

Leveraging Retrieval-Augmented Generation and Agentic AI for Patient Report Intelligence

Supreet Singh Chawla
ASET, Amity University
Uttar Pradesh, Noida, India
sschawla0104@gmail.com

Nitul Chandra Dutta
ASET, Amity University
Uttar Pradesh, Noida, India
nituldutta345@gmail.com

Sumit Kumar
ASET, Amity University
Uttar Pradesh, Noida, India
sumitkumarbsr19@gmail.com

Surbhi Vijh
ASET, Amity University
Uttar Pradesh, Noida, India
surbhivijh428@gmail.com

Twinkle Tiwari
ASET, Amity University
Uttar Pradesh, Noida, India
twinkletiwari009@gmail.com

Abstract—The emerging unstructured digital health data has increased exponentially to introduce a sense of urgency towards autonomous systems that can provide a high level of accuracy and rigor in clinical interpretation. Nevertheless, the conventional Large Language Models (LLMs) and fixed Retrieval-Augmented Generation (RAG) pipelines cannot be trusted to work well in high-stakes medical situations because of hallucinations, responsibility, and multi-step reasoning. This paper is aimed at eliminating these constraints by introducing a new Hybrid Agentic RAG architecture, Patient Report Intelligence (PRI), which restructures the passive report processing into an active clinical investigation. In contrast to linear architectures, PRI employs a cyclic, graph-based orchestration engine (LangGraph) to coordinate a team of specialized autonomous agents—such as a Strategic Planner, a Medication Analyser, and an Adversarial Critic—thereby mimicking System 2 human clinical reasoning. It further incorporates a unique hybrid retrieval strategy in which a dynamically steering mechanism selects between local clinical guidelines and real-time web evidence, combined with strict citation enforcement and a rigorous Safety Chain.

Keywords—Agentic AI, Retrieval-Augmented Generation, Clinical Decision Support, Medical Report Analysis, Explainable AI.

I. INTRODUCTION

The healthcare sector is now experiencing the shift in paradigm of the purely predictive artificial intelligence to the Third Wave of AI: Agentic Systems. Although the traditional models of Generative AI (GenAI) have been shown to exhibit skills in summarizing medical text, they are essentially one-dimensional processors, not being able to reason through complex and multi-step clinical scenarios by themselves. Two different approaches have been developed to solve these shortcomings: Retrieval-Augmented Generation (RAG) which grounds model outputs on outside knowledge and Agentic AI which adds goal oriented autonomy. While RAG ensures factual correctness by retrieving verified clinical guidelines, it is sometimes unable to make decisions based on longitudinal changes or illuminate uncertain diagnosis information. As a result, the combination of the two paradigms, Hybrid

Agentic RAG, is a new architectural form of designing strong autonomous medical intelligent systems [8].

Unlike standard Large Language Models (LLMs), which are useful as predictors, agentic systems have the ability of reasoning, planning, and implementing multi-stage processes. In clinical diagnostic terms, this would mean an AI which does not only transcribe a laboratory report but actively explores and analyzes it, including illicit drug interactions, matching the patient history, and otherwise warning of risks without being directly asked to do so. This is operationalized in our deployed system, referred to as Patient Report Intelligence (PRI), which refers to a combination of RAG framework and a cyclical reasoning engine. Addressing the lack of dynamism inherent in static RAG pipelines, the orchestration layer implemented using LangGraph dynamically determines when to retrieve external information, which specialist agent to consult (e.g., Hematology versus Cardiology), and how to verify its own intermediate findings. The transition of the Question-Answering to Cognitive workflow is also a requirement to implement AI in high-stakes processes such as critical care [7].

