
https://iphopen.org/index.php/se
Online ISSN: 3050-8797 Print ISSN: 3050-9270

# DESIGNING TRUSTWORTHY AI IN HEALTHCARE: EXPERIENCES WITH COPILOT AGENTS, AGENTIC MODELS, AND RAG INTEGRATION

**Venkata Babu Mogili***
*Independent Researcher, USA

***Corresponding Author:** **Venkata Babu Mogili**

## ABSTRACT

Healthcare systems need to reduce administrative burden and support decision-making for clinical practice. Artificial intelligence approaches have the potential to reduce documentation and support diagnosis. Copilot Agents are in-app assistants that enable users to ask questions, automate documentation tasks, and coordinate clinical work processes without interrupting their current tasks within electronic health record systems. Agentic AI is not limited to single-turn questions and responses but also includes goal-aware reasoning during multi-turn tasks. Examples of such tasks span from processing prior authorizations to transitioning care calls and quality measurement documentation. Further, the clinical review at several checkpoints in the architecture is important to the implementation. For the RAG to be factually correct, language model outputs are grounded in validated institutional knowledge bases and clinically accepted guidelines. Source attribution mechanisms enable clinicians to trace model outputs to their respective information sources or references. Critical to the architecture of the RAG are security, privacy, and interpretability constraints in medical practices. Governance frameworks created by ongoing monitoring, responding to incidents, and involving stakeholders are essential for successfully using AI solutions in a way that supports rather than replaces clinical decision-making.

**Keywords:** Retrieval-Augmented Generation, Clinical Decision Support, Electronic Health Records, Healthcare Artificial Intelligence, Agentic Systems, Knowledge Grounding

DOI:-10.5281/zenodo.18851338 Manu script # 415

## I. Introduction

In healthcare, a meaningful amount of unstructured data is created in the backend, requiring interpretation, coding, and documentation. This causes clinicians to spend excessive time on data and not enough time with patients. Just as well-being-driven strategic models in holistic industries emphasize long-term relationship building and user-centered value creation, healthcare AI systems must prioritize clinician well-being, trust, and sustained engagement to ensure successful technological adoption [13]. Chart review, documentation, and care coordination are time-consuming tasks. Customary rule-based automation attends to a narrow set of tasks, lacking flexibility in complex, contextualized clinical scenarios.

Artificial intelligence has many proposed uses in medicine, including pattern recognition in medical images, mining the huge amount of unstructured data existing in diverse data sources, and identifying clinically meaningful relationships in electronic health records [1]. Machine learning algorithms can be used to assist in diagnosis, prediction and recommendation, creating a powerful man-and-machine partnership in high-performance medicine. Realizing this potential requires addressing these implementation challenges.
In a medical setting, using machine learning comes with limitations because doctors' decisions can greatly affect patients, and there are rules about how data can be used and how systems must be checked; still, the doctor is always responsible for the patient, no matter what calculations are done. Ethical challenges include transparency about the algorithmic limitations and bias [2], machine learning challenges such as fairness and accountability, and proper limits to automation.

Modern AI architectures provide a way to offer adaptive support. Large language models display natural language understanding and generation that can be considered similar to other modalities for communication about healthcare. Deploying in the clinical setting raises accuracy issues. Technical hurdles include compliance, integration into existing systems, and the gulf between experimental AI capabilities and actual clinical tools that have the requisite production quality. Systems should be approached in a systematic and holistic manner.
This paper outlines a practical framework for examining three complementary AI models: Copilot Agents, embedded directly into clinical software. Agentic AI performs multi-step reasoning under human supervision, while Retrieval-Augmented Generation uses verified knowledge sources to augment large language models. The analyses of architectural design choices, governance structures, and lessons learned from the deployments illustrate the operational requirements for each model. This combination provides a coherent support architecture for the healthcare domain.

