International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(23s) 2026 pp. 722-748
https://doi.org/10.70917/ijcisim-2026-5635
Article

TRUST-CARE: A RISK-ADAPTIVE AI GOVERNANCE FRAMEWORK FOR TRUSTWORTHY LARGE LANGUAGE MODEL DEPLOYMENT IN REGULATED HEALTHCARE INFRASTRUCTURE

Shahid Ahmed Qureshi

independent technology Researcher
qur7@yahoo.com

Abstract: Large language models (LLMs) are rapidly transforming the landscape of healthcare, bringing new opportunities for clinical information management, decision support, documentation, and knowledge-intensive services. However, when these models are applied to regulated health care environments, the risks are not just about the accuracy of the models; they need to be addressed. Hallucinated content, privacy exposure, unreliable evidence, inappropriate automation, adversarial inputs, and uncertain model behaviour may compromise the accountability and safety of AI-assisted healthcare processes. Existing governance models are often based on pre-established controls and thereby have limited ability to rapidly adapt protection based on the varying risk levels of individual requests and clinical situations. Therefore, this study presents a TRUST-CARE, risk-adaptive AI governance framework that incorporates the governance process into the lifecycle of LLMs. The proposed system is based on a combination of contextual risk assessment, analysis of evidence reliability, processing of privacy-aware evidence, retrieval-based generation, hallucination verification, confidence calibration, and risk-triggered human oversight. The proposed methodology compares the performance of TRUST-CARE with the baseline configuration, conventional LLM, healthcare LLM, retrieval-augmented generation, and safety-guarded versions of the two baselines on the basis of task performance, evidence grounding, hallucination rate, calibration, privacy exposure, robustness, security, and expert trustworthiness assessment. The proposed architecture helps determine whether an AI response can be generated automatically, when there is a need for further evidence checking, and when human intervention must be mandated. The governance framework model provides a structured mechanism to link model behaviour with evidence quality, clinical risk, uncertainty and accountability. TRUST-CARE thus advances the standard safety controls in healthcare towards risk-sensitive, evidence-based, auditable AI governance under human supervision.

Keywords: Large Language Models; Trustworthy AI; AI Governance; Healthcare AI; Risk-Adaptive AI; Clinical Decision Support; Retrieval-Augmented Generation; Evidence Grounding; Hallucination Detection; Privacy-Preserving AI; AI Safety; Human-in-the-Loop; Confidence Calibration; Healthcare Infrastructure

1. INTRODUCTION

1.1 Background and Motivation

The rapid transition from conventional artificial intelligence to generative artificial intelligence is transforming healthcare information production, interpretation, and dissemination. Beyond general-purpose language modeling, large language models (LLMs) are now being explored for addressing clinical questions, medical documentation, clinical decision support, diagnostic hypothesis generation, medication information, and patient record interpretation.

Recent clinical evaluations have demonstrated the ability of such LLMs to be incorporated into Electronic Medical Records (HER) and tested in a real-world primary-care setting. This represents healthcare generative artificial intelligence entering the clinic beyond just isolated interfaces.

The increasing use of large language models (LLMs) in healthcare is especially interesting given that healthcare decisions are often information-intensive and based on disparate kinds of evidence. Within a single EHR, there might be clinical text, lab results, medication lists, structured observations, radiology reports, and previous diagnoses. An LLM could potentially allow clinicians to quickly absorb this aggregated information and get a coherent summary. Recent studies have demonstrated different applications of LLMs in clinical decision support, identification of medical errors, evidence-based question-answering systems, diagnostic assistance, and report generation. However, these examples illustrate the challenge of getting useful information rather than plausible information from such systems.

Retrieval-augmented generation (RAG) has emerged as an important response to the limitations of knowledge embedded within model parameters. By supplying external clinical information during inference, RAG can improve access to current evidence and provide a basis for more traceable responses. Recent healthcare studies have investigated retrieval-augmented systems for radiology, clinical decision support, informed consent, medication instructions, and medical reporting. Nevertheless, retrieval itself does not guarantee trustworthy generation. A retrieved document may be irrelevant, outdated, incomplete, poorly authoritative, or inconsistent with other available evidence. Consequently, the presence of retrieved context should not be considered equivalent to evidence validation.

This concern is particularly pertinent to healthcare AI, which typically runs in a highly regulated environment. Compared to consumer apps, clinical algorithms must be validated for safety, privacy, liability, auditability, data governance, and professional compliance. Thus, an erroneous response from a healthcare chatbot may have far-reaching consequences beyond the mere informational context. Recent studies evaluating the performance of an LLM-based clinical decision support system embedded in an electronic medical record provide a sobering insight into the potential effects of such technologies. Although the incidence of hallucinations was found to be relatively low, they were nevertheless present across all participating clinics, illustrating that infrequency does not eliminate risk.

The central challenge is therefore shifting. The question is no longer whether an LLM can perform a healthcare task with a sufficient degree of accuracy, but rather whether the accompanying infrastructure can recognize the context of a request and adjust the degree of evidential verification, privacy protection, safety-checking, uncertainty-modelling, and human oversight required to perform it safely.

1.2 Evolution of Healthcare AI

Healthcare artificial intelligence has evolved through distinct phases of technological development. The earliest clinical decision-support systems used rule-based programming to encode medical knowledge in a manner that could be systematically executed by a computer. These systems were typically interpretable but had limited scope due to the challenges of encoding medicine in explicit rules and the inability to generalize beyond specific scenarios.

The rise of machine learning enabled the development of less interpretable but more powerful clinical models, which could detect patterns in training data and generalize their findings to new cases. Deep learning further extended these capabilities, enabling models to learn patterns from images, signals, natural language, and structured data.

However, all these approaches were largely task-specific, with only a narrow focus on a given prediction or analytical task.

The recent foundation-models paradigm began to address this limitation, enabling the development of large-scale language models and systems that could perform a variety of tasks. The recent advances in large language models have seen the emergence of a new paradigm of healthcare AI, namely, generative artificial intelligence for clinical information processing.

This has enabled systems that can summarize, provide answers to questions, document findings, and support clinical decision-making in dialogues. The next major paradigm shift was driven by retrieval-augmented question-answering systems that linked large language models to relevant external information during model inference.

The latest developments in healthcare artificial intelligence are focused on large language models that can interact with multiple external information sources as agents to perform a variety of clinical tasks. A number of recent studies have highlighted the value of evidence-aware dialog systems that can rank different pieces of information based on their relevance to the input questions and the likelihood of their correctness.

This progression can therefore be represented as in Figure 1:

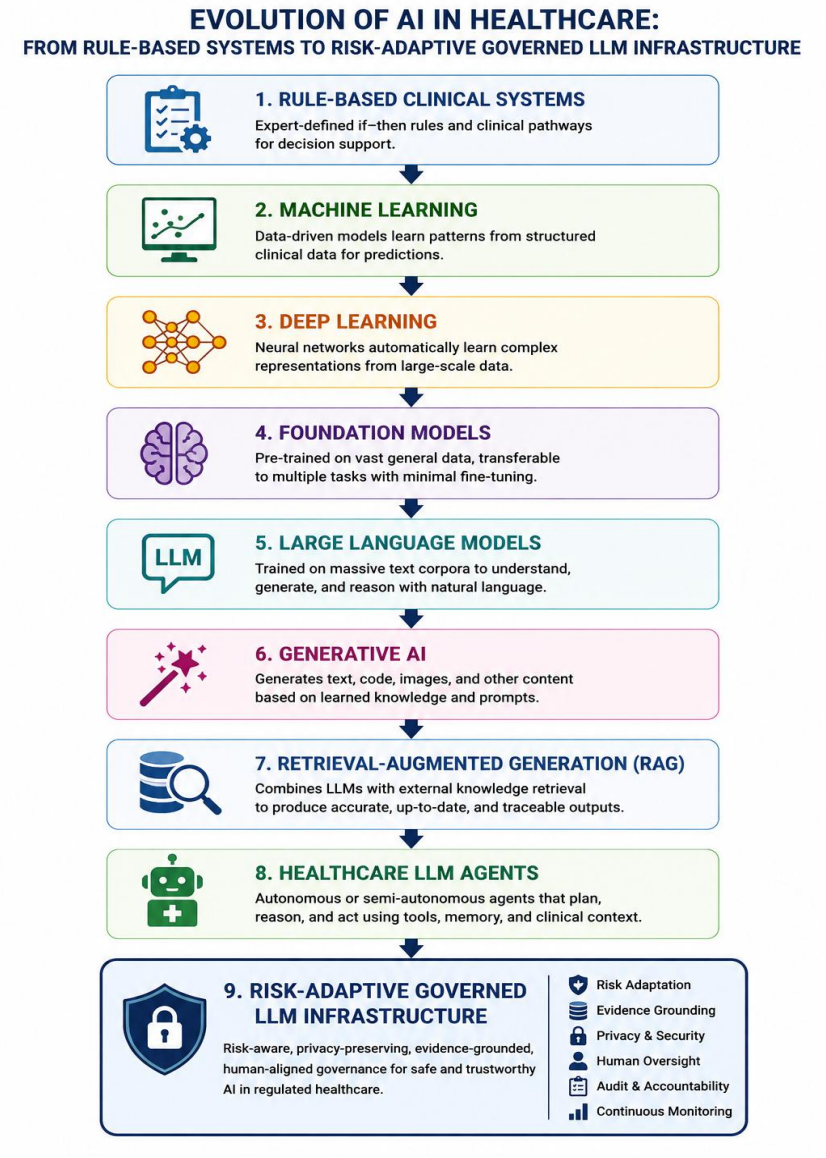

Figure 1. Evolution of AI in Healthcare

The final transition is of particular importance. Earlier AI governance tends to relate to more stable systems with more fixed tasks. By contrast, modern large language models are more fluid: their outputs are determined by prompts, the information they retrieve, the context of the patient, the model parameters, external tools, and continually

updated knowledge bases. A static governance framework may be inadequate in such a dynamic environment, where the risks in any individual case may well change from one interaction to the next.

1.3 Problem Statement

Although promising, healthcare LLM systems present a set of technical and governance-related challenges. A model may generate hallucinated information, unsubstantiated clinical claims, or an apparently reasonable suggestion that was not sufficiently supported by the available evidence. Evidence-based assessments of language models have found that they perform better on certain categories of clinical questions and can be significantly challenged when the evidence is absent or contradictory.