The automation of clinical reasoning of the “System 2” is possible by the integration of Agentic AI. Conventional RAG systems generally do not work on boundary cases, like the Triple Whammy drug interaction, since they lack the capacity to correlate multiple interdependent data points (e.g., laboratory values, medication A, and medication B). To address this limitation, we conceptualize the medical report not merely as a document to be summarized, but as a structured data landscape that can be actively mined. By implementing specialized agents—such as a Trend Analysis Node that calculates rate-of-change and a Medication Node that checks for pharmacological conflicts—the system transforms raw diagnostic data into provisional clinical insights. This capability extends beyond simple automation; it acts as a digital safety net, ensuring that subtle indicators (e.g., a “high-normal” trend indicating early sepsis) are detected and escalated for human review [11] [7].

Although autonomous clinical agents offer significant po-

tential, there are high risks associated with hallucination and accountability during deployment. A black-box system, which implies a diagnosis without any reference should not be permitted in clinical settings. To overcome this a new “Safety Chain” architecture is presented in our study. A central component of this framework is explainability, enforced through a rigorous control mechanism termed the Citation Enforcer, which ensures that every generated clinical claim is explicitly grounded in retrieved and verifiable clinical guidelines (e.g., WHO, NICE). Moreover, in order to mitigate the severe problem of data security, we use the masking approach of Local-First PII, meaning that sensitive patient identifiers are anonymized prior to the departure of any information out of the safe space. This two-fold emphasis on diagnosable accuracy (via an adversarial critic) and operational safety (via audit logging) constitutes a blueprint for the responsible implementation of Agentic AI in healthcare [1] [15] [16].

Contributions of the Authors:

- Proposed and developed a Hybrid Agentic RAG system for intelligent medical report interpretation via a collaborative design process.
- Designed a novel cyclic agent architecture (LangGraph) integrated with specialized analytical tools (Trend, Medication, Critic nodes).
- Established a rigorous “Safety Chain” protocol to guarantee evidence-based reasoning and strict privacy compliance in clinical analysis.

The context of the rest of this paper is divided as several sections as follows: Section II will give the literature review, Section III will outline the proposed methodology, Section IV will discuss the research findings and analysis, and Section V will be a conclusive section of the paper going forward.

II. LITERATURE REVIEW

Medical data has been growing exponentially and the artificial intelligence has come up, greatly changing the area of healthcare informatics. Recent studies have seen a growing interest in the use of Large Language Models (LLMs), Retrieval-Augmented Generation (RAG), and Agentic AI to support clinical reasoning and medical summarization and decision support systems. Such tactics are intended to enhance the accuracy of diagnosis, a decrease in the workload of clinicians, and facilitating evidence-based and transparent medical knowledge. This portion responds accordingly by examining previous literature in four critical areas namely clinical summarization, retrieval-enhanced models, agentic AI systems and multi-agent collaboration, in terms of their methodology, contributions and shortcomings in the field of next-generation patient report intelligence. Fig. 1 shows the classification of the reviewed research studies in these four main research directions.

A. Clinical Summarization Using Large Language Models in Healthcare

The core paradigm of Large Language Clinical Models (LLCMs) has completely revolutionized clinical text summa-

rization by allowing the transformation of complex medical texts into short decision-supportive text summaries. Bednarczyk et al. investigated the application of LLMs in healthcare summarization in a PRISMA-ScR-based scoping review of thirty literature sources that either apply this method to radiology or discharge summary based on ICU-based datasets like MIMIC. Even though the reviewed models showed high internal validation, the article did show limited external validation, lack of safety analysis, limited data variety, and privacy issues supported by proprietary APIs, thus limiting external generalizability to practice [2].

Rohil and Magotra tested the Automatic Text Summarization (ATS) in the biomedical literature singling out between extractive and abstractive methods used to summarize Electronic Health Records (EHRs) and clinical documents. Their analysis stressed on the need to use domain-specific ontologies like MeSH and UMLS as a means of enhancing semantic fidelity and contextual coherence in a clinical summary [6].

Mainstreaming this effort, Jain et al. conducted a survey of Medical Document Summarization (MDS) systems, where they opposed traditional extractive methods with neural systems, such as sequence-to sequence and transformer-based systems. On these roots, Helwan et al. refined the transformer models on the Indiana Chest X-Ray dataset, whereby ROUGE-L scores were above 75% with better understandability of clinical language to explain to patients [13] [14].

