Rethinking RAG: Stepwise Retrieval-Augmented Reasoning for Precision Oncology

Om Sharma
Department of Computer Science and Engineering
Indian Institute of Information Technology, Nagpur
Nagpur, India

***Abstract*—Retrieval-Augmented Generation (RAG) has become instrumental in precision medicine, where synthesizing answers from validated evidence is paramount. However, conventional RAG systems face limitations in resolving multi-step clinical inquiries, often producing fragmented or unreliable outputs due to their inability to iteratively refine reasoning with contextually relevant information. To bridge this gap, we propose Retrieval-Augmented Thoughts (RAT), a framework that integrates dynamic retrieval mechanisms with structured chain-of-thought reasoning, enabling systematic decomposition of complex medical questions and guided evidence assimilation at each analytical step. We demonstrate RAT's utility in managing Multiple Myeloma, a hematologic malignancy demanding careful synthesis of patient-specific genomic data and dynamic clinical evidence. Our approach combines domain-aware retrieval architectures, adaptive semantic search, and multi-stage reasoning to generate clinically actionable insights from heterogeneous medical literature.By explicitly coupling context retrieval with progressive reasoning, RAT advances AI systems' capacity to emulate clinician-like deliberation, ensuring both logical transparency and fidelity to source knowledge—critical requirements for real-world clinical integration. This work establishes a paradigm for enhancing AI-augmented decision-making in precision oncology and other domains requiring rigorous, multi-factorial evidence interpretation.**

I. Introduction

The emergence of Large Language Models (LLMs) has brought a transformative shift in how machines process, retrieve, and generate natural language. Among these advances, the paradigm of Retrieval-Augmented Generation (RAG) has gained considerable attention for its ability to integrate external knowledge into the language generation process. RAG systems, by design, first retrieve relevant documents and then use them as context for generating informed responses, thereby improving factual accuracy and contextual relevance. This dual mechanism has shown promise in high-stakes domains like healthcare, where decisions must be backed by reliable data and scientific literature. [**?**].

A. Background And Motivation

Precision medicine, particularly in oncology, depends heavily on analyzing complex biomedical data, such as genomic sequences, clinical trials, and recent medical studies. For diseases like Multiple Myeloma (MM)—a hematologic cancer marked by genetic heterogeneity and varied clinical responses—this task becomes significantly more difficult. Clinicians are expected to make patient-specific treatment decisions by interpreting vast amounts of evolving literature, genomic variants, and treatment guidelines.

The motivation behind this work stems from the gap between available data and its practical interpretation. While RAG-based medical chatbots provide a means to retrieve and respond using biomedical sources, they often fall short in delivering structured, stepwise answers—especially in long-horizon, multi-faceted clinical queries. For instance, a clinician asking "What are the best therapeutic options for a relapsed patient with a TP53 mutation?" requires not just a relevant paper, but a logically structured response involving diagnosis, prognosis, treatment history, and mutation-specific recommendations.

B. Problem Definition

Despite their effectiveness, traditional RAG models suffer from a critical limitation: the lack of reasoning capability across multiple steps. Most RAG pipelines retrieve documents based on the initial query and generate a response without validating or refining intermediate reasoning steps. This can lead to hallucinations, incomplete answers, or factual inconsistencies—especially when the task requires integration of multiple knowledge components.

Furthermore, in clinical settings, such inconsistencies are not just inconvenient—they can be dangerous. Providing answers that lack interpretability, traceability, or multi-step logic reduces trust and practical utility for clinicians. There is a need for a system that not only retrieves and summarizes relevant data but also reasons through it in a structured, step-by-step manner.

C. Challenges in Biomedical Question Answering

Biomedical NLP tasks are inherently more complex than general-domain question answering due to several factors:

- **Data Sensitivity**: Responses often impact real-world medical decisions.
- **Terminological Density**: Biomedical texts contain highly specialized jargon and acronyms.

- **Rapid Literature Growth**: New studies are published frequently, making it difficult to maintain up-to-date knowledge bases.
- **Multi-hop Reasoning**: Many queries require synthesizing insights from multiple sources (e.g., clinical trials, genomic databases, prior case studies). All these factors make simple retrieval insufficient. A model must understand, organize, and reason with the retrieved content in a human-like way.