## II. Related Work

Most prior work on healthcare AI has focused on isolated clinical automation tasks in specific domains and uses single-modality rather than multimodal implementations. This article presents a unified architectural framework consisting of Copilot Agents, Agentic AI, and Retrieval-Augmented Generation. Though common in electronic health record systems, few implementation studies exist in the academic literature. Additionally, this literature presents functional architectures for a Copilot Agent built to preserve clinical workflows. Despite the emphasis on agentic AI in the literature, there has been insufficient attention given to the governance of oversight in healthcare. A checkpoint structure for human oversight of consequential clinical decisions is an important application. Even though a lot of work has been done on Retrieval-Augmented Generation (RAG) for tasks that need a lot of knowledge in natural language processing, there is still a need for more research focused on healthcare situations that require specific knowledge bases and clinical guidelines, even with the framework providing source attributions Security and privacy are not added on after the fact but baked into the design of the architecture. Interpretability needs inform design choices from the outset. The framework connects what AI can theoretically do with the limits it faces in regulated healthcare, allowing for gradual use and learning within institutions.

## III. Copilot Agents in Clinical Environments

### A. Functional Architecture and Design Principles

Copilot Agents are embedded generative AI digital assistants that integrate within existing software, with electronic health record (EHR) systems being the primary example. They assist clinicians in answering clinical questions, surfacing relevant clinical knowledge and patient data, and documenting care. The architectural pattern designates the agent as an augmentation of human capabilities, where the human retains control rather than the agent.

Machine learning is a class of algorithms for building medical systems that learn from experience. Classification and regression are the workhorses of predictive models [3]. Supervised learning methods are the most common in clinical practice and use labeled examples to make predictions, with performance dependent on how well they

are trained. Feature selection selects and determines which variables influence predicted results. Copilot implementations use these techniques for information retrieval and ranking.
In this model, when the agent is aware of the clinical task at hand, it provides actionable information to the user without being prompted (e.g., surfacing historical information, reviewing medication lists and lab trends during the review of a patient encounter). Natural language processing decodes clinical notes. Information extraction identifies clinical entities from unstructured text.

Neural networks can approximate higher-level concepts or patterns. Deep learning networks have been applied to images, text, and other high-dimensional subjects. Convolutional networks excel at visual pattern recognition. Recurrent architectures are aimed at sequential data. For natural language tasks, such as those seen in Copilot Agents, transformer models are frequently employed. "These models can generate contextually appropriate responses to clinical queries.

**B. Workflow Preservation and Integration Considerations**
A clinical workflow is another important design principle. Healthcare professionals pursue efficient workflows in these practice environments. Systems that require behavioral changes face barriers. Human -factor research has identified critical factors for AI adoption [4]. Trust in recommendations impacts use, while perceived usefulness drives continued use, and ease of use stimulates initial adoption.

Clinicians use an AI-enabled decision support program differently from a customary software program due to different contexts and impacts on the decision, such as time pressure and cognitive load in clinical settings. On the negative side, overload may affect performance [4]. Good Copilot design minimizes cognitive workload.
The agent interface is often presented as a sidebar or an overlay for clinicians to use as needed. So the idea is to preserve workflow and to assist without taking the focus away from this complicated clinical task. Progressive disclosure offers information when needed. Default behaviors assume user intents in common situations.
Integration with existing data sources commonly requires compliance with interoperability standards. For example, standards exist for health information exchange. Application programming interfaces aid machines in communicating, and authentication mechanisms secure sensitive information. Role-based access controls (RBAC) manage visibility to data, while audit logging tracks system activities to meet compliance.

| Component | Function | Clinical Application |
|---|---|---|
| Contextual Awareness Module | Monitors current clinical task and patient context | Surfaces relevant historical information during encounter review |
| Natural Language Processor | Interprets clinical notes and extracts key concepts | Identifies medications, diagnoses, and laboratory findings |
| Information Retrieval Engine | Searches patient records for pertinent data | Retrieves medication lists, laboratory trends, and prior assessments |
| Documentation Assistant | Generates and refines clinical documentation | Automates progress notes and discharge summaries |
| Interface Overlay | Presents assistance within existing software | Sidebar or overlay component for selective clinician engagement |
| Feature Selection Module | Determines relevant variables for predictions | Identifies clinical factors informing decision support |

