improves fit with real-world safety operations
Accountability and auditability	AI-assisted reports may later influence committee decisions, corrective actions, or regulatory review	Log AI-generated text, retrieved sources, user edits, and final approvals	Safety governance committee reviews audit trail when needed	Version control, AI contribution log, approval timestamping	Preserves institutional accountability for final report content
Implementation readiness	A technically strong assistant may fail if it disrupts workflow or increases burden	Pilot in existing safety-event workflow with usability review and staff training	QI leadership monitors adoption, burden, and user trust	Phased deployment, training sessions, workflow integration testing	Supports practical adoption rather than isolated technical demonstration

Integration into Quality Improvement Workflows

Seamless integration with safety event management systems

The assistant would operate as a module within existing safety event management or QI software rather than as a disconnected chatbot. It could be triggered when an incident reaches closure, when an RCA is completed, or when a recurring metric crosses a locally defined review threshold. Prior work on NLP integration with patient safety event review committees suggests that computational tools become more useful when embedded in established review workflows rather than used as isolated analytic systems [1]. Integration should therefore prioritize single sign-on, role-based permissions, source-system traceability, and compatibility with existing report approval processes.

Supporting a learning culture and continuous improvement

By reducing the manual effort required to assemble a thorough QI report, the assistant could encourage more frequent, timely, and consistent documentation of safety learning. However, the system should be framed as a support for learning culture rather than a surveillance or blame tool, because patient safety reporting depends on trust, psychological safety, and credible governance. Ethical discussions of generative AI in healthcare emphasize that organizational context, transparency, and accountability shape whether AI tools are experienced as helpful or threatening [21, 22]. A carefully

governed assistant could therefore support continuous improvement by making learning easier while leaving judgment, prioritization, and accountability with human safety leaders.

Evaluation Strategy

Report quality and completeness

Evaluation should assess whether AI-assisted reports are complete, accurate, actionable, clear, and faithful to the underlying incident record and RCA evidence. Blinded QI experts could compare assistant-generated drafts with human-written reports using structured review criteria, but the evaluation should avoid presenting conceptual claims as experimental outcomes. Prior evidence summarization and clinical documentation studies provide useful evaluation domains, including source fidelity, omission of relevant information, usability, and documentation quality [14, 20]. In this article, the evaluation strategy is proposed conceptually and should be tested prospectively before operational adoption.

User experience and conversational efficiency

User experience evaluation should examine whether quality officers find the assistant efficient, understandable, trustworthy, and easy to correct during report drafting. Measures could include perceived workload, usefulness of clarification questions, ease of editing, and whether the assistant helps users maintain a coherent report structure without losing important context. Clinical workflow assessments of LLMs and broader reviews of generative AI in healthcare suggest that adoption depends not only on technical capability but also on usability, fit with existing work, and confidence in oversight mechanisms [12, 28]. The assistant should therefore be evaluated as a socio-technical intervention, not merely as a text-generation model.

Impact on QI cycle time and safety culture

A future implementation study could examine whether the assistant affects QI cycle time, report completion, action-plan documentation, and staff perceptions of reporting burden. Reviews of LLMs across healthcare applications and current clinical trials indicate that rigorous evaluation is needed before generative AI tools are treated as mature clinical or operational infrastructure [29]. For safety culture, evaluation should consider whether staff perceive the assistant as improving learning, transparency, and follow-through, or whether it creates concerns about automation, monitoring, or loss of narrative nuance. These outcomes should be interpreted cautiously because organizational readiness and governance quality may strongly influence the assistant's real-world effects.

Limitations

Data sensitivity and privacy

Incident narratives and RCA notes may contain protected health information, staff identifiers, location details, and sensitive organizational learning content. Because generative AI systems can introduce privacy, security, and governance risks, the assistant should be deployed within a controlled local or private environment with role-based access, audit logs, and strict data-retention policies [21, 28]. De-identification would reduce risk but cannot be assumed to remove all sensitive information, especially when narratives contain rare events or recognizable contextual details. Privacy protection should therefore be treated as a continuous governance requirement rather than a one-time preprocessing step.

Dependency on input quality and organizational variation

The assistant would depend on the quality, completeness, and consistency of incident narratives, classification fields, RCA templates, and metric systems. Prior studies of incident report NLP indicate that narrative variability, local terminology, and inconsistent documentation can affect the reliability of automated classification and extraction [4, 6, 15].

Organizational variation may also limit portability because each hospital may use different safety taxonomies, QI templates, RCA conventions, and approval workflows. Local customization, user training, and ongoing monitoring would therefore be necessary before the assistant could be responsibly embedded in routine QI work.

Conclusion

A conversational artificial intelligence assistant for structured QI report generation could help healthcare organizations transform fragmented safety information into coherent, actionable learning documents. By engaging quality officers in guided dialogue, the assistant would support synthesis of incident narratives, classifications, RCA findings, and performance metrics without replacing expert judgment.

The key strength of the proposed framework is its combination of interactive drafting, template-driven structure, retrieval grounding, and embedded human oversight. This design would allow the assistant to generate useful report drafts while preserving source traceability and accountability for final interpretation.

Important challenges remain, including privacy protection, organizational readiness, local workflow integration, and the need to evaluate whether generated reports are accurate, useful, and safe. The assistant should be introduced cautiously, with governance processes that prevent unsupported claims from entering the safety record and that maintain trust in reporting systems.