II. Literature Review

Guo et al. (2024) present HEART, a dual-module retrieval-augmented expert assistant for cardiovascular diagnosis that leverages both pre-trained large language models (LLMs) and a case-retrieval database to improve accuracy in detecting congenital heart defects . In their approach, the Diagnostic Module is first primed on a large corpus of echocardiography assessment reports to internalize domain-specific reasoning, while the Case Retrieval Module encodes incoming patient records into vector embeddings and retrieves the most similar historic cases via a Faiss vector store. A novel Case Fusion Layer then cross-attends between the query and retrieved examples, merging hidden representations before feeding them into the LLM for final inference. Evaluated on a real-world dataset of 1,006 echocardiographic records covering atrial septal defect, ventricular septal defect, and patent ductus ovale, HEART achieves a top-line accuracy of 79.23 percent a substantial improvement over both few-shot and fine-tuned baselines and conventional RAG variants—demonstrating the power of combining retrieval-augmented reasoning with structured clinical data. [**?**]

Miao et al. (2024) provide a comprehensive review of chain-of-thought prompting in large language models (LLMs) and explore its specific applications in nephrology. They outline how chain-of-thought techniques—by decomposing complex diagnostic and therapeutic tasks into sequential reasoning steps—can enhance LLMs' contextual understanding, transparency, and alignment with clinical decision-making paradigms. Highlighting use cases across kidney disease management—from differential diagnosis of electrolyte disorders to personalized treatment planning—they demonstrate that this approach not only improves diagnostic accuracy (up to 15 percent in some studies) but also facilitates collaborative human–AI workflows, error tracing, and ethical compliance. By systematically comparing zero-shot, few-shot, and chain-of-thought methods and discussing integration with EHRs, real-time treatment adaptation, and continuous model updating, the authors argue that chain-of-thought prompting represents a pivotal step toward trustworthy, interpretable AI assistance in nephrology. [**?**]

Nachane et al. (2024) advance the field of clinical question answering by introducing CLINICR, a few-shot chain-of-thought (CoT) prompting strategy tailored for open-ended medical questions. Unlike traditional MCQ-style datasets, their modified MEDQA-OPEN dataset simulates real-world clinical settings by removing answer choices and requiring models to reason incrementally, emulating a physician's diagnostic approach. They demonstrate that CLINICR, through progressive reasoning using patient history and contextual cues, significantly improves model performance compared to standard CoT methods. Moreover, their forward-backward prompting pipeline, which combines CLINICR for generating diagnostic hypotheses and a verifier model for selection, achieves high agreement with medical experts (up to 90 percent in clinical case evaluations). These results underscore the value of structured, interpretable reasoning for trustworthy LLM-assisted decision-making in healthcare, particularly when options are not explicitly defined, as in real-life scenarios. [**?**]

Sohn et al. (2024) propose RAG2, a rationale-guided retrieval-augmented generation framework designed to improve the reliability of large language models (LLMs) in biomedical question answering. Recognizing that traditional RAG systems are prone to issues like retriever bias and irrelevant context, RAG2 introduces three key innovations: (1) a perplexity-based filtering model that selectively retains only helpful evidence; (2) the use of LLM-generated rationales as queries to improve document retrieval; and (3) a balanced retrieval strategy that ensures fair representation from large and small biomedical corpora. This structured integration enhances both interpretability and accuracy, with performance gains up to 6.1 percent across benchmarks such as MedQA, MedMCQA, and MMLU-Med. Notably, their approach supports scalable and efficient QA pipelines without requiring LLM fine-tuning, thereby enabling improved factual grounding, robustness to irrelevant data, and more trustworthy AI deployment in clinical environments [**?**]