**Table 1.** Copilot Agent Functional Components and Clinical Integration Characteristics [3, 4].

**IV. Agentic AI for Multi-Step Healthcar j e Processes**
**A. Goal-Oriented Reasoning and Task Decomposition**
Agentic AI pipelines need to go beyond multi-turn dialogues to multi-step tasks that require planning, execution, and error recovery, naturally fitting into healthcare workflows. An example is prior authorization processing, having multiple decisions (e.g., treatment, inpatient admission, authorization) across multiple systems for care transition coordination. Quality measure documentation requires evidence retrieval and synthesis.

Retrieval-Augmented Generation (RAG) improves a language model's ability to handle knowledge-heavy tasks by blending learned information with information that is retrieved from other sources. Agentic systems also use retrieval mechanisms in both planning and acting. Task decomposition makes use of procedural information, and execution depends on obtained guidance.

An agentic model is characterized by breaking down high-level goals into subtasks and actions whose execution proceeds based on intermediate outcomes. Such a model allows reacting to variability in healthcare processes.

These clinical practice guidelines include exceptions, as patient factors can result in variations in clinical management.

Knowledge-intensive applications require grounding in domain-specific corpora. For instance, healthcare agents rely on clinical guidelines, formulary data, and institutional policies. Retrieval-based mechanisms extract relevant passages from document collections, subsequently guiding actions and fine-tuning parameters [5]. Factual grounding reduces the likelihood of hallucinations.
Planning algorithms sequence actions. The goal specification specifies an end state the agent desires. State space search is used to find action sequences achieving particular goals. Constraint satisfaction handles operational constraints. Resource allocation improves efficiency. Error recovery allows graceful execution failures.

**B. Human Oversight Mechanisms and Accountability Structures**
In healthcare, quality control is high. Human oversight of important decisions is essential. Agentic systems have checkpoint architectures, whereby automated processes pause at points of decision-making. A clinical review is done before reaching the checkpoint to ensure efficiency, accountability, and patient safety.
AI decision support systems need rules to ensure responsibility and to confirm their accuracy in healthcare before they are used, along with ongoing checks to see if their performance declines. The notes provide an important form of feedback to clinicians. Incident reporting is a form of continuous quality improvement. Regulatory requirements are associated with an audit trail.

Checkpoints align with clinical workflow, and critical decisions trigger human review. Normal operations still occur with interruptions at intervals as low as the threshold, with only a minor effect. Checkpoint order is determined by risk. Intensity of supervision depends on the severity of the patient and the complexity of the procedure [6]. Adaptive checkpoint policies balance performance and safety.

Audit logging allows for the complete execution trace of operations to be reviewed for quality assurance and records for auditing purposes. Similar to real-time payment monitoring systems in public finance that enhance transparency and traceability through continuous transaction logging, healthcare agentic AI systems benefit from real-time execution monitoring and revenue-like workflow tracking mechanisms to strengthen accountability and compliance [12]. Timestamp recording creates a temporal relationship, and user identifiers associate actions with their actors. System state snapshots allow capturing decision-making contexts.

To help clinicians understand system behavior, transparency mechanisms are used, such as explanation generation. Confidence indicators signal uncertainty, alternative actions allow choice, transparency of limitations sets expectations, and trust calibration aligns reliance with true performance and automated behavior.

| Process Element | Description | Healthcare Application |
|---|---|---|
| Goal Specification | Defines desired end states for automated processes | Prior authorization approval, care transition completion |
| Task Decomposition | Breaks high-level objectives into actionable subtasks | Sequential steps for quality measure documentation |
| Execution Monitoring | Tracks outcomes during multi-step process execution | Identifies failures requiring alternative action paths |
| Checkpoint Architecture | Pauses automation at defined decision points | Mandatory clinical review before consequential actions |
| Audit Logging | Captures complete execution traces with timestamps | Supports regulatory compliance and quality assurance |
| Adaptive Execution | Adjusts subsequent actions based on intermediate results | Accommodates patient-specific circumstances |