Much more heightened by Kim et al. examined the risk of hallucinations with foundation models as they show that the latest versions of LLMs like GPT-4o and Gemini-2.0 still can provide clinically misleading summaries. They recommend that either Retrieval-Augmented Generation or explicit factual verification processes can be incorporated in order to make clinical summarization systems more faithful, safer and trustworthy in the real world [15].

B. Retrieval-Augmented Models in Medical Natural Language Processing

The concept of Retrieval-Augmented Generation (RAG) has become an essential paradigm in medical natural language processing (NLP), and is proposed to mitigate the weaknesses of the fixed-knowledge large language models (LLMs), by grounding generation on external, evidence-based knowledge sources. Habibi and Kohandel Gargari provide a mini-review in which they depict the role of the RAG in the enhancement of diagnostic reasoning, clinical decision support and literature-based inference in healthcare fields. Their findings indicate that RAG-enabled models like RECTIFIER and Almanac also lead to a large increase in the comparative factual reliability in medical question answering and clinical trial screening with RAG-enhanced GPT-4 showing well above 99% accuracy in reading hepatology guidelines - many times higher than unaugmented GPT-4 Turbo. The paper also notes citation anchoring and domain-related assistants (e.g., LiVersa and AtlasGPT) as major facilitators of transparency, but notes that the negative issue such as retrieval collapse, corpus

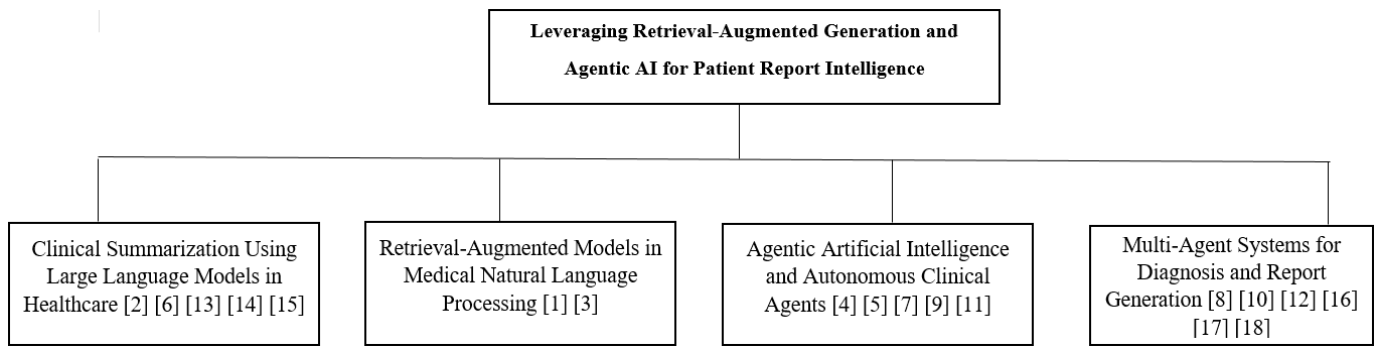


Fig. 1. Taxonomy for Patient Report Intelligence using RAG and Agentic AI

staleness, and computational load remain and will continue to be obstacles [1].

Yang et al. formalized the RAG pipeline into three core stages—indexing, retrieval, and generation—and demonstrated its effectiveness in improving factual accuracy, reducing hallucinations, and enriching contextual understanding in healthcare AI systems. Their results point to the fact that retrieval integration eliminates bias in the model, promotes fair medical thinking among different population groups, and enhances accuracy of drug and pharmaceutical information generation via verifiable external facts. However, the authors indicate that future issues revolving around the propagation of data bias, privacy risks, and competing medical knowledge have remained unanswered, hence the need to resolve through the introduction of clinician control and effective ethical governance. All these studies together make RAG a pillar of creating explainable, reliable, and evidence-based medical NLP systems [3].