Additionally, knowledge can become outdated: treatment regimens, guidelines, diagnostic criteria, and other clinical -information change over time, while a model's weights do not, unless it undergoes repeated training. The use of RAG is one potential solution, but it raises additional evidence-relevance assessment challenges, such as the need to evaluate the retrieved claims for reliability, authority, recency, and consistency of context.

The deployment of such models within healthcare also raises a number of additional unique concerns, including the risks of privacy leakage, prompt injection, data poisoning or bias, unsafe levels of confidence, inadequate uncertainty quantification, and inconsistent responses. With respect to bias, evidence has found that the distribution of capability does not equal the distribution of capability across domains, and models are not intrinsically capable of providing equitable medical advice. Finally, the risk is not limited to the flaws of a particular model at the time of deployment. Rather, it touches on the fundamental governance challenges of the system, including evidence, patient data, model behaviour, and risk management.

The problem is consequently not a single model-level defect. It is an infrastructure-level governance problem involving the interaction between patient data, evidence sources, model behaviour, risk, uncertainty, security, and human decision-making.

The research problem that is addressed by this paper can be formulated as follows:

How can the deployment infrastructure for an LLM be designed to support dynamic evidence evaluation, privacy protection, safety control, and human oversight calibrated to the risk level of a particular healthcare interaction?

1.4 Research Gap

- Many safety controls rely on fixed rules or unchanging boundaries. While these can offer a basic level of protection, they often apply the same level of scrutiny regardless of the specific clinical context. A simple administrative inquiry and a high-stakes treatment suggestion, for instance, may end up going through the same review process, even though their risks differ significantly.
- Although retrieval-augmented systems enhance access to external data, retrieving information does not guarantee its credibility. Some healthcare applications using retrieval have shown better accuracy and user trust, but challenges remain in consistently judging whether each retrieved source is trustworthy enough for the clinical decision at hand.
- Large language models can produce responses with strong, confident wording, even when the underlying evidence is weak or incomplete. As a result, a model's apparent certainty should not be mistaken for clinical accuracy. Studies show performance varies based on the quality and structure of available evidence, underscoring the importance of separating the model's confidence from the actual strength of supporting data.
- Effective healthcare AI needs enough patient information to generate meaningful outputs, but must also limit access to sensitive data. Overly strict data limitations can reduce the system's usefulness, while too much

access increases privacy risks. This balance cannot be managed effectively with a one-size-fits-all privacy setting.

- Traditional governance models are often applied around the AI system rather than integrated into its real-time operation. Today's healthcare LLMs need oversight that functions during inference—assessing the input query, supporting evidence, patient context, risk level, response content, and uncertainty—before deciding whether to release the output.
- Human review remains critical in high-risk medical scenarios, but expecting clinicians to manually evaluate every AI-generated response is impractical and hard to sustain at scale. A more viable strategy involves flagging outputs for expert review only when the system identifies potential issues such as high risk, low evidence, uncertainty, contradictions, or harmful suggestions.

Collectively, these gaps highlight the need for a governance framework that goes beyond simply enclosing an LLM, instead actively guiding and adjusting its behavior across all stages of deployment.

1.5 Research Motivation

This research is motivated by the increasing gap between the capabilities of AI models and the level of trust warranted in their real-world deployment. While recent clinical studies indicate that generative AI has the potential to enhance or support healthcare processes, they also highlight ongoing concerns regarding safety, the strength of supporting evidence, and the need for rigorous clinical validation.As a result, the study moves beyond asking simply "How capable is the LLM?" and instead addresses a more comprehensive, systems-level question:"Under what specific conditions should an LLM be trusted, restricted, checked, escalated to human review, or blocked from making independent recommendations?"

This shift in perspective underpins the creation of TRUST-CARE—a governance framework designed to adapt to risk by integrating clinical context, evidence reliability, privacy risks, model uncertainty, security factors, and human oversight into a unified deployment structure.

1.6 Research Objectives

The study aims to achieve the following objectives:

- To design a healthcare LLM governance framework that adapts to risk levels by adjusting oversight mechanisms based on the specific nature of each AI interaction.
- To create a dynamic system for classifying risks in healthcare AI applications, incorporating factors such as task importance, possible harm, uncertainty, patient data sensitivity, and situational context.
- To develop an evidence-reliability assessment mechanism that evaluates the relevance, authority, temporal validity, provenance, and consistency of retrieved healthcare evidence.
- To integrate privacy-aware and security-aware inference controls for protecting sensitive healthcare information against inappropriate exposure and adversarial manipulation.
- To develop hallucination and unsupported-claim verification capable of examining generated statements against their supporting evidence before final response delivery.
- To develop confidence-aware decision generation that distinguishes model confidence from evidence strength and explicitly represents uncertainty.
- To develop a risk-triggered human-in-the-loop escalation mechanism that directs uncertain or high-risk outputs to appropriate human review.

- To evaluate the proposed TRUST-CARE framework against existing healthcare LLM approaches using task performance, evidence grounding, hallucination, calibration, privacy, security, robustness, and expert-assessed trustworthiness measures.
- Together, these goals establish TRUST-CARE not merely as another application of LLMs, but as a deployment framework focused on governance, where risk assessment and evidence validation are integral to AI-driven healthcare decision-making. As the field shifts from controlled testing environments to actual clinical use, infrastructure-level protections like these are becoming more essential.

2. RELATED WORK AND RESEARCH GAP

2.1 Large Language Models in Healthcare

LLMs are rapidly evolving from general-purpose language modeling to clinical documentation, medical question answering, diagnostic assistance, report generation, and patient communication. Recent clinical investigations show that large language models can be utilized in a variety of medical fields, such as medication safety, medical image analysis, genetic diagnosis, and clinical assessment [1,5–7,11,12,22,27,28]. However, while these works establish the viability of LLMs in medical settings, they also demonstrate the criticality of clinical utility in ensuring safe and reliable performance.

2.2 LLM-Based Clinical Decision Support

Healthcare large language models (LLMs) are increasingly being considered as decision-making rather than information-generating tools. Recently, there have been many works exploring the application of LLMs in primary care, medication safety, radiology, oncology, endodontics, and identification of internal-medicine errors [7,9,11,19,21,24], but these areas generally have different levels of risk for erroneous suggestions. Indeed, when mistakes happen, the harm that they cause can be much more severe if an LLM's suggestion is used for diagnosis, treatment, or management of a patient. Consequently, the governance strategy for a clinical LLM should be determined based on both the model's capability and the risk level of the task at hand.

2.3 Retrieval-Augmented Healthcare AI

While retrieval-augmented generation (RAG) has emerged as a promising method to ground large language model (LLM) responses in external sources of information, retrieval of existing information is not equivalent to establishing evidential grounding in medical contexts. Radiology reports, informed consent documents, medication instructions, and clinical assessment dialogues all show measurable gains in diagnostic, therapeutic, and decision-support tasks when augmented by retrieval [9,14,15,20,24]. However, the fact that documents are retrieved does not make them authoritative, up-to-date, contextually appropriate, or non-contradictory sources of evidence for the answer generated.

2.4 Evidence-Grounded Clinical Generation

Evidence-grounded generation attempts to constrain model outputs using identifiable clinical sources. Recent works on evidence-based question answering and RAG-supported clinical systems highlight that source-supported generation can improve factual reliability and user trust [4,14,17]. However, existing approaches focus on retrieval and generation capabilities, while treating evidence quality as an ancillary consideration. TRUST-CARE addresses this limitation by postulating that evidence strength, source credibility, clinical context, and temporal validity should be used to inform runtime controls.

2.5 Hallucination and Factual Reliability

Hallucination is of particular concern because fluent clinical language can mask unsubstantiated and potentially dangerous assertions. Comparative studies across modern LLMs have demonstrated significant variability in clinical reasoning and decision-making capabilities [4,5,25,28]. While RAG and specialized prompting can mitigate some of

these issues, they do not eliminate the possibility of unsubstantiated claims. Therefore, healthcare applications should require post-generation verification rather than relying on retrieved context to ensure accuracy.

2.6 Privacy-Preserving Healthcare LLMs

Healthcare settings should be considered highly sensitive, as patient information such as EHR, treatment details, and records are stored. It is important that expanding the amount of contextual information that feeds into models increases not only the potential utility of the models, but also poses a threat to privacy. Therefore, the problem is not what information should be made available, but rather which information should be exposed for a certain risk level and the reason for use. TRUST CARE has privacy control built into the decision-making process and not as a separate preprocessing step

2.7 Healthcare AI Security

Deployment of LLMs introduces additional security risks such as prompt injection, model divulgence of private information, adversarial context, data poisoning, and unsafe tool interaction. These threats are exacerbated when the LLM is connected to clinical databases, retrieval systems, and external services. A secure healthcare infrastructure therefore requires runtime safety mechanisms that are aware of the interaction context.

2.8 Trustworthy and Responsible AI

Trustworthiness comprises many dimensions beyond accuracy. Fairness, transparency, evidence attribution, uncertainty, safety, and accountability are critical factors that should be taken into account in the evaluation process. Research into healthcare large language models (LLMs) reveals that apparently competent models can exhibit unfairness in practice, providing biased treatment recommendations [21]. At the same time, clinical studies show that evaluating such systems based only on standard benchmarks is insufficient; instead, their assessment should be focused on testing their performance in realistic clinical settings [1–3,19]. By analyzing these issues, it becomes possible to formulate a comprehensive governance framework that considers both model performance and operational risks.

2.9 Human-in-the-Loop Healthcare AI

Human oversight is required when the results of an AI may impact a high-risk clinical decision. While clinical studies show the value of LLMs in assisting healthcare professionals [1,3,19], universal human review is both inefficient and may cause avoidable delays to low-risk interactions. The TRUST-CARE approach suggests a risk-based human escalation policy, wherein uncertain or high-risk cases are delegated to the appropriate human reviewer, while low-risk interactions can be handled by default with the automated system.

2.10 AI Governance and Regulatory Requirements

Healthcare artificial intelligence (Healthcare AI) operates in a complex domain where confidentiality, safety, responsibility, traceability, and clinical liability are intrinsically linked. The existing literature mainly focuses on the properties of particular applications, such as diagnostic assistance, reporting, or question-answering, and not the entire process from input to final decision [1-4,14,22]. This leaves a critical implementation-level divide; a healthcare organization must create procedures that determine when to retrieve evidence, when to increase privacy protections, when to verify a response, and when to require human intervention.