**Table 2.** Agentic AI Process Characteristics and Oversight Mechanisms are detailed in references [5, 6].

**V. RAG Integration for Output Reliability**
**A. Knowledge Grounding Architecture and Retrieval Mechanisms**
Retrieval-Augmented Generation augments generative models with a retrieval mechanism, constraining the model's outputs to relevant documents retrieved from verified knowledge sources. Knowledge bases created from institutional sources are authoritative, and clinical guidelines provide evidence-based standards of care and approved reference material to the organization.

Retrieval-augmented generation (RAG) architectures have shown strong performance on knowledge-intensive tasks by integrating external knowledge bases into a single generation process [7]. Then, the query formulation extracts the main concepts of the user query, and the document retrieval extracts relevant passages. Passage ranking ranks the most informative passages, while context integration integrates the retrieved content into the prompt.

The querying component retrieves documents that match input queries, often using vector similarity search. Dense retrieval models embed queries and documents into a common space and use nearest neighbor search. An additional re-ranking model may improve on the initial results [7]. The top-ranked passages are used as context. Healthcare applications need domain-specific retrieval corpora. Terminology in health is domain-specific. Medical ontologies organize relationships between concepts. Synonym expansion improves recall. Abbreviation handling deals with clinical shorthand. Negation detection prevents misinterpretations. Temporal reasoning finds application within patient timelines.

During generation, the retrieved context is condensed into an answer and processed with retrieval results and a query. In faithful generation, the model focuses its attention on the salient content, while in abstractive summarization and format adaptation, models are required to synthesize and transform the generated content.

**B. Source Attribution and Knowledge Base Maintenance**

Certain RAG implementations additionally offer source attribution, enabling users to discern the sources utilized. This allows users to trace back statements to the source documents, enabling clinical verification. Evidence-based practice patterns are computationally implemented.

Large language models can hold medical information, as shown by tests where they answer medical questions using facts from their training. Parametric knowledge is trained on cutoff dates, and, for rare conditions, it may lack media coverage. Institutional coverage is not well represented, and RAG architectures tackle this issue using dynamic retrieval.

Attribution mechanisms guide the models. They specify which part of the sources was used (e.g., span-level citation) or simply indicate which of the sources is appropriate. Confidence scoring indicates certainty [8]; verification interfaces allow inspection of the source. Correction reduces attribution.

The retrieval corpus and clinical guidelines are periodically updated, and institutional policies are modified as organizations change and new evidence is published. Outdated documentation may become reflected in system outputs. Document lifecycle management. Version control tracks changes, while expiration policies remove outdated entries.

Quality assurance procedures verify corpus integrity either through content review, completeness assessment (for missing coverage), or consistency checking (for contradictions). Authority checks if the source has the appropriate credentials. Currency checks if the source is timely. Governance oversight maintains the curation quality.

| Component | Function | Output Reliability Contribution |
|---|---|---|
| Query Formulation | Extracts key concepts from user input | Ensures retrieval targets relevant knowledge domains |
| Vector Similarity Search | Encodes queries and documents in shared embedding spaces | Enables semantic matching beyond keyword overlap |
| Passage Ranking | Prioritizes most informative retrieved content | Surfaces authoritative sources for generation context |
| Knowledge Base Curation | Maintains currency with evolving clinical guidelines | Prevents outdated materials from contaminating outputs |
| Source Attribution | Links generated statements to specific source documents | Enables clinical validation and evidence tracing |
| Faithful Generation | Constrains outputs to adhere to retrieved source content | Reduces hallucination risks in clinical responses |

**Table 3.** Retrieval-Augmented Generation Architecture Components [7, 8].

## VI. Architectural and Compliance Considerations

### A. Security, Privacy, and Interpretability Requirements

Deployments of healthcare AI models are subject to regulations on protected health information, and system architectures implement access control mechanisms. Data is encrypted at rest and in transit. Auditable capabilities can be used to comply with regulations. Privacy-preserving techniques reduce disclosure risks.