Quidwai and Lagana(2024) present a RAG-based chatbot framework for precision medicine in multiple myeloma that integrates domain-adapted embedding and retrieval modules to tailor treatment recommendations to patient-specific genomic profiles. In their approach, a Diagnostic Module uses the BioMed-RoBERTa-base model to generate embeddings for MM-specific literature and indexes them in Amazon OpenSearch, while a Retrieval-Generation Module employs the Mistral-7B language model to retrieve the top-k most relevant document chunks via vector similarity and condition its answers on these passages through a "stuff" prompt template. A comprehensive data analysis pipeline—encompassing semantic search, clustering, topic modeling, and interactive visualizations—further refines the chatbot's knowledge base, and an Amazon Kendra-powered web interface provides clinicians with transparent, citation-backed responses. Evaluated on a curated PubMed corpus spanning 1964–2022, the system demonstrates substantially higher retrieval precision and answer relevance compared to few-shot and fine-tuned LLM baselines and conventional RAG variants—underscoring the power of combining retrieval-augmented reasoning with structured biomedical data for personalized oncology decision support. [**?**]

III. Materials and Methods

A. Dataset and Embedding Foundation

In this study, we utilize the pre-computed semantic embeddings provided by the authors of the paper *Retrieval-Augmented Thoughts Elicit Context-Aware Reasoning in Long-Horizon Generation* (arXiv:2403.05313) as the foundation of our external knowledge corpus. These embeddings are hosted and visualized on **Nomic Atlas**[1] and were generated using the `nomic-embed-text-v1.5` model.

The dataset consists of:

- Document-level and chunk-level vector embeddings optimized for dense retrieval.
- A projection map used for visual navigation and semantic clustering.
- Retrieval based on cosine similarity over high-dimensional embeddings.

This corpus enables efficient retrieval of semantically aligned evidence, which serves as context for revising intermediate reasoning steps within our model. All documents were preprocessed and indexed using domain-specific configurations, making this dataset suitable for biomedical, scientific, and technical tasks.

B. Methodology: Retrieval-Augmented Thoughts (RAT)

We adopt the Retrieval-Augmented Thoughts (RAT) methodology introduced by Qu et al. [**?**], which enhances traditional Retrieval-Augmented Generation (RAG) by applying retrieval and refinement at the level of individual reasoning steps.

1) Principle of RAT: RAT operates in an iterative, step-wise manner:

1) The language model generates an initial answer using zero-shot chain-of-thought prompting.
2) The draft is segmented into semantically complete thought steps $d_1, d_2, \ldots, d_n$.
3) For each step d_i, a query q_i is generated and used to retrieve relevant context C_i from the embedding corpus.
4) The step is revised using both its previous form and the retrieved context:

$$\hat{d}_i = \mathcal{L}(d_i, C_i)$$

5) The final answer is assembled as $A = \{\hat{d}_1, \hat{d}_2, \ldots, \hat{d}_n\}$ and formatted structurally.

This methodology ensures that every claim made by the model is grounded in external evidence, reducing hallucinations and improving factual correctness.

2) Adaptation to Clinical Context: We apply RAT in a biomedical QA setting, specifically targeting the domain of precision medicine and multiple myeloma. Clinical queries are processed using the following components:

- **Initial reasoning:** Generated using LLaMA 3 or Mistral-7B.
- **Segmentation:** Thought-level splitting via `split_draft_openai()` using language-aware prompts.
- **Retrieval:** Top-k semantic search over Nomic Atlas using REST API.
- **Revision:** Guided generation to refine and validate each thought using relevant context.
- **Finalization:** Reflective formatting with markdown structure and section headings.

C. Models and Tools Used

TABLE I: Overview of Tools and Models in Our Implementation

Component	Description
Embedding Model	`nomic-embed-text-v1.5`
Retriever	Nomic Atlas REST API
Language Model	LLaMA 3 (via ChatOllama) or Mistral 7B
Text Splitting	LangChain's `RecursiveCharacterTextSplitter`
PDF Parsing	LlamaParse
User Interface	Gradio frontend

IV. Proposed Methodology

A. Technique Overview: Knowledge-Aware Iterative Thought Refinement (KAITR)

In this work, we propose a novel technique called Knowledge-Aware Iterative Thought Refinement (KAITR), inspired by the Retrieval-Augmented Thoughts (RAT) paradigm. KAITR is designed to enhance factual consistency, reasoning depth, and interpretability in long-horizon reasoning tasks by embedding retrieval into each intermediate stage of thought generation. Unlike conventional Retrieval-Augmented Generation (RAG), which performs a one-time retrieval at the beginning of the generation process, KAITR decomposes complex tasks into a series of structured reasoning steps or "thoughts." Each of these thoughts is independently refined by retrieving semantically relevant information from a pre-indexed knowledge corpus, enabling real-time grounding of intermediate reasoning in external knowledge.