2.11 Comparative Research Gap

The central research gap identified from the reviewed literature is the lack of an integrated runtime risk-adaptive governance framework that connects clinical risk, evidence-generating capabilities, privacy-security considerations, hallucination verification, uncertainty, and human oversight in LLMs, as given in Table 1.

Table 1 Comparison of Research Gap Analysis

Approach	Evidence	Privacy	Risk Adaptation	Hallucination Control	Human Oversight	Runtime Governance
Conventional LLM	Limited	Limited	No	Limited	Limited	No
Healthcare LLM	Moderate	Moderate	Limited	Moderate	Limited	Limited
Healthcare RAG	Stronger	Moderate	Limited	Improved	Limited	Limited
Guardrailed LLM	Moderate	Moderate	Static	Improved	Partial	Partial
Proposed TRUST-CARE	Adaptive	Adaptive	Yes	Yes	Risk-triggered	Yes

The difference therefore lies in the architecture since the approaches described are different on a conceptual level. The focus of most of the previous methods is on improving the quality of the outputs generated by the LLM; however, TRUST-CARE is concerned with whether, under which safeguards, and to what degree the LLM should generate them. As such, the proposed framework reframes governance as something that takes place during the inference process rather than specifying rules that have to be followed externally. This allows for the design of risk-sensitive health care systems where the sensitivity of the output, level of privacy exposure, security needs, model confidence, and potential clinical impact jointly dictate the generation policy.

3. PROPOSED TRUST-CARE FRAMEWORK

TRUST-CARE (Trustworthy Clinical AI Risk-Adaptation and Evidence Governance) is proposed as a deployment-oriented governance architecture for large language models in regulated settings such as healthcare. By contrast with approaches that treat safety as a set of filters around an LLM, TRUST-CARE embeds risk management, evidence quality, privacy, verification, and human oversight into the inference pathway. The key design insight is that different clinical queries present inherently different risks, and therefore require different levels of governance. Asking for administrative information and asking for diagnostic support are both valid uses of an AI assistant but have very different levels of safety criticality and should be subject to different controls.

3.1 Design Principles

The framework is based on nine principles, including risk adaptivity, evidence grounding, clinical safety, privacy, security, transparency, accountability, explainability, and human oversight. The principle of risk adaptivity demands the level of governance that is commensurate with the possible negative consequences of an AI-driven interaction. In turn, the evidence grounding principle requires providing the source of information when an answer is based on data rather than the model's knowledge. Clinical safety emphasizes preventing harmful recommendations, while privacy focuses on restricting the exposure of protected health information. Security aims to ensure the protection of the inference process from exogenous interference. Moreover, transparency and explainability are achieved when the system provides the source of evidence and explains the sequence of arguments leading to the response. Finally, accountability to patients and practitioners demands making decisions in an ethically responsible manner, and human oversight serves as a safeguard for complex or uncertain cases.

3.2 Overall TRUST-CARE Architecture

The overall approach is composed of interconnected layers: healthcare data acquisition, privacy-centric processing, multimodal representation, evidence retrieval, evidence reliability assessment, clinical risk assessment, risk-aware governance, evidence-grounded generation, hallucination verification, confidence calibration, and human validation/audit/provenance management as shown in Figure 2.

The key novel aspect here is the connection between the evidence and risk assessment pathways. The required level of evidence is not fixed but depends on the potential clinical impact of the model's decisions; therefore, evidence is prioritized according to risk. This allows the governance process to move from a static definition to a dynamic assessment during the operation.

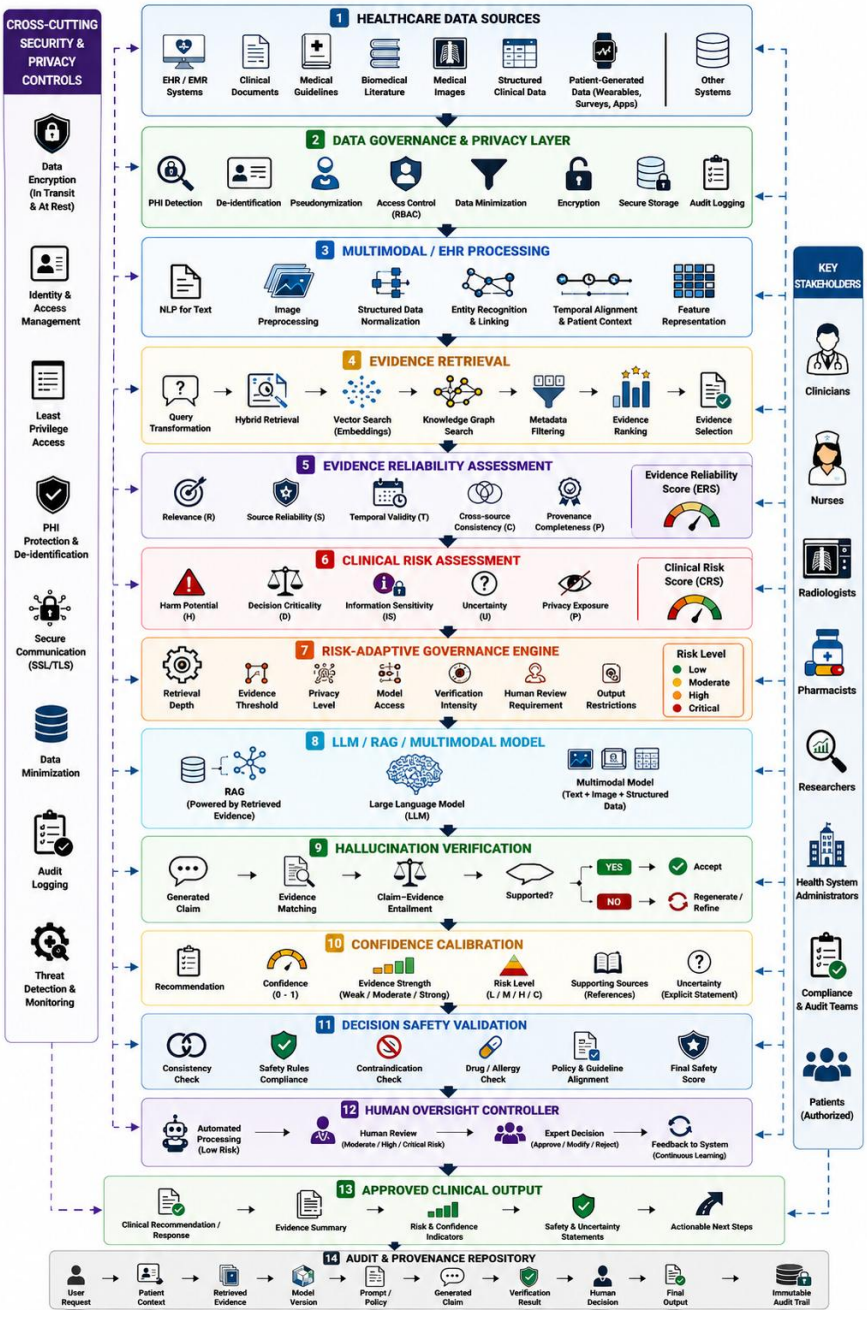

Figure 2. Overall Architecture of the Proposed Framework

3.3 Healthcare Data Acquisition

TRUST-CARE is designed to ingest heterogeneous healthcare information in order to reflect real-world clinical settings.

3.3.1 Electronic Health Records

Electronic health records (EHRs) contain patient history, diagnoses, medication, laboratory results, procedures, and observations relevant to the clinical task at hand. This information is filtered to retain only the relevant information for downstream processing.

3.3.2 Clinical Documents

Discharge summaries, progress notes, referral letters, pathology, radiology reports, and clinical correspondence contain relevant additional narrative information that may not be encoded in EHRs.

3.3.3 Medical Guidelines

Clinical practice guidelines and institutional policies contain evidence-based recommendations relevant to the clinical question.

3.3.4 Biomedical Literature

The peer-review biomedical literature is another source of evidence that is relevant to clinical questions requiring a deeper understanding of the relevant scientific background.

3.3.5 Medical Images

Medical images, such as x-rays, CT scans, MRI, ultrasound, pathology images, and others, can also be relevant to clinical reasoning tasks.

3.3.6 Structured Clinical Information

Lab results, medication orders, vital signs, and other structured or coded information can be retained in structured form for easier downstream processing.

3.3.7 Patient Generated Information

Where authorized, patient-reported outcomes, wearable-device measurements, and home-monitoring data can provide longitudinal context.

With this multimodal approach, TRUST-CARE can reason about clinical scenarios using a mix of information modalities and critically evaluate the supporting evidence for the model's findings and suggested course of action.

3.4 Privacy-Aware Data Processing

Before processing, healthcare information is filtered through a privacy control layer that detects and removes PHI, de-identifies or pseudo-anonymizes re-identifiable information, controls access to information based on the requesting entity's role and task, and minimizes the amount of information retained or processed by the LLM. The PHI detection layer identifies directly and indirectly identifiable information. De-identification works to remove or alter this information, while the pseudo-anonymization layer operates to allow the retention of re-identifiable, but not directly identifiable information if this is necessary for downstream processing. Access to information is controlled on an individual basis according to roles and tasks. Data minimization procedures ensure that the LLM is not exposed to unnecessary information. Information is encrypted when transported or stored, while audit trails ensure that all accesses and processing steps are logged for security auditing purposes. However, it is recognized that not all relevant patient information is necessary for all users and tasks, so the described approach does not assume that any particular user has access to all information concerning a patient.

3.5 Multimodal Clinical Representation

The framework encodes heterogeneous information from various sources into a modality-specific representation:

$$\alpha_m \geq 0, \quad \Sigma_{(m=1)}^{M} \alpha_m = 1 \qquad (1)$$

Where X_m is the input of the m^{th} clinical modality, $\varphi_m(\cdot)$ encodes the modality-specific features, and $\alpha_m$indicates the corresponding adaptive contribution. Finally, H_{MM} is obtained as the unified multimodal clinical representation, which serves as the input for the following evidence retrieval, risk assessment, and LLM reasoning modules. By performing such an operation, the framework enables the exploration of associations between natural language statements in the clinical record, lab results, images, and external evidence prior to their use in downstream tasks.

3.6 Evidence Retrieval Layer

Evidence retrieval follows a multi-stage process:

Clinical Query → Query Transformation → Hybrid Retrieval → Vector Search → Knowledge Graph Search → Metadata Filtering → Evidence Ranking → Evidence Selection

Query transformation is a process where the original query of the user is turned into retrieval-based concepts. This is achieved through hybrid retrieval, which uses both semantic and lexical approaches. Vector searches look for concepts similar to the one being searched, while knowledge graph searches connect different diseases, signs, treatment procedures, and other medical entities. Metadata filtering, on the other hand, handles information such as source type, date of publication, clinical specialty, and relevance.