Pilot implementation should begin in safety-net hospitals and other high-volume environments where incident review backlogs and documentation burden are especially consequential. Future randomized or stepped-wedge evaluations should assess report quality, QI cycle time, user trust, and downstream safety-learning outcomes before broader deployment.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Jul 2024

Revised: 09 Sep 2024

Accepted: 02 Nov 2024

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

- Fong A, Harriott N, Walters DM, Foley H, Morrissey R, Ratwani RR. Integrating natural language processing expertise with patient safety event review committees to improve the analysis of medication events. *Int J Med Inform.* 2017;104:120-5.
- Wang Y, Coiera E, Runciman W, Magrabi F. Using multiclass classification to automate the identification of patient safety incident reports by type and severity. *BMC Med Inform Decis Mak.* 2017;17(1):84.
- Wang Y, Coiera E, Magrabi F. Using convolutional neural networks to identify patient safety incident reports by type and severity. *J Am Med Inform Assoc.* 2019;26(12):1600-8.
- Young IJ, Luz S, Lone N. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *Int J Med Inform.* 2019;132:103971.
- Murphy DR, Savoy A, Satterly T, Sittig DF, Singh H. Dashboards for visual display of patient safety data: a systematic review. *BMJ Health Care Inform.* 2021;28(1):e100437.
- Tetzlaff L, Heinrich AS, Schadewitz R, Thomeczek C, Schrader T. Die Analyse des CIRSmedical.de mittels Natural Language Processing. *Z Evid Fortbild Qual Gesundheitswes.* 2022;169:1-1.
- Ozonoff A, Milliren CE, Fournier K, Welcher J, Landschaft A, Samnaliev M, et al. Electronic surveillance of patient safety events using natural language processing. *Health Informatics J.* 2022;28(4):14604582221132429.
- Large Language Models in medicine: Thirunavukarasu AJ, Ting DS, Elangovan K, et al. Large language models in medicine. *Nat Med.* 2023;29(8):1930-40.
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620(7972):172-80.
- Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. *J Med Internet Res.* 2023;25:e48568.
- Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell.* 2023;6:1169595.
- Rao A, Pang M, Kim J, Kamineni M, Lie W, Prasad AK, et al. Assessing the utility of ChatGPT throughout the entire clinical workflow: development and usability study. *J Med Internet Res.* 2023;25:e48659.
- Baker HP, Dwyer E, Kalidoss S, Hynes K, Wolf J, Strelzow JA. ChatGPT's ability to assist with clinical documentation: a randomized controlled trial. *JAAOS.* 2024;32(3):123-9.
- Yim WW, Fu Y, Ben Abacha A, Snider N, Lin T, Yetisgen M. ACI-Bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Sci Data.* 2023;10(1):586.
- Fong A. Realizing the power of text mining and natural language processing for analyzing patient safety event narratives: the challenges and path forward. *J Patient Saf.* 2021;17(8):e834-6.
- Tabaie A, Sengupta S, Pruitt ZM, Fong A. A natural language processing approach to categorise contributing factors from patient safety event reports. *BMJ Health Care Inform.* 2023;30(1):e100731.

Evans HP, Anastasiou A, Edwards A, Hibbert P, Makeham M, Luz S, et al. Automated classification of primary care patient safety incident report content and severity using supervised machine learning approaches. *Health Informatics J.* 2020;26(4):3123-39.

Wang Y, Coiera E, Magrabi F. Can Unified Medical Language System–based semantic representation improve automated identification of patient safety incident reports by type and severity? *J Am Med Inform Assoc.* 2020;27(10):1502-9.

Boxley C, Fujimoto M, Ratwani RM, Fong A. A text mining approach to categorize patient safety event reports by medication error type. *Sci Rep.* 2023;13(1):18354.

Tang L, Sun Z, Ilday B, Nestor JG, Soroush A, Elias PA, et al. Evaluating large language models on medical evidence summarization. *NPJ Digit Med.* 2023;6(1):158.

Oniani D, Hilsman J, Peng Y, Poropatich RK, Pamplin JC, Legault GL, et al. Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare. *NPJ Digit Med.* 2023;6(1):225.

Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt JN, Laleh NG, et al. The future landscape of large language models in medicine. *Commun Med (Lond).* 2023;3(1):141.

Chen H, Cohen E, Wilson D, Alfred M. A machine learning approach with human-AI collaboration for automated classification of patient safety event reports: algorithm development and validation study. *JMIR Hum Factors.* 2024;11(1):e53378.

Denecke K, Paula H. Analysis of critical incident reports using natural language processing. *Ind Health.* 2024:1-6.

Tabaie A, Tran A, Calabria T, Bennett SS, Milicia A, Weintraub W, et al. Evaluation of a natural language processing approach to identify diagnostic errors and analysis of safety learning system case review data: retrospective cohort study. *J Med Internet Res.* 2024;26:e50935.

Heilmeyer F, Böhringer D, Reinhard T, Arens S, Lyssenko L, Haverkamp C. Viability of open large language models for clinical documentation in German health care: real-world model evaluation study. *JMIR Med Inform.* 2024;12:e59617.

Hölzing CR, Rumpf S, Huber S, Papenfuß N, Meybohm P, Happel O. The potential of using generative AI/NLP to identify and analyse critical incidents in a Critical Incident Reporting System (CIRS): a feasibility case-control study. *Healthcare (Basel).* 2024;12(19):1964.

Moulaei K, Yadegari A, Baharestani M, Farzanbakhsh S, Sabet B, Afrash MR. Generative artificial intelligence in healthcare: a scoping review on benefits, challenges and applications. *Int J Med Inform.* 2024;188:105474.

Omar M, Nadkarni GN, Klang E, Glicksberg BS. Large language models in medicine: a review of current clinical trials across healthcare applications. *PLOS Digit Health.* 2024;3(11):e0000662.