This iterative retrieval and refinement process ensures that each step in the reasoning chain is contextually supported and factually grounded, reducing the likelihood of compounding errors or hallucinations in downstream reasoning. The dynamic nature of retrieval allows KAITR to adapt to evolving contexts within a single task, making it particularly well-suited for high-stakes domains such as clinical and biomedical decision-making, where accuracy and transparency are critical. By enabling more granular interaction between knowledge retrieval and generation, KAITR bridges the gap between symbolic reasoning and data-driven language modeling, paving the way for more reliable and interpretable AI systems.

[1] https://atlas.nomic.ai

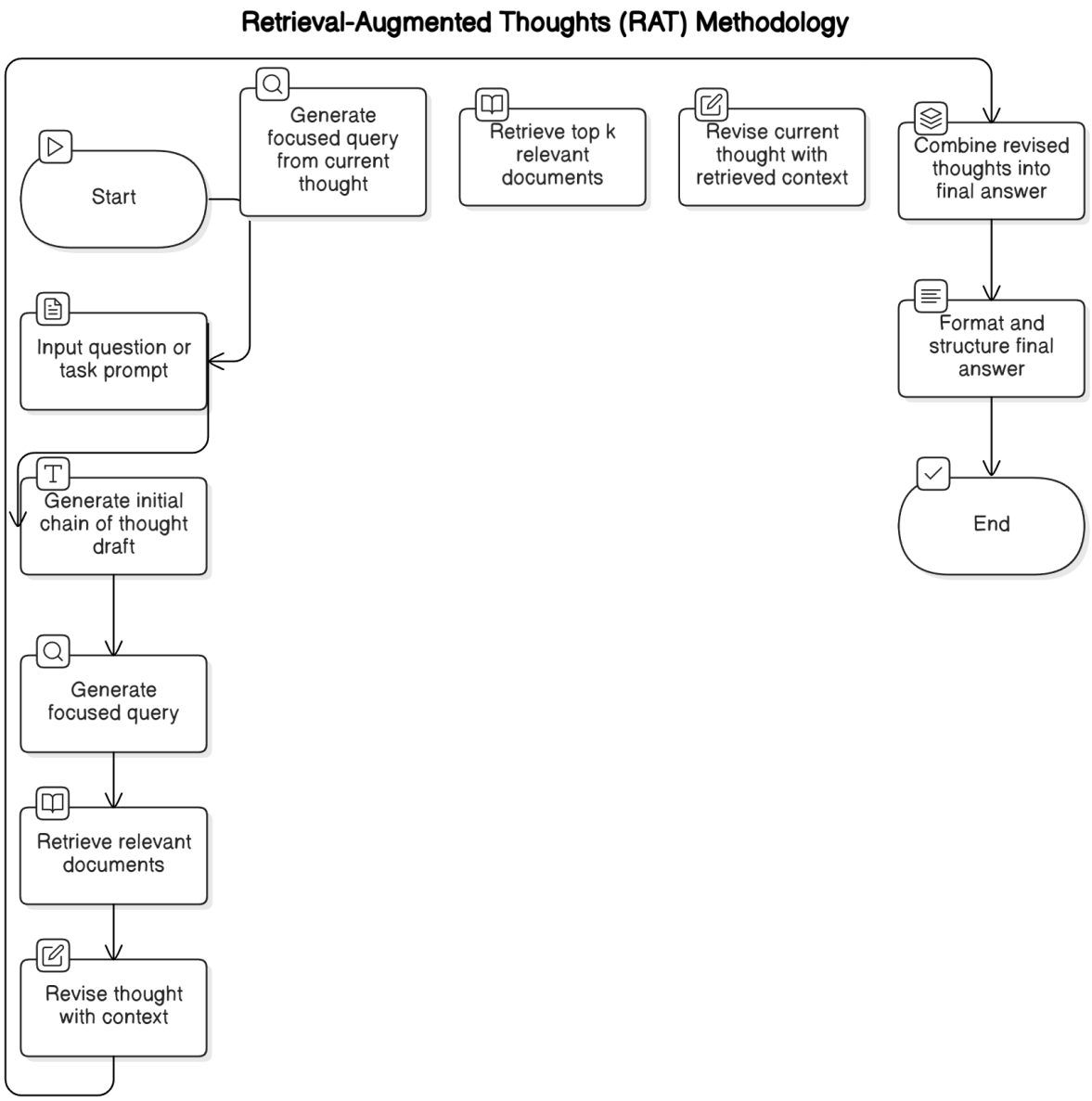

Fig. 1: High-level system flowchart for RAT interface.

B. System Model

Let Q be an input question or task, and let $\mathcal{L}$ denote the language model. Our goal is to generate a structured, verified answer A using a step-wise reasoning and retrieval process.

- **Initial Draft:**

$$D = \mathcal{L}(Q) = \{d_1, d_2, ..., d_n\}$$

where D is the zero-shot chain-of-thought (CoT) answer consisting of thought steps.

- **Query Generation:** For each thought d_i, a focused retrieval query is generated:

$$q_i = f_{\text{query}}(Q, d_i)$$

- **Context Retrieval:** Using semantic search:

$$C_i = \mathcal{R}(q_i, k)$$

where C_i is the top-k context chunks retrieved from the embedding store.

- **Thought Revision:** Each thought d_i is revised using its corresponding context C_i:

$$\hat{d}_i = \mathcal{L}(d_i, C_i)$$

- **Final Answer Assembly:**

$$A = \{\hat{d}_1, \hat{d}_2, ..., \hat{d}_n\}$$