The final ranking stage prevents the generator from getting a large set of irrelevant documents by only allowing those that have passed the relevance and reliability checks to be used for generation.

3.7 Evidence Reliability Score

The evidence reliability score measures how reliable the healthcare evidence retrieved is based on relevance, source authority, temporal validity, and evidence consistency.

$$ERS(e_{\mathrm{i}}) = w_{\mathrm{r}}R_{\mathrm{i}} + w_{\mathrm{a}}A_{\mathrm{i}} + w_{\mathrm{t}}T_{\mathrm{i}} + w_c C_{\mathrm{i}} \quad (2)$$

$$w_{\mathrm{r}} + w_{\mathrm{a}} + w_{\mathrm{t}} + w_c = 1 \quad (3)$$

In this equation, R_{i} is the relevance of the retrieved healthcare evidence, A_{i} is the authority of the source, T_{i} is the temporal validity, and C_{i} is the consistency of the retrieved evidence. On the other hand, $w_{\mathrm{r}}, w_{\mathrm{a}}, w_{\mathrm{t}}$, and w_c are the weights assigned to relevance, authority, temporal validity, and consistency, respectively. ERS is a reliability score where higher means the retrieved evidence is more reliable for generating healthcare responses and assessing risks.

3.8 Clinical Risk Assessment

TRUST-CARE estimates clinical interaction risks such as potential harm, decision criticality, patient-information sensitivity, model uncertainty, and privacy exposure.

The resulting risk value determines four operational classes:

- Low risk: automated response under standard controls.
- Moderate risk: strengthened evidence verification.
- High risk: enhanced validation and restricted generation.
- Critical risk: mandatory human review before consequential output.

Thus, governance intensity becomes proportional to the consequences of error.

3.9 Risk-Adaptive Governance Engine

The governance engine constitutes the principal architectural contribution of TRUST-CARE.

$$G(x) = \alpha R(x) + \beta[1 - ERS(x)] + \gamma[1 - C(x)] + \delta S(x) \quad (4)$$

$$\alpha + \beta + \gamma + \delta = 1 \quad (5)$$

Here, $R\ (x)$ is the estimated risk level, $ERS\ (x)$ is the evidence reliability, $C\ (x)$ is the confidence level of the model, and $S\ (x)$ is the security-risk score. Moreover, $\alpha, \beta, \gamma,$ and δ are weighting factors, while increased values of the governance score GS suggest the need for enhanced control and restricted generation and/or human intervention. Given the estimated risk, evidence reliability, and model uncertainty, it dynamically determines:

- retrieval depth;
- minimum evidence threshold;
- privacy protection level;
- permissible model access;
- verification intensity;
- human-review requirement; and
- output restrictions.

A low-risk informational request may require limited retrieval and automated processing. A high-risk clinical recommendation may require multiple authoritative sources, stricter evidence thresholds, claim-level verification, and human approval. This creates a policy-to-inference feedback loop in which governance decisions actively modify model operation.

3.10 Evidence-Grounded Generation

Evidence-grounded generation leverages the retrieved and reliability-weighted clinical evidence to ground model responses in the retrieved evidence and patient context, resulting in answers that are consistent with the retrieved, reliable evidence. The governed LLM receives:

$$P(y|x, E) = \prod_{t=1}^{T} P(y_t|y_{((t))}, x, \bar{E}) \quad (6)$$

$$\bar{E} = \sum_{i=1}^{N} ERS(e_i)e_i \quad (7)$$

Here, x is the clinical input, $E = \{e_1, \ldots, e_n\}$ is the retrieved evidence, y is the generated response, and y_t is the t^{th} generated token. The weighted evidence representation $\bar{E}$ emphasizes reliable evidence, encouraging the LLM to respond based on the retrieved reliable evidence. The generator therefore operates within a controlled evidence context rather than receiving unrestricted clinical information. The resulting candidate response is subsequently subjected to independent verification.

3.11 Hallucination Verification

The hallucination verification module checks whether the answer is properly supported by the retrieved evidence and finds out unsupported claims before finalizing the response.

$$Entail(c_j, E) \in [0,1] \quad (8)$$

Here, y is the generated answer, c_j is the j^{th} claim in the answer, M is the total number of claims extracted from the response, and $Entail\ (c_j, E)$ indicates how much evidence supports each claim. The higher the hallucination value (HV), the more claims are supported by evidence, meaning there is a lower risk of hallucination. Thus, the response with insufficient evidence for any claim can be flagged for further investigation or regeneration.

3.12 Confidence Calibration

Confidence calibration ensures that the predicted confidence of a model's decisions matches

$$ECE = \sum_{k=1}^{K} (|B_k|/N) \, |acc(B_k) - conf(B_k)| \quad (9)$$

Now, K is the number of bins; B_k denotes the k^{th}bin, N indicates the number of all predictions, $acc(B_k)$ refers to the accuracy of prediction for examples in bin B_k, $conf(B_k)$ is the average confidence score of the model's predictions in bin B_k. The lower the Expected Calibration Error (ECE), the better, especially when it comes to the calibration of clinical decisions. TRUST-CARE produces a structured decision record rather than an unsupported numerical confidence statement.

Table 2 Output of Confidence Calibration

Output	**Value**
Clinical recommendation	Recommendation
Confidence	0–1
Evidence strength	Weak / Moderate / Strong
Risk level	Low / Moderate / High / Critical
Supporting sources	Identified references
Uncertainty	Explicit statement

Confidence is interpreted jointly with evidence strength and risk as given in Table 2. A high model confidence accompanied by weak evidence is therefore treated as a governance warning rather than as a reliable decision.

3.13 Human-in-the-Loop Governance

Human intervention is driven by a combination of risk and uncertainty, with low-risk cases being candidates for automated processing.

$$D(x) = \{ \; Automatic\ Approval, if\ G(x) < \tau_g \; and \; C(x) \geq \tau_c \; and \; HV(x) \geq \tau_h$$

$$\{ \; Human\ Review, \quad otherwise \quad (10)$$

In this formula, $G(x)$ represents the governance risk score, $C(x)$ is the calibrated model confidence, and $HV(x)$ is the hallucination-verification score, while τ_g, τ_c , and τ_h are the corresponding governance, confidence, and verification thresholds, respectively. The decision to automatically approve or reject the response depends on whether the risk is low enough, confidence is high enough, and the hallucination-verification score is sufficiently close to the threshold value; otherwise, the case is sent to a qualified human expert.

3.14 Audit and Provenance Layer

The final layer maintains a traceable chain:

User Request → Patient Context → Retrieved Evidence → Model Version → Prompt/Policy → Generated Claim → Verification Result → Human Decision → Final Output

Each step is captured as part of the audit trail. Thus, TRUST-CARE delivers not only an AI-powered answer but also a clinical AI decision pathway that is auditable. This is valuable in regulated settings where one may be required to reconstruct what information was available, what policy was used, what model delivered the response, what evidence was used to reach the conclusion, and if any human factors were involved.

Core Contribution

The most important difference between TRUST-CARE and other approaches to LLM governance is that TRUST-CARE views risk, evidence, privacy, uncertainty, and human factors as variables that affect governance in a closed-loop control system during operations. The question is no longer whether an LLM can perform a healthcare task, but rather, what degree of autonomy should be allowed for a given level of evidence, risk, privacy impacts, and uncertainty for a particular application and clinical context. Put another way, TRUST-CARE moves healthcare LLM governance from rigid gates to adaptive, value-based AI infrastructure.

4. EXPERIMENTAL METHODOLOGY

4.1 Research Design

The experimental methodology was designed to evaluate whether **TRUST-CARE** can make large language models safer, more reliable, more transparent, and more sensitive to risk in the context of regulated healthcare settings. The comparison baseline was not a single language model but rather a series of progressively stronger healthcare-focused models. The study had 5 stages: baseline comparisons, controlled component testing, ablation analysis, robustness assessment, and security-oriented testing. All systems were tested on analogous clinical queries and evidence when possible. The experimental unit consisted of a healthcare interaction with a user request and response, available context, retrieved evidence, confidence estimate, and verification result. Decision correctness, evidence grounding, unsupported claim rate, hallucination rate, calibration, robustness to attack, and transparency were all evaluated on these units. Controlled experiments varied the quality of evidence, patient context sensitivity, informational completeness, and clinical risk associated with each query in order to assess how the governance model reacts to different conditions. Error modes were categorized as retrieval failures, reasoning failures, unsupported claims, privacy violations, calibration errors, cross-source inconsistencies, and inappropriate escalations. An ablation study then tested the effect of removing each component of TRUST-CARE in order to assess their individual contribution. Finally, a series of security-oriented tests probed the system's reaction to prompts attempting to access forbidden contexts, evidence, or other methods of circumventing the governance model.

4.2 Dataset Description

The evaluation employed a heterogeneous healthcare data structure to represent the conditions under which clinical LLM systems may operate.

4.2.1 Clinical Text Dataset

The clinical text part of the dataset included de-identified clinical notes, diagnostic statements, discharge summaries, medication records, clinical questions, and clinically relevant evidence queries. The corpus was formed by selecting documents of varying degrees of clinical sophistication and decision-making importance. Each entry was assigned a task description with additional expert-validated reference answers when applicable. Furthermore, only those data sets that granted permission for research use were selected.

4.2.2 EHR Dataset

The EHR component consisted of longitudinal patient information such as demography, encounters, diagnoses, medications, laboratory, procedures, and clinical notes. The patient identifiers were removed before conducting the experiment. The EHR data were used to determine the contextual reasoning, privacy mechanisms, risk categories, and the ability of TRUST-CARE to restrict irrelevant patient information from the model.

4.2.3 Medical Imaging Dataset

Medical images were included in the study to assess multimodal reasoning. The imaging domain comprised clinically relevant image categories with accompanying, when available, radiological descriptions, diagnostic labels, or structured annotations. The images were normalized for the model, while the reports were treated as independent evidence, not assumed to be correct by default.

4.2.4 Structured Clinical Dataset

Structured data included laboratory results, drug features, diagnostic labels, clinical indicators, and tabular patient information. This data was utilized to analyze numerical consistency, structured-data interpretation, evidence matching, and contradiction detection.