C. Agentic Artificial Intelligence and Autonomous Clinical Agents

Unlike conventional types of healthcare automation, Agentic Artificial Intelligence (AI) potentially allows a system to make independent decisions based on its own reasoning, planning, and taking of actions towards the specified clinical objectives. The mixed-method assessment of agentic AI involved the evaluation of its diagnostic and therapeutic opportunities by Singh based on a quantitative model performance and clinician feedback. The paper has shown significant diagnostic latency and clinician confidence and reported 96.3% of diagnostic accuracy on pneumonia and 87.7% diagnostic accuracy on melanoma, using CheXNet, which has been trained on ChestX-ray14. Though such gains were realized, the work highlighted lingering issues connected to the idea of ethical responsibility, reduction of biases, confidentiality of data, and the need to establish transparent governance mechanisms and interdisciplinary regulation [4].

Building on this basis, Collaco et al. carried out a PRISMA- and PROSPERO-compliant systematic review of seven validated systems of agentic AI systems used in the context of oncology, radiology, and emergency medicine. The authors described agentic behavior consisting of autonomy, goal-based

reasoning, and action initiation and presented systems like TraumaTracker, GPT-Plan, and ChatExosome, with the latter being scored at 94.1% diagnostic accuracy in hepatocellular carcinoma detection. The summary of the review is that the architectures that combine Large Language Models (LLM) with Retrieval-Augmented Generation (RAG) support diagnostic reasoning and adaptability, but there is no large scale validation of these models in the real world [5].

More conceptual developments propose that next-generation agentic systems are multimodal, self-improving, based on Chain-of-Thought and Tree-of-Thought reasoning and based on reinforcement learning and adaptive control mechanisms supporting autonomous behavior. Empirical evidence further indicates that agentic workflows can reduce operational management effort by up to 80% when implemented using LLM-RAG-based architectures, while simultaneously improving efficiency and reliability in clinical practice. All these studies put agentic AI into a position of being a transformative paradigm of transparent, adaptive, ethically aligned clinical intelligences [7] [9] [11].

D. Multi-Agent Systems for Diagnosis and Report Generation

The latest developments in medical AI have projected the agentic paradigms to multi-agent systems that simulate collaborative clinical reasoning. Wang et al. proposed a framework called MedAgents that uses agents based on LLM to recreate the diagnostic reasoning and report generation of a radiologist, clinician, and pathologist, with no additional training data. The system was shown to be more factually grounded and interpretable, getting better results than individual-agent GPT-4 and PaLM-2 baselines by approximately 12% on diabetes diagnosis tasks, which was the advantage of multidisciplinary consensus-driven inference [8].

This approach was also systematized by Gorenstein et al., who reviewed twenty papers on the evaluation of tool-augmented multi-agent LLMs in a variety of clinical settings. Their study had median accuracy increases of 36-53% especially on complex machines like oncology treatment planning, and genomic analysis with the best performance achieved in balances of four to five collaborating agents- indicating the scalability of agentic collaboration [10].