Interpretable machine learning is intriguing because clinical decision support systems should be able to explain their recommendations [9]. Black boxes can be difficult to adopt; clinicians must understand the algorithm's reasoning. Regulation is pushing for explanation, and patients have a right to know how they were treated.
Architectural best practices recommend data minimization so that a single component only has access to and uses the data that it needs to perform its task and the AI component only receives the required data. Basic controls, like retention limits and purpose limitations, can prevent the aggregation of data.

Access control mechanisms enforce authorization policies. Role-based access is aligned with clinical duties. Attribute-based controls enable fine-grained restrictions, context-based decisions, and emergency access provisions in times of need. Audit logging provides an overview of activity patterns.
Encryption protects sensitive data throughout processing; transport layer security protects network communications. Storage encryption protects persistent data. Key management secures cryptographic material. Hardware security modules protect against tampering. Cryptographic agility updates algorithms.

### B. Governance Frameworks and Ethical Considerations

Governance structures need to be established for the model lifecycle and to monitor model performance drift. Incident response procedures handle failures. Clinical and technical participants collaborate. Policies define system function. Acceptable use sets parameters. Escalation procedures address concerns. Governance in healthcare AI can also draw lessons from sustainable financing and risk management frameworks used in global property markets, where long-term accountability, structured oversight, and stakeholder transparency are critical for system sustainability and trust [11].

Healthcare AI governance has drawn inspiration from topics in other high-stakes fields, such as military applications, where accountability, fairness, traceability, dependability, and governability are essential [10]. In healthcare, accountability and fairness are common areas of focus. Established frameworks have allowed for adaptation, while stakeholders ensure various perspectives in governance.

Monitoring detects changes in performance, while SPC is suitable for detecting changes in distribution. Outcome monitoring validates predictions and collects user feedback [10]. Alert systems notify the right people, and investigations identify root causes.

Health professionals can provide feedback to the system via reporting forms. Submissions are prioritized in the triage process. Resolution tracking for closure. Channels to reporters for communicating results. Feedback is incorporated iteratively.

Model lifecycle management includes versioning, validation before deployment, and comparing the behavior of parallel models. Rollback reduces failures. Deprecation policies signal which versions are out of date, while documentation holds institutional knowledge. Training prepares users in anticipation of changes.

| Framework Element | Implementation Requirement | Compliance Contribution |
|---|---|---|
| Access Control Mechanisms | Role-based and attribute-based permission systems | Ensures appropriate data visibility per clinical responsibility |
| Encryption Protocols | Transport layer security and storage encryption | Protects sensitive information at rest and in transit |
| Audit Capabilities | Comprehensive logging of system interactions | Satisfies regulatory documentation requirements |
| Interpretability Features | Explanation generation and confidence indicators | Supports clinician understanding of algorithmic reasoning |
| Performance Monitoring | Statistical process control and outcome tracking | Identifies degradation requiring intervention |
| Incident Response Procedures | Investigation protocols and resolution tracking | Addresses failures systematically with stakeholder communication |