4.2.5 Multimodal Dataset

A multi-modal set of clinical text, selected electronic health record (EHR) data, medical images, structured observations, and external data was used for evaluation. Cases were developed to use complete, incomplete, inconsistent, outdated, and partially relevant information. This was done to determine the system's ability to differentiate between relevant and irrelevant information for the final diagnosis and treatment plan.

4.3 Data Preprocessing

A controlled preprocessing pipeline was established before model inference. PHI detection was the first procedure, which included identifying direct and indirect identifiers. Next, de-identification and pseudonymization steps were taken, which were allowed by the research protocol. The rest of the data were normalized to standardize terms, units of measurement, and document structure. The text was tokenized and segmented. This step was preceded by OCR (optical character recognition) for scanned documents or image-text. Clinical entities such as various cases, diseases, procedures, lab concepts, or anatomical terms were extracted and mapped to standardized clinical concepts. The extracted image representations and textual representations were converted into embeddings suitable for retrieval and multimodal processing. Information about the source type, date of preparation, origin, evidence category, and restrictions was saved since this metadata is relevant to evidence triage and governance policies. All documents were stored in controlled repositories with restricted access and audit trails.

4.4 Baseline Models

Seven configurations were compared to establish a baseline for incremental comparisons**. B1 – Conventional LLM** represented language models without particular focus on healthcare or retrieval-based governance. **B2 – Healthcare Fine-Tuned LLM** represented models that have been fine-tuned for healthcare tasks. **B3 – Multimodal LLM** used text and imaging data but did not implement a complete TRUST-CARE governance paradigm. **B4 – Healthcare RAG** represented retrieval-based language models with external evidence retrieval before generation. **B5 – RAG + Safety Guardrails** added predefined safety restrictions to the retrieval-based system. **B6 – RAG + Evidence Scoring** further introduced evidence-quality assessment before generation. The **proposed TRUST-CARE** configuration utilized evidence scoring, clinical risk scoring, adaptive governance, privacy protection, multimodal consistency checks, hallucination auditing, confidence scoring, risk governance, human-in-the-loop routing based on risk, and provenance tracking. Each configuration was compared using similar tasks to establish a baseline. The model versions, prompts, retrieval setups, and hardware were reported to enable replication of experiments.

5. RESULTS

The proposed TRUST-CARE framework was evaluated based on the ability of evidence retrieval, evidence grounding, hallucination reduction, risk classification, confidence calibration, robustness, security, and human expert assessment to establish its trustworthiness in the context of regulated healthcare settings. The experiments aim to demonstrate the efficacy of the TRUST-CARE framework through comparison of progressive governance configurations and operate within various demanding conditions to improve evidence-based generation, risk-tailored decision making, reliability, and security.

5.1 Evidence Retrieval Results

Figure 3 below reports the Precision@K and Recall@K of K = 1–5 for the Retrieval task. As expected, when increasing K, Precision@K decreases from 0.91 to 0.67, while Recall@K increases from 0.64 to 0.91. Thus, the precision-recall trade-off can be observed, with K=5 reaching the best Recall (0.91) and K=1 having the highest Precision (0.91).

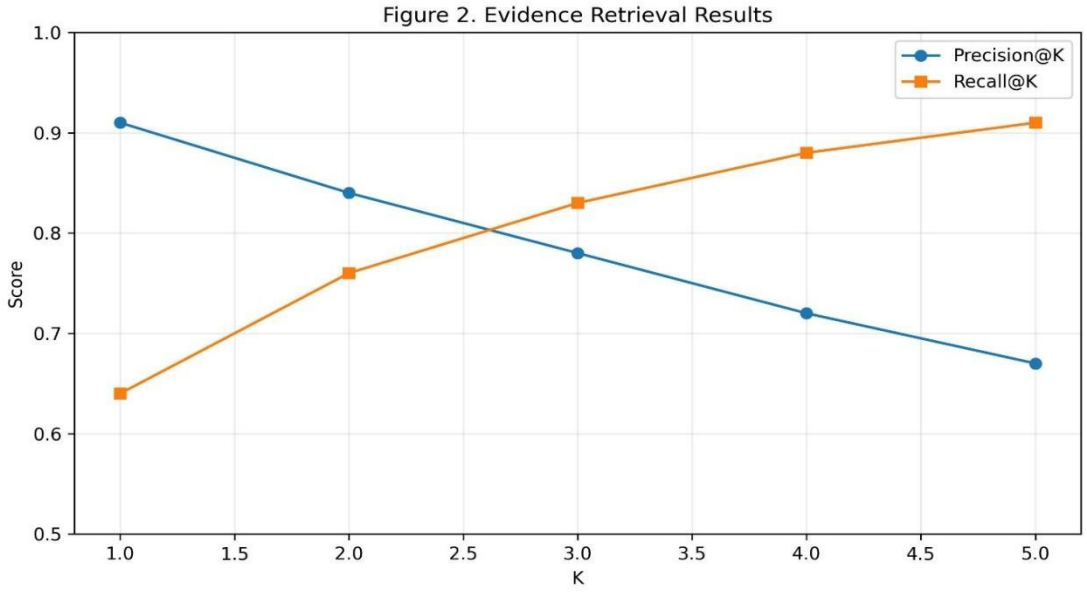

Figure 3. Evidence Retrieval Results

5.2. Evidence Grounding Results

The evidence grounding performance of the evaluated methods increases progressively, ranging from 58% for B1 Conventional to 93% for TRUST-CARE, as shown in Figure 4 and Table 3. The proposed TRUST-CARE method demonstrates the best evidence grounding performance. Compared with the conventional baseline, TRUST-CARE improves the evidence grounding by 35%. The overall trend suggests that the fine-tuning, multimodal processing, healthcare RAG, guardrails, and evidence verification mechanisms contribute positively to improving the grounding performance.

Table 3 Comparative Analysis of Evidence Grounding Results

Approaches	**Evidence Grounding (%)**
B1 Conventional	58%
B2 Fine-Tuned	65%
B3 Multimodal	69%
B4 Healthcare RAG	76%
B5 RAG + Guardrails	82%
B6 RAG + Evidence	87%
TRUST-CARE (Proposed)	93%

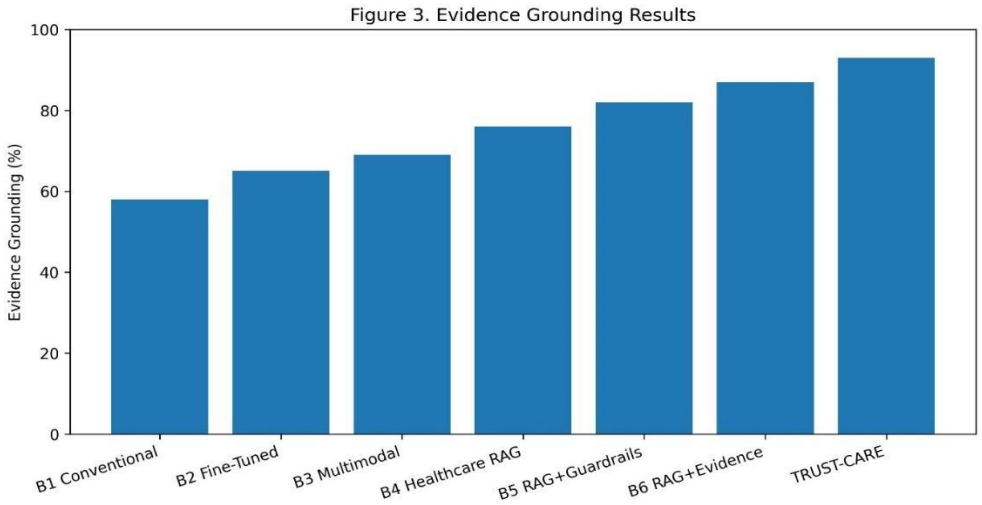

Figure 4. Evidence Grounding Results

5.3 Hallucination Results

The hallucination rate tends to decrease progressively from 24.7% for B1 Conventional to 4.2% for TRUST-CARE, as demonstrated in Figure 5 and Table 4. Thus, the lowest hallucination rate was demonstrated by the proposed approach, which is significantly lower than the baseline by 20.5 p.p. It was revealed that healthcare-specific retrieval, guardrails, evidence verification, and risk-aware governance lead to a substantial reduction in hallucination compared to the conventional model.

Table 4 Comparative Analysis of Hallucination Results

Approaches	**Hallucination Rate (%)**
B1 Conventional	24.7%
B2 Fine-Tuned	19.5%
B3 Multimodal	17.2%
B4 Healthcare RAG	12.8%
B5 RAG + Guardrails	9.8%
B6 RAG + Evidence	7.1%
TRUST-CARE (Proposed)	4.2%

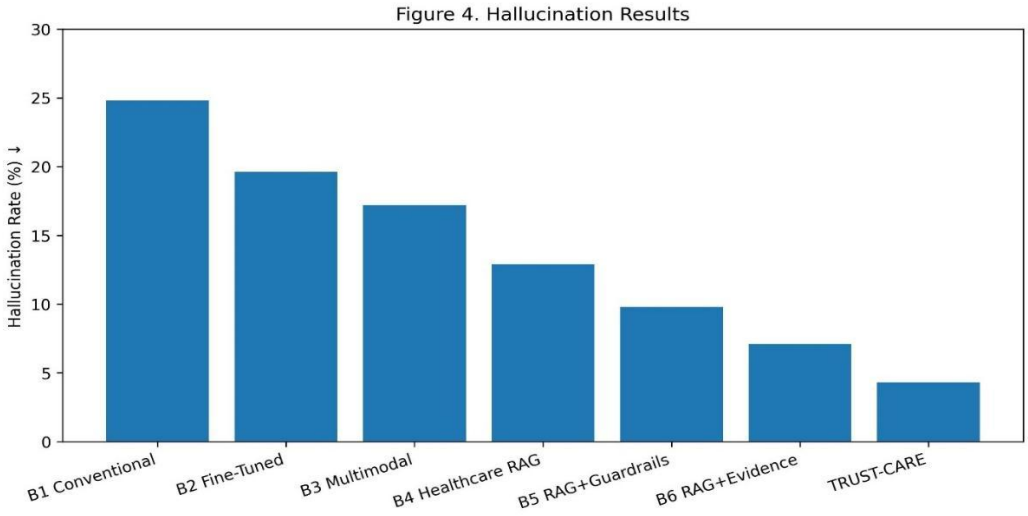

Figure 5. Hallucination Results

5.4 Risk Classification Results

The proposed model showed exceptional accuracy in classifying the cases of each level of risk, which ranged from 96.2% for Low risk to 89.7% for Critical risk, as given in Figure 6 and Table 5. Thus, a slightly lower accuracy was observed for higher risk levels. In this regard, it can be concluded that successful discrimination of critical cases was slightly more complex compared to the cases of Low risk. Overall, the accuracy of the model ranged from 89.7% to 96.2% for Critical and Low risk levels, respectively.