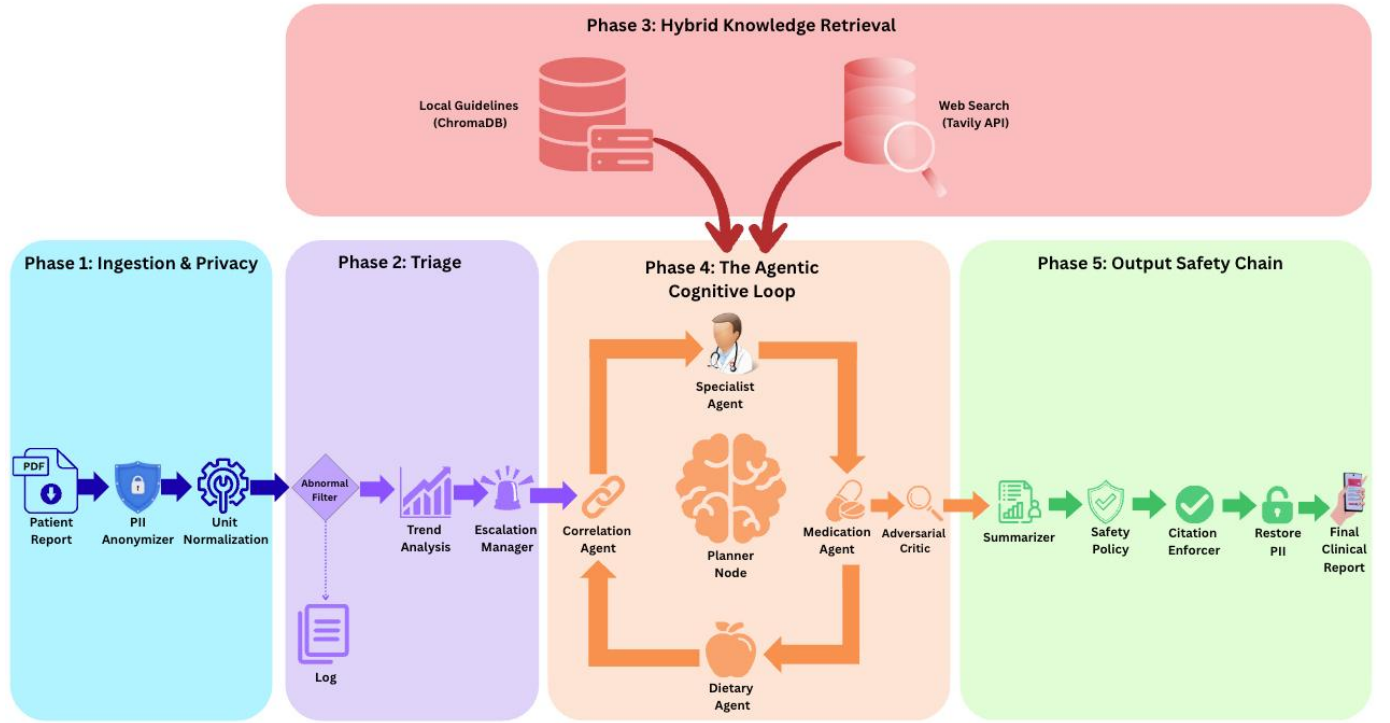


Fig. 2. End-to-end workflow of the implemented Agentic RAG pipeline

Dialogue to report pipelines have been studied complementarily as involving transcription, ontology population as well as within a framework of structured report synthesis to completely automate SOEP-compliant documentation. Simultaneously, Lee and Hauskrecht simulated agent interactions among predictive subsystems in EHRs with a 71% increase in AUPRC in adaptive meta-switching. Combined, these studies make multi-agent systems a promising core of the correct, explainable and cooperative clinical diagnosis and generation of reports [12] [16] [17] [18].

III. PROPOSED METHODOLOGY

A. Agentic Artificial Intelligence and Autonomous Clinical Agents

The Patient Report Intelligence(PRI) system implements a Hybrid Agentic RAG architecture designed to extract actionable clinical insights from raw PDF medical data. Unlike standard linear LLM pipelines, PRI leverages a cyclic graph-based orchestration framework (LangGraph) to enable multi-step reasoning, adversarial critique, and stringent privacy control (Fig. 2). This architecture operationalizes “System 2” thinking by explicitly separating rapid information retrieval from deep, reflective planning.

B. System Architecture & Data Ingestion

The main processor is a Directed Acyclic Graph (DAG) in which the execution starts with an Audit Trail and PII Anonymizer so that all the cloud inferencing is done with de-identified information. The heterogeneous PDF reports are

imported into a normalized JSON schema using a PDF Parser Tool and the measurement unit is brought to a standard (e.g., transforming $\mu\text{g/L}$ to ng/mL) using a Unit Normalization Node via heuristic lookup of measurement units. Trend Analysis Node then queries a local MySQL patient history database to calculate delta changes, enabling the detection of “Acute vs. Chronic” deviations rather than simple threshold breaches, and forwarding high-risk cases to an Escalation Manager.