**Table 4.** Security, Privacy, and Governance Framework Elements [9, 10].

**Conclusion**

Copilot Agents, Agentic AI, and RAG can help healthcare organizations automate smartly, addressing different operational issues and supporting a full clinical assistance system. Copilots need to deliver value in the moment by providing contextualized assistance directly to parts of the clinical workflow (e.g., embedded in EHRs) without requiring behavior changes on the part of clinicians. Agentic capabilities solve the uncertainty triggered by complex workflows that require multiple decisions and interactions with other systems. Checkpoints preserve human agency in consequential clinical decisions. Audit trails provide accountability and traceability for the automated execution sequence. Recall-Augmented Generation is an approach to language model reliability. Verified knowledge bases connect generated results to trustworthy information used by the organization, and citation methods connect the results to reliable sources to support evidence-based practice. The knowledge base is updated with new guidelines from time to time and institutional policy through curation processes. To protect patient privacy, security architectures and access controls are proposed. To achieve interpretability, post-hoc explanations and confidence scores are proposed, and, to ensure governance, responsibility for medical decisions is clearly defined to ensure accountability. This includes performance monitoring and incident management processes, enabling organizations to show the advantages of implementing these systems gradually and building familiarity. AI's role in health care is to augment clinical capability, not to substitute for clinicians. cognition. Systems informed by human-centered design principles help ensure that clinicians trust reliable and transparent systems, which persist over time.

**References**

[1] Eric J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, 2019. [Online]. Available: https://www.sbra.be/sites/default/files/1._topol_j._e._high-performance_medicine_the_convergence_of_human_and_artificial_intelligence_nature_medicine_volume_25_january_2019.pdf

[2] Danton S. Char et al., "Identifying Ethical Considerations for Machine Learning Healthcare Applications," American Journal of Bioethics, 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7737650/pdf/nihms-1648280.pdf

[3] Jenni A. M. Sidey-Gibbons and Chris J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," BMC Medical Research Methodology, 2019. [Online]. Available: https://link.springer.com/content/pdf/10.1186/s12874-019-0681-4.pdf

[4] Michael Knop et al., "Human Factors and Technological Characteristics Influencing the Interaction of Medical Professionals With Artificial Intelligence–Enabled Clinical Decision Support Systems: Literature Review," JMIR Hum Factors, 2022. [Online]. Available: https://humanfactors.jmir.org/2022/1/e28639

[5] Aparna Vinayan Kozhipuram et al. conducted a study titled "Retrieval-Augmented Generation vs. Baseline LLMs: A Multi-Metric Evaluation for Knowledge-Intensive Content," published by MDPI in 2025. [Online]. Available: https://www.mdpi.com/2078-2489/16/9/766

[6] Ariowachukwu Divine Sopruchi and Anguzu Rashid, "The Integration of AI-Driven Decision Support Systems in Healthcare: Enhancements, Challenges, and Future Directions," IDOSR JOURNAL OF COMPUTER AND APPLIED SCIENCES, 2024. [Online]. Available: https://www.researchgate.net/profile/Rashid-Anguzu/publication/385379245

[7] YUANJIE LYU et al., "CRUD-RAG: A Comprehensive Chinese Benchmark for Retrieval-Augmented Generation of Large Language Models," ACM Transactions on Information Systems, 2025. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3701228

[8] Karan Singhal et al., "Large Language Models Encode Clinical Knowledge," arXiv, 2022. [Online]. Available: https://arxiv.org/pdf/2212.13138

[9] Muhammad Aurangzeb Ahmad et al., "Interpretable Machine Learning in Healthcare," ACM, 2018. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3233547.3233667

[10] David Oniani et al., "Adopting and expanding ethical principles for generative artificial intelligence from military to healthcare," Nature, 2023. [Online]. Available: https://www.nature.com/articles/s41746-023-00965-x.pdf

[11]Mintah, P. A. (2022). Sustainable Construction Financing Models In The Global Property Market. *Journal of International Crisis and Risk Communication Research* , 71–81. https://doi.org/10.63278/jicrcr.vi.3423

[12]Darteh, F. K. (2022). Real-time payment systems as a tool for improving government revenue monitoring. *Journal of Computational Analysis and Applications (JoCAAA)*, *30*(2), 590–603

[13]Guarin, A. Y. L. (2023). Well-being-driven branding strategies: Connecting holistic fitness and long-term customer relationships. *Sarcouncil Journal of Economics and Business Management*, 2(3), 11–18.