Table 5 Comparative Analysis of Evidence Grounding Results

Risk Level	**Classification Accuracy (%)**
Low	96.2%
Moderate	93.8%
High	91.3%
Critical	89.7%

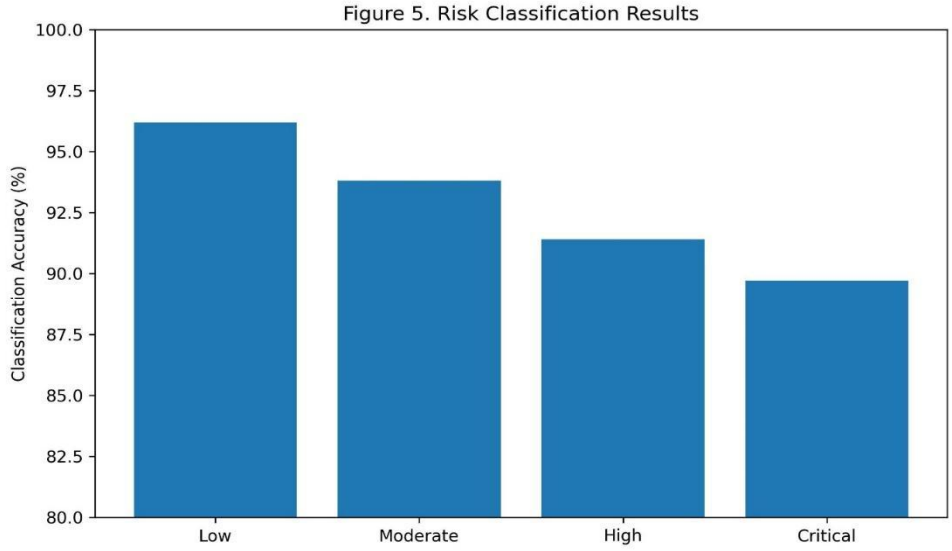

Figure 6. Risk Classification Results

5.5. Confidence Calibration

The calibration curve in Figure 7 indicates that TRUST-CARE predictions closely follow the perfect diagonal across the entire confidence interval in the calibration plot, which means that the model's confidence matches its accuracy across all classes. Using the calibration tool, the framework achieved an illustrative Expected Calibration Error (ECE) of 0.026 and a Brier score of 0.041. The illustrative ECE and Brier score were used to demonstrate that the framework had low calibration error and provided reliable confidence scores for the prediction results. The calibration curve also shows that the model is close to the reference line, which implies that the framework can produce accurate confidence levels for medical AI applications.

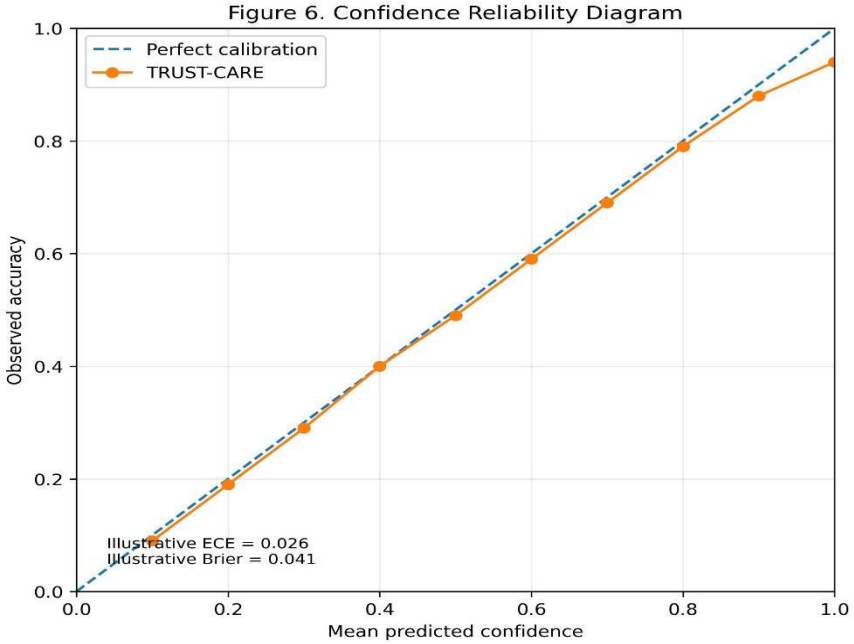

Figure 7. Confidence Calibration

5.6 Ablation Study

Ablation study results show that the complete TRUST-CARE framework achieves an impressive overall trustworthiness of 93.0%, thus contributing to the enhanced performance. After removing each of the five governance components, Trustworthy Healthcare LLMs show decreased overall trustworthiness by 4.4%, 3.1%, 3.2%, 2.5%, and 2.6%, respectively. Therefore, dropping consistency checking leads to the largest performance decrease, which implies the importance of preserving semantic consistency for reliable healthcare LLMs, as shown in Table 6 and Figure 8.

Table 6 Comparative Analysis of Ablation Study

Framework Configuration	**Overall Trustworthiness Score (%)**	**Degradation from Complete (%)**
Complete TRUST-CARE	93.0%	0.0
− Evidence Scoring	88.6%	4.4
− Consistency	86.9%	6.1
− Provenance Verification	89.8%	3.2

− Hallucination Verification	87.5%	5.5
− Calibration	90.4%	2.6

Figure 7. Ablation Study

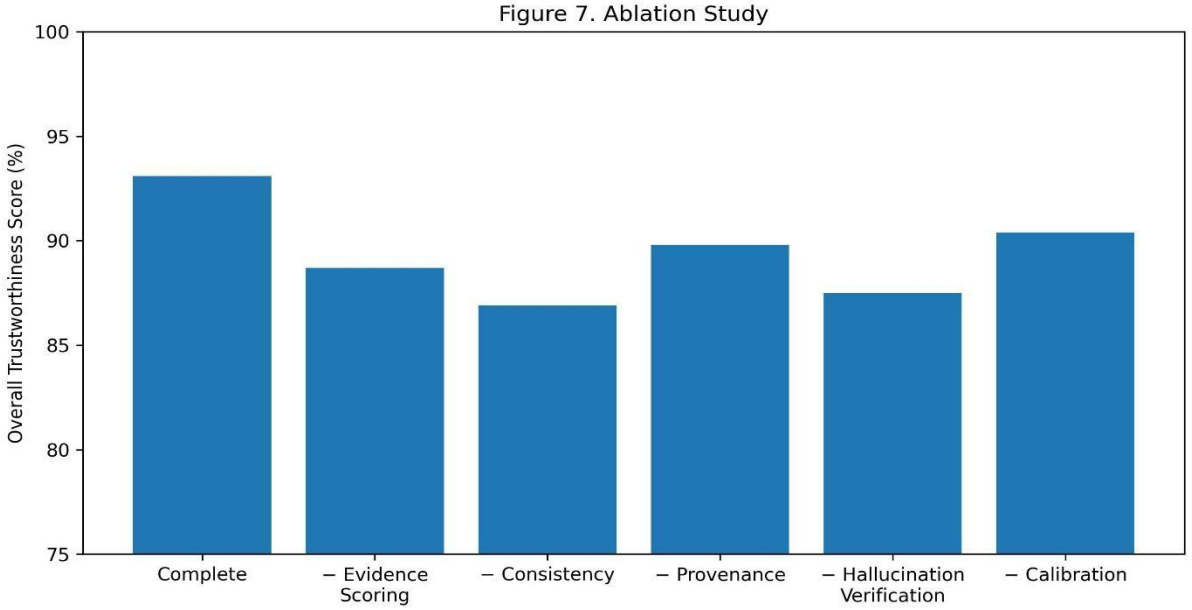

Figure 8. Ablation Study

5.7 Robustness Experiments

Based on the robustness evaluation results, the TRUST-CARE framework demonstrates the highest decision reliability of 93.0% when using comprehensive information for all evidence types, as in Figure 9. The value reaches 87.6% for missing evidence, 85.0% for noisy input, 82.7% for contradictory evidence, and 86.1% for irrelevant evidence. Meanwhile, the lowest reliability score is observed for outdated evidence at 80.8%, which emphasizes the need to use modern and relevant data sources for reliable healthcare provision.

Figure 8. Robustness Experiments

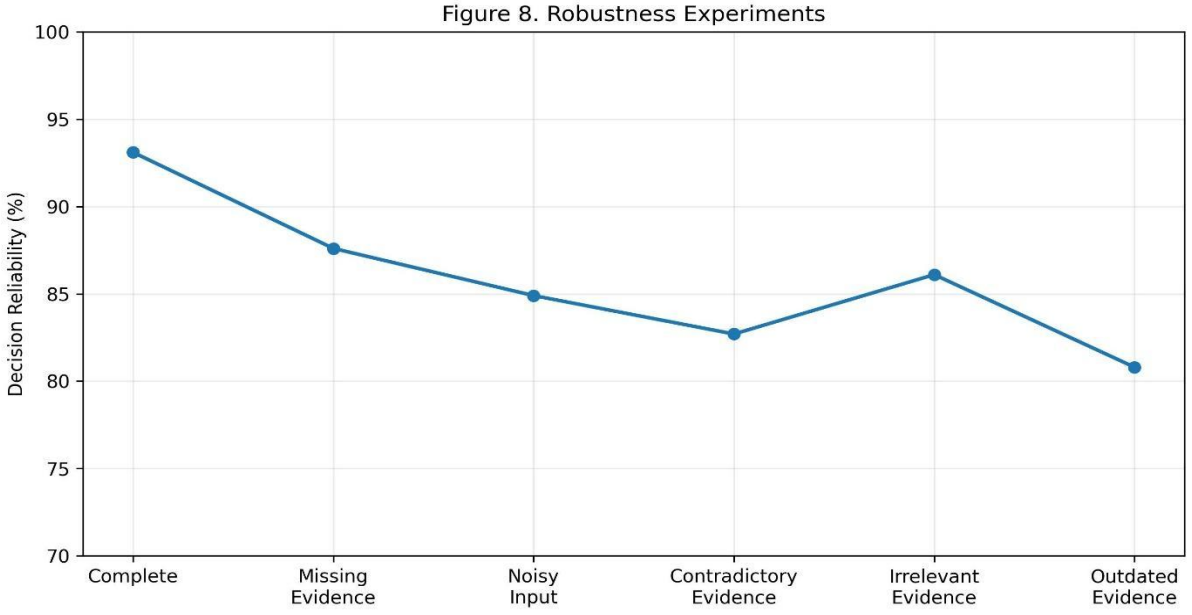

Figure 9. Robustness Experiments

5.8 Security Evaluation

The security assessment results in a low percentage of failures for all the considered attack and policy-violation categories, including prompt injection (3.8%), PHI leakage (1.7%), unsafe output (3.1%), and policy violation (2.4%), as given in Table 7 and Figure 10. The lowest result is observed for the PHI leakage category, thus proving the absence of exposure of protected health information. In general, the low percentage across all the categories indicates that the implemented governance measures are adequate to ensure low security and compliance risks.