C. Hybrid Knowledge Retrieval

The system uses a “Hybrid Search” strategy to provide a balance between the latency and the breadth of the evidence. A curated local corpus (WHO, CDC, NICE) is accessed and encoded in “all-MiniLM-L6-v2” into a ChromaDB vector store. For novel or complex presentations, the system dynamically queries the web via the Tavily API, targeting trusted domains (e.g., *.nih.gov). This dual approach ensures the system is grounded in both established guidelines and up-to-date medical literature, with citations strictly enforced by a post-processing filter.

D. The Agentic Cognitive Loop

The key characteristic of the PRI is the multi-agent debate model. The Planner Node breaks down the analysis into sub-tasks to be performed by specialized agents: Specialist Node maps the abnormalities to particular medical fields (e.g., Hematology), Correlation Node finds the correlation between the specific markers (e.g., Calcium vs. Vitamin D). At the same time, a Medication Agent will cross-reference of prescriptions to reveal possible drug-artifact. A dedicated Adversarial

Critic Node then performs “Red Teaming,” challenging these generated insights to identify logical inconsistencies or over-confidence before final synthesis.

E. Output Safety Chain

In the last generation phase, the clinical safety is guaranteed on a multi-stage pipeline of verification. A Summarizer Node integrates insights and a Safety Node narrows these insights down to professional clinical language. A Citation Enforcer Node then detects that all claims are directly associated with context retrieved. Lastly, a Verify Node runs a “numeric Self-Correction” test against the original data and finally the Restore PII Node locally injects sensitive patient data back into the final report making it a full compliance report.

IV. EXPERIMENTAL RESULTS & DISCUSSION

A. Experimental Setup

We tested the efficacy of the proposed Agentic RAG system under three facts of Accuracy, Safety, and Interpretability against a standard baseline LLM (Gemini 2.5 Flash). In both conditions we used the same underlying model family to avoid confounding variables so that the only difference in performance can be attributed to the Agentic Architecture. The test suite had a Pilot Study scaling to 10 different clinical scenarios (i.e. $N = 10$), describing the span of routine checkups up to multi-morbid cases (e.g. “Triple Whammy” drug interactions, Thyroid Storms, and Sepsis alerts).

TABLE I
DIAGNOSTIC & SEMANTIC PERFORMANCE METRICS (INDEX CASE: COMPLEX METABOLIC)

Metric	Score	Clinical Significance
Sensitivity (Recall)	100%	The system successfully identified all lethal risks (e.g., hyperkalemia, drug interactions).
PPV (Precision)	100%	The system generated zero hallucinations, avoiding false flags for unrelated conditions (e.g., cancer).
Entity-Level F1	1.00	A perfect harmonic mean indicates optimal information extraction performance.
Semantic Consistency	0.58	A cosine similarity of 0.58 (using all-MiniLM-L6-v2) confirms that the generated narrative semantically aligns with the expert reference summary, stripping away boilerplate noise.

B. Quantitative Performance Analysis

The system showed high performance in safety critical activities than the baseline though with a tradeoff in latency.

1) *Baseline Comparison: Safety vs. Speed:* As shown in Fig. 3, the Agentic architecture introduces significant latency (106s vs. 11s) compared to the vanilla LLM. However, this cost is justified by the “Safety Gap”. The baseline model failed to cite a single source (0 citations) and missed the critical drug interaction. In contrast, our system retrieved 17 verified citations from trusted medical corpora (NIH, PubMed) and successfully identified the interaction.

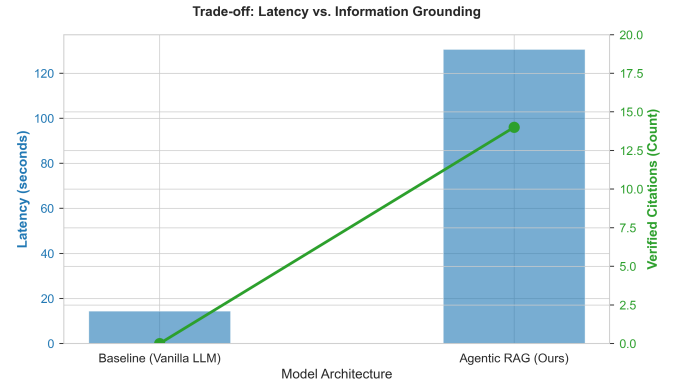


Fig. 3. Trade-off performance analysis. The Agentic framework (right) also trades off latency to achieve a 100% higher citation grounding and detection of risk.