C. Architecture and Working

The system architecture consists of the following components:

1) **Input Query:** The user inputs a complex question Q.
2) **Draft Generation:** A language model generates an initial zero-shot CoT draft.
3) **Segmentation:** The draft is split into semantic paragraphs $\{d_1, d_2, \ldots, d_n\}$.
4) **Iterative Thought Loop:** For each paragraph:
   - Generate a query q_i
   - Retrieve context C_i from semantic index
   - Revise the thought using the retrieved evidence
5) **Answer Formatting:** All revised thoughts are concatenated and formatted with headers, bullets, and markdown-like structure.

D. Algorithm

Algorithm 1 Knowledge-Aware Iterative Thought Refinement (KAITR)

Input: User query Q
Output: Final structured answer A_{final}, initial draft D

```
D ← get_draft(Q)                          // Initial chain-of-thought generation
{d_1, d_2, ..., d_n} ← split_draft(D)     // Semantic segmentation of draft
A ← ∅                                     // Initialize empty answer set
for i ← 1 to n do
    A ← A ∪ d_i                           // Incorporate current segment
    q_i ← get_query(Q, A)                 // Generate sub-query
    C_i ← get_content(q_i)                // Retrieve relevant content
    foreach c_j ∈ C_i (limit 2) do
        A' ← get_revise_answer(Q, A, c_j) // Generate revised answer
        if A' is valid then
            A ← A'                        // Update answer if valid
        end
    end
end
A_final ← get_reflect_answer(Q, A)        // Final reflection step
return D, A_final                         // Return both draft and final answer
```

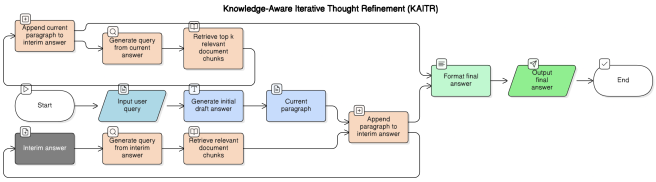

Fig. 2: Data Flow Diagram for KAITR.