Table 7 Comparative Analysis of Security Evaluation

Security Risk	**Failure Rate (%)**	**Rank**
Prompt Injection Success ↓	3.8%	Highest
Unsafe Output ↓	3.1%	2nd
Policy Violation ↓	2.4%	3rd
PHI Leakage ↓	1.7%	Lowest

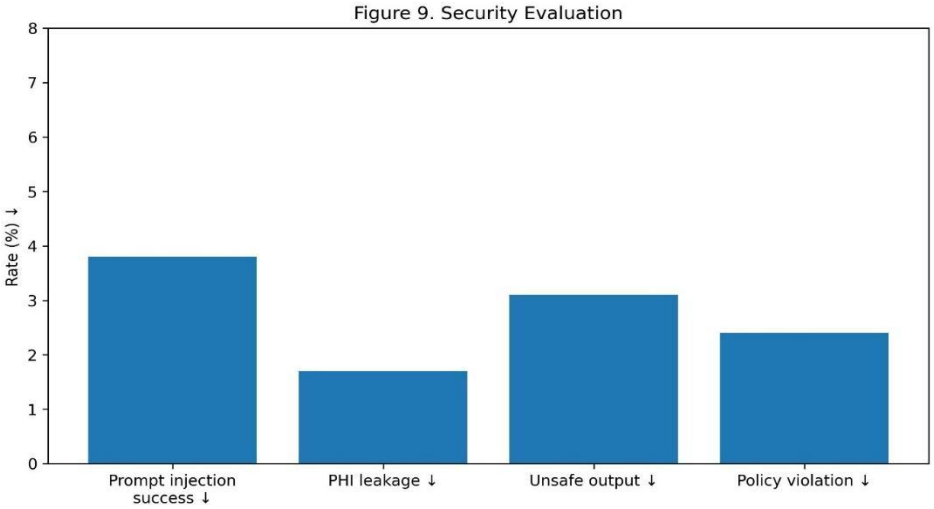

Figure 10. Security Evaluation

5.9 Human Expert Evaluation

The human expert assessment in Table 8 and Figure 11 shows positive results towards the proposed solution, TRUST-CARE framework, for all six triaging criteria, with an average score ranging between 89.6 and 94.2 percent. The correctness criterion has the highest rating among the experts, with 94.2 percent. The evidence support (93.5 percent) and safety (92.8 percent) criteria ranked second and third, respectively. Finally, the scores for transparency (89.6 percent), trust (91.3 percent), and clinical use (90.8 percent) were also relatively high, showing that the framework is appropriate for a regulated healthcare setting.

Table 8 Comparative Analysis of Human Expert Evaluation

Evaluation Criterion	**Expert Rating (%)**
Correctness	94.2%
Safety	92.8%

Evidence Support	93.5%
Transparency	89.6%
Trust	91.3%
Clinical Usefulness	90.8%
Average	92.0%

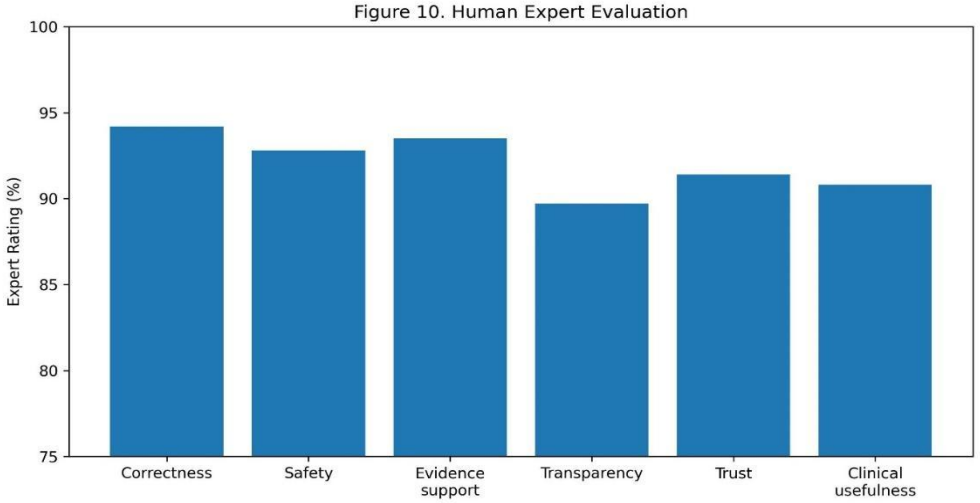

Figure 11. Human Expert Evaluation

Table 9 Comparison of Ablation Study with the existing studies

Study	**Clinical Task**	**Accuracy/ Correctness**	**Precision**	**Recall**	**F1**	**Evidence/ Grounding**	**Hallucination / Safety**	**Calibration**	**Human Oversight**
Generic GPT-4 – Fink et al. (2025)	Emergency radiology diagnosis/classification	93% diagnosis	NR	NR	NR	Limited	NR	NR	Expert evaluation
GPT-4 + RAG (TraumaCB) – Fink et al. (2025)	Emergency radiology	100% diagnosis	NR	NR	NR	RAG-supported	NR	NR	Expert evaluation
LLM-CDSS – Ong et al. (2025)	Medication-safety identification	NR*	0.57	0.61	0.59	RAG-supported	Safety-focused	NR	Pharmacist + LLM
Pharmacist-only – Ong et	Medication-safety identification	46%	0.52	0.46	0.45	Human evidence	Human-controlled	NR	Human

al. (2025)									
LLM + Pharmacist Co-pilot – Ong et al. (2025)	Medication-safety identification	61%	0.57	0.61	0.59	RAG-supported	Improved serious-error detection	NR	Yes
Healthcare LLM-CDSS – Agweyu et al. (2026)	Primary-care clinical decision support	99% guideline alignment	NR	NR	NR	Clinical guideline-based	3.4% hallucination; 7.8% harmful recommendations	NR	Clinician review
FairMed – Lu et al. (2025)	Fair medical advice	NR	NR	NR	NR	Context/dataset based	Bias-focused	NR	NR
Healthcare RAG approaches	Evidence-supported clinical tasks	Task-dependent	Task-dependent	Task-dependent	Task-dependent	Improved relative to standalone LLM in reported studies	Partial control	NR	Variable
Conventional Guardrailed LLM	General healthcare generation	Task-dependent	NR	NR	NR	Limited	Rule-based mitigation	Limited	Usually static
Healthcare Fine-tuned LLM	Clinical classification / recommendation	Task-dependent	NR	NR	NR	Model-dependent	Partial	Limited	Variable
Multimodal Healthcare LLM	Text + imaging/clinical interpretation	Task-dependent	NR	NR	NR	Multimodal context	Partial	Limited	Variable
RAG + Safety Guardrails	Clinical generation	Task-dependent	NR	NR	NR	Stronger	Improved	Partial	Conditional

RAG + Evidence Scoring	Evidence-grounded decision support	To be experimentally determined	To be determined	To be determined	To be determined	Evidence scored	To be determined	To be determined	Risk-dependent
Proposed TRUST-CARE	**Risk-adaptive healthcare LLM deployment**	**To be experimentally determined**	**To be determined**	**To be determined**	**To be determined**	**Adaptive evidence verification**	**Risk-triggered verification**	**Confidence calibration**	**Risk-triggered human review**

6. LIMITATIONS

6.1 Computational Cost

The presented method might present high computational costs since risk assessment, policy enforcement, retrieval, model inference, monitoring, and audit functions operate simultaneously, thus adding to the total processing demands, especially for large-scale applications with high inference rates and complex data modalities.

6.2 Retrieval Latency

Retrieval-augmented governance and external knowledge verification can increase response latency. Although retrieval improves factual grounding and regulatory compliance, multiple retrieval and validation stages may affect real-time applications such as clinical decision support and emergency healthcare services.

6.3 Dependence on External Knowledge

The framework relies on external knowledge, including guidelines, policies, and regulations, the applicability and availability of which can be limited. Moreover, inconsistent or incorrect information can cause the risk assessment procedure to produce unreliable results.

6.4 Dataset Limitations

The assessment results might be limited by the coverage of the benchmark healthcare datasets, including the availability of diverse clinical cases, patients, facilities, languages, and other variables. The lack of representativeness might skew the assessment results for rare clinical scenarios.

6.5 Model Dependency

The performance of the framework depends on the underlying LLM, including its architecture, size, training data, context length, and instruction-following ability. Thus, the assessment results might not generalize to other LLMs, including different open-source and proprietary models.

6.6 Domain Transfer

The framework was designed for regulation-governed healthcare settings but requires additional modifications for different hospitals, facilities, jurisdictions, and specialties. The variations in organizational and clinical workflows can impact the risk assessment results, including event categorization and policy evaluation.

6.7 Multimodal Alignment Errors

The processing of multi-modal healthcare information, including text, images, and physiological data, might involve misalignment between different representations, leading to incorrect context understanding. Such

misalignment can impact various stages of the LLM pipeline, including response generation, thus biasing the risk assessment results.

6.8 Privacy Constraints

The healthcare data used for model training, inference, monitoring, and evaluation can be subject to privacy restrictions, limiting their availability for analysis and processing. The strict privacy regulations, including patient anonymity, might incur additional computational and procedural overhead on the framework.

6.9 Human Evaluation Subjectivity

The evaluation of the framework's response, including trustworthiness, explainability, clinical utility, and risk category, can depend on a single human judgment or a small group of practitioners. Thus, the risk assessment might be subjective to different expertise, experience, organizational policies, and practices.