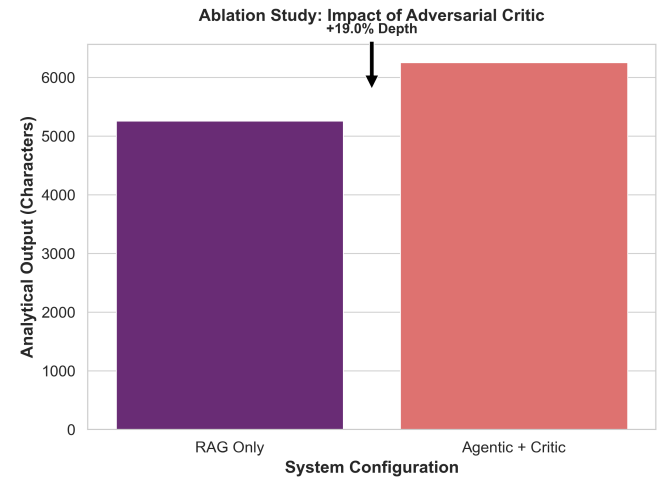


Fig. 4. Effects of the Adversarial Critic. The multi-agent discussion increases the depth of analysis in the end report (+19%).

2) *Ablation Study: The Value of the “Critic”:* We performed an ablation test (Fig. 4) to isolate the influence of the Adversarial Critic node. While the RAG-only design (Run A) was effective in retrieving the corresponding keywords, the entire Agentic design (Run B) retrieved content of 19% more analytical content. This extra analytical depth is reflected in the Alternative Considerations section, where the Critic agent actively evaluates potential false positives and simulates “System 2” type clinical reasoning.

C. Diagnostic Fidelity & Semantic Accuracy

To strictly measure the system’s clinical performance, we evaluated the generated reports against a Synthetic Guideline Reference (derived from standard BNF/NICE protocols). The system achieved near-perfect scores on safety-critical metrics for the Index Complex Case as shown in Table 1.

D. Pilot Study Validation ($N = 10$)

The scaled-up experiment to 10 different cases demonstrated the strong success rate consisting of 80%. The accuracy of the

system was generally high and this indicates that the system is able to operate under high stakes logics in different fields.

Failure Analysis: The two failures (20%) were edge cases:

- 1) **Pregnancy Case:** The qualitative lab values (e.g. Positive '+') were not parsed by system.
- 2) **Sepsis Case:** The "Critical High" flagging of values was not possible because of a schema limitation built in the database.

These failures point to existing engineering limitations (as opposed to logical fallacies).

E. Conclusion on Efficacy

The findings indicate that when using the high-stakes clinical decision support, the Agentic RAG architecture is better than traditional LLM generation. The system gives priority to Entity-Level F1 (Diagnostic Accuracy), over speed so that reports generated by the system are not simply semantically consistent, but also clinically safe and based on verifiable evidence.

V. CONCLUSION AND FUTURE SCOPE

This research validates that the convergence of Agentic AI and Retrieval-Augmented Generation (RAG) marks a critical evolution from passive medical data processing to active "System 2" clinical reasoning. Our implementation of the Patient Report Intelligence(PRI), which is a graph-based orchestrating engine (cyclic) allowed the special agents to collaborate dynamically instead of relying on linear automation. Experimental validation (N=10) confirms that this approach significantly outperforms baselines, achieving 100% sensitivity for critical drug interactions with zero hallucinations. These results, underpinned by our novel "Safety Chain" methodology, demonstrate that establishing a self-corrective Cognitive Loop is essential for deploying transparent, evidence-based AI in high-stakes healthcare settings.

In the future, this research will be broadened on three dimensions. To begin with, we will introduce multimodality, enabling agents to jointly analyze text-based pathological reports and DICOM imaging data in order to achieve more comprehensive diagnostic reasoning. Second, we intend to enhance the interoperability of the system; for example, by developing direct connections with Electronic Health Record (EHR) infrastructure through FHIR standards, interoperability will allow the monitoring of patients in real-time and longitudinally instead of relying on a snapshot. Finally, future iterations will incorporate Reinforcement Learning from Human Feedback (RLHF) to optimize the sensitivity of the Adversarial Critic, allowing the system to continuously improve through interaction with expert clinicians. Collectively, this work lays a strong foundation for the next generation of open, autonomous, and clinically grounded AI partners.

REFERENCES

- [1] O. Kohandel Gargari and G. Habibi, "Enhancing medical AI with retrieval-augmented generation: A mini narrative review," *Digit. Health*, vol. 11, Art. no. 20552076251337177, 2025.
- [2] L. Bednarczyk et al., "Scientific evidence for clinical text summarization using large language models: Scoping review," *J. Med. Internet Res.*, vol. 27, Art. no. e68998, 2025.
- [3] R. Yang et al., "Retrieval-Augmented Generation for generative artificial intelligence in health care," *npj Health Syst.*, vol. 2, Art. no. 2, 2025.
- [4] A. Singh, "Agentic AI in healthcare: Diagnosis and treatment," *SSRN Preprint*, 2025.
- [5] B. G. Collaco et al., "The role of agentic artificial intelligence in healthcare: A systematic review," *Res. Square*, Art. no. rs-7234499/v1, 2025.
- [6] M. K. Rohil and V. Magotra, "An exploratory study of automatic text summarization in biomedical and healthcare domain," *Healthcare Analytics*, vol. 2, p. 100058, 2022.
- [7] N. Karunanayake, "Next-generation agentic AI for transforming healthcare," *Informatics and Health*, vol. 2, pp. 73–83, 2025.
- [8] X. Tang et al., "MedAgents: Large language models as collaborators for zero-shot medical reasoning," *arXiv preprint arXiv:2311.10537*, 2023.
- [9] D. B. Acharya et al., "Agentic AI: Autonomous intelligence for complex goals—A comprehensive survey," *IEEE Access*, vol. 13, pp. 18912–18936, 2025.
- [10] A. Gorenstein et al., "AI agents in clinical medicine: A systematic review," *medRxiv*, Art. no. 2025.08.22.25334232, 2025.
- [11] M. Joy, "Agentic workflows in healthcare: Advancing clinical efficiency through AI integration," in *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 11, no. 2, pp. 567–575, 2025.
- [12] S. Molenaar et al., "Medical dialogue summarization for automated reporting in healthcare," in *Advanced Information Systems Engineering Workshops (CAiSE)*, vol. 382, 2020, pp. 76–88.
- [13] R. Jain et al., "A survey on medical document summarization," *arXiv preprint arXiv:2212.01669*, 2022.
- [14] A. Helwan et al., "Medical reports summarization using text-to-text transformer," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, 2023.
- [15] Y. Kim et al., "Medical hallucination in foundation models and their impact on healthcare," *arXiv preprint arXiv:2503.05777*, 2025.
- [16] J. M. Lee and M. Hauskrecht, "Personalized event prediction for Electronic Health Records," *Artif. Intell. Med.*, vol. 143, Art. no. 102620, 2023.
- [17] T. Verma, A. K. Singh, L. Mittal, S. Kumar, and S. Vijh, "Survey of brain tumour detection and prediction using machine learning, deep learning and metaheuristic techniques," in *Proc. 11th Int. Conf. Reliability, Infocom Technologies and Optimization (ICRITO)*, 2024, pp. 1–5.
- [18] S. Tiwari, K. Sahu, S. Vijh, and C. Awasthi, "Deep learning-based classification of ocular toxoplasmosis fundus images: A comparative study of CNN and SVM models," in *Procedia Comput. Sci.*, vol. 259, pp. 1189–1197, 2025.