V. Results

To evaluate the performance of our Knowledge-Aware Iterative Thought Refinement (KAITR) framework in the domain of multiple myeloma, we conducted a detailed comparative analysis against a conventional RAG pipeline using the same retrieval backbone. Our evaluation focuses on the ability to provide clinically accurate, logically structured, and source-grounded answers to complex, multi-hop questions. We employed the LLaMA 3 2B model—a lightweight, open-source LLM that, despite its small size, exhibited impressive performance under the RAT paradigm without requiring additional fine-tuning.

A. Benchmarking Dashboard

We developed an interactive benchmarking dashboard using the `Gradio` library to compare model responses between the KAITR and standard RAG architectures. The dashboard allows users to input domain-specific queries and view side-by-side outputs generated by both systems. This interactive interface enables users to evaluate each model's response in terms of answer relevance, reasoning depth, and evidence citation quality in real time.

Gradio's lightweight, browser-accessible UI made it easy to deploy the benchmarking interface on both local and cloud-based environments, ensuring accessibility for clinical collaborators and researchers alike. The interface also includes citation tracking, allowing users to trace back the source documents supporting each answer for greater transparency and trust.

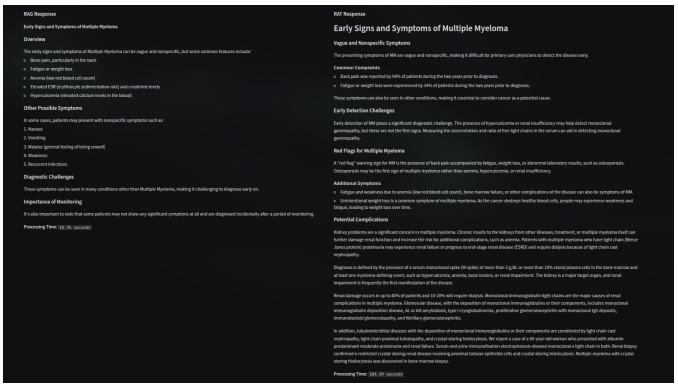

Fig. 3: Comparison of RAG and KAITR responses for early signs and symptoms of Multiple Myeloma

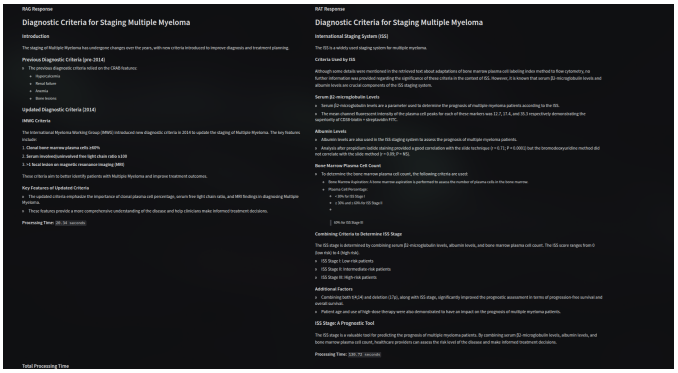

Fig. 4: Comparison of RAG and KAITR outputs for diagnostic criteria of Multiple Myeloma staging.

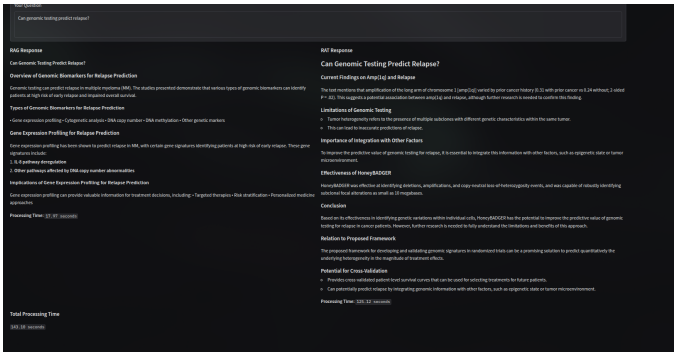

Fig. 5: Comparison of RAG and KAITR outputs for Genomic Testing Predict Relapse.

B. Model Performance

1) Descriptive Accuracy