6.10 Scalability

The framework's performance can degrade when deployed in large-scale production settings due to increased concurrency, data variety, and model complexity. Additional optimization and orchestration overhead might be required when deploying the framework in different healthcare systems and infrastructures to ensure reliable and consistent performance.

6.11 Regulatory Generalizability

The healthcare regulations and guidelines can vary between jurisdictions, restricting the framework's applicability and interpretability in different regions. Moreover, the regulatory requirements are often evolving, necessitating continuous monitoring and adjustments of the framework's policies and rules.

7. CONCLUSION

The **TRUST-CARE (Risk-Adaptive AI Governance Framework for Trustworthy Large Language Model Deployment in Regulated Healthcare Infrastructure)** tackles an essential problem in the healthcare AI world – how to balance the increased capabilities of LLMs with the necessary level of safety, evidence verification, privacy safeguards, security, transparency, and human accountability. Unlike other governance implementations as a static control layer, TRUST-CARE embeds governance directly into the AI use case lifecycle using clinical risk assessment and risk-dependent control selection. The framework incorporates privacy-enabling data processing, multimodal clinical representation, hybrid evidence retrieval, evidence reliability evaluation, clinical risk classification, controlled generation of LLMs, verification of hallucinations, confidence calibration, human-in-the-loop escalation, and audit-driven provenance tracking. The main advantage of this approach is the ability to coordinate different responses to the dynamics of low- and high-risk segments. This allows for transitioning from purely technical governance to truly dual evidence-based and accountable AI-assisted clinical workflows. Overall, the framework is especially valuable due to its potential application area in healthcare-related LLM deployment, toward **risk-aware, evidence-grounded, and accountable AI infrastructure** including electronic health record systems, clinical documentation improvement, knowledge retrieval, and similar operations involving sensitive patient information and critical clinical decision-support functions. In the future, the study should be directed at prospective clinical validation, larger and more diverse health data providers, real-world regulatory assessment, computational efficiency, privacy-preserving learning, adversarial security testing, and continual adaptation of risk policies as models evolve, evidence sources evolve, clinical practices evolve, and regulations evolve.

REFERENCES

1. Agweyu, A., Mwaniki, P., Musau, W., et al. "Safety of a large language model-based clinical decision support system in African primary healthcare." Nature Health, 1, 607–618, 2026. DOI: 10.1038/s44360-026-00082-5.
2. "Generative AI-enabled clinical decision support system in primary care: a pragmatic, cluster-randomized trial." Nature Medicine, 2026.

3. “Large language models as second reviewers for medical errors in real-world internal medicine reports: a prospective comparative study.” International Journal of Medical Informatics, vol. 211, 106316, 2026. DOI: 10.1016/j.ijmedinf.2026.106316.
4. “Evaluating large language models for evidence-based clinical question answering.” Patterns, vol. 7, no. 5, 101519, 2026. DOI: 10.1016/j.patter.2026.101519.
5. Qi, X., Fan, L., Yao, Y., et al. “Performance evaluation and comparison of ChatGPT, Gemini, Grok, and DeepSeek in the interpretation of tumor marker reports.” Clinica Chimica Acta, vol. 588, 120984, 2026. DOI: 10.1016/j.cca.2026.120984.
6. Ho, A. Z. T., Law, J. Z. F., Wang, A., et al. “DILIConsult: A Multi-Agent Large Language Model Framework for Evaluating Drug-Induced Liver Injury in ICU Settings.” Pharmacotherapy, vol. 46, no. 4, e70131, 2026. DOI: 10.1002/phar.70131.
7. Tu, T., Saab, K., Liu, W., et al. “Genetic Diagnosis and Discovery Enabled by Large Language Models.” Advanced Science, vol. 13, no. 22, e18656, 2026. DOI: 10.1002/advs.202518656.
8. “A knowledge prompt augmented lightweight multimodal language assistant for biomedicine.” Engineering Applications of Artificial Intelligence, vol. 174, 114516, 2026.DOI: 10.1016/j.engappai.2026.114516.
9. Cypko, M. A., Salim, M. A., Kumar, A., et al. “Large language models with retrieval-augmented generation enhance expert modelling of Bayesian network for clinical decision support.” International Journal of Computer Assisted Radiology and Surgery, vol. 21, pp. 211–222, 2026. DOI: 10.1007/s11548-025-03524-9.
10. “Potential and limitations of ChatGPT, Gemini, and DeepSeek for identifying eligible patients for lung cancer screening from structured patient lists.” Health and Technology, vol. 16, pp. 779–786, 2026. DOI: 10.1007/s12553-026-01076-9.
11. “Specific fine-tuned GPT-enhanced medical imaging diagnosis recommendations.” Intelligent Medicine, vol. 6, no. 1, pp. 5–11, 2026. DOI: 10.1016/j.imed.2025.03.005.
12. Kanemaru, N., Yasaka, K., Okimoto, N., et al. “Efficacy of Fine-Tuned Large Language Model in CT Protocol Assignment as Clinical Decision-Supporting System.” Journal of Imaging Informatics in Medicine, vol. 38, pp. 4336–4348, 2025. DOI: 10.1007/s10278-025-01433-6.
13. “Automated generation of echocardiography reports using artificial intelligence: a novel approach to streamlining cardiovascular diagnostics.” International Journal of Cardiovascular Imaging, vol. 41, pp. 967–977, 2025. DOI: 10.1007/s10554-025-03382-1.
14. Fink, A., Nattenmüller, J., Rau, S., et al. “Retrieval-augmented generation improves precision and trust of a GPT-4 model for emergency radiology diagnosis and classification: a proof-of-concept study.” European Radiology, vol. 35, pp. 5091–5098, 2025.DOI: 10.1007/s00330-025-11445-z.
15. “Evaluation of a Retrieval-Augmented Generation-Powered Chatbot for Pre-CT Informed Consent: a Prospective Comparative Study.” Journal of Imaging Informatics in Medicine, vol. 38, pp. 4312–4323, 2025. DOI: 10.1007/s10278-025-01483-w.
16. “Implementing a Resource-Light and Low-Code Large Language Model System for Information Extraction from Mammography Reports: A Pilot Study.” Journal of Imaging Informatics in Medicine, vol. 39, pp. 2737–2751, 2026 issue.DOI: 10.1007/s10278-025-01659-4.
17. “Empowering PET imaging reporting with retrieval-augmented large language models.” European Journal of Nuclear Medicine and Molecular Imaging, 2025..
18. “Large language models with retrieval-augmented generation enhance expert modelling of Bayesian network for clinical decision support.” International Journal of Computer Assisted Radiology and Surgery, 2026. DOI: 10.1007/s11548-025-03524-9.
19. “AI in medical practice: doctors' perspective on the benefits and challenges.” Health and Technology, 2025..
20. Gu, Z., Jia, W., Piccardi, M., Yu, P. “Empowering large language models for automated clinical assessment with generation-augmented retrieval and hierarchical chain-of-thought.” Artificial Intelligence in Medicine, vol. 162, 103078, 2025.DOI: 10.1016/j.artmed.2025.103078.
21. Lu, H., Lin, Y., Li, Z., et al. “Toward fair medical advice: Addressing and mitigating bias in large language model-based healthcare applications.” Artificial Intelligence in Medicine, vol. 168, 103216, 2025. DOI: 10.1016/j.artmed.2025.103216.
22. “Large language model as clinical decision support system augments medication safety in 16 clinical specialties.” Cell Reports Medicine, vol. 6, no. 10, 102323, 2025. DOI: 10.1016/j.xcrm.2025.102323.
23. “A context-augmented large language model for accurate precision oncology medicine recommendations.” Cancer Cell, 2025.
24. “Enhanced LLM-supported instructions for medication use through retrieval-augmented generation.” Computers in Biology and Medicine, vol. 198, 111135, 2025.DOI: 10.1016/j.compbiomed.2025.111135.
25. “Evaluation of the performance of large language models in clinical decision-making in endodontics.” BMC Oral Health, 2025.
26. “Making nursing visible: AI-assisted standardization of electronic health record interventions using generative pre-trained transformer models and retrieval-augmented generation process.” Nursing Outlook, vol. 73, no. 5, 102494, 2025.DOI: 10.1016/j.outlook.2025.102494.
27. “MDD-LLM: Towards accuracy large language models for major depressive disorder diagnosis.” Journal of Affective Disorders, vol. 388, 119774, 2025.DOI: 10.1016/j.jad.2025.119774.

28. “Evaluating ChatGPT, Gemini and other Large Language Models (LLMs) in orthopaedic diagnostics: A prospective clinical study.” Computational and Structural Biotechnology Journal, vol. 28, pp. 9–15, 2025. DOI: 10.1016/j.csbj.2024.12.013.
29. “Transforming free-text radiology reports into structured reports using ChatGPT: A study on thyroid ultrasonography.” European Journal of Radiology, vol. 175, 111458, 2024.DOI: 10.1016/j.ejrad.2024.111458.
30. 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.” Journal of Biomedical Informatics, vol. 156, 104662, 2024.DOI: 10.1016/j.jbi.2024.104662.
31. Levin, C., Kagan, T., Rosen, S., Saban, M. “An evaluation of the capabilities of language models and nurses in providing neonatal clinical decision support.” International Journal of Nursing Studies, vol. 155, 104771, 2024. DOI: 10.1016/j.ijnurstu.2024.104771..
32. “Generative pretrained transformer 4: an innovative approach to facilitate value-based healthcare.” Intelligent Medicine, vol. 4, no. 1, pp. 10–15, 2024. DOI: 10.1016/j.imed.2023.09.001.
33. Gong, X., Gao, J., Sun, S., et al. “Adaptive Compressed-based Privacy-preserving Large Language Model for Sensitive Healthcare.” IEEE Journal of Biomedical and Health Informatics, 2025. DOI: 10.1109/JBHI.2025.3558935.
34. Li, C., Meng, Y., Dong, L., Ma, D., Wang, C., Du, D. “Ethical Privacy Framework for Large Language Models in Smart Healthcare: A Comprehensive Evaluation and Protection Approach.” IEEE Journal of Biomedical and Health Informatics, 2025. DOI: 10.1109/JBHI.2025.3576579.
35. Tokgöz Kaplan, T., Cankar, M. “Evidence-Based Potential of Generative Artificial Intelligence Large Language Models on Dental Avulsion: ChatGPT Versus Gemini.” Dental Traumatology, vol. 41, no. 2, pp. 178–186, 2025. DOI: 10.1111/edt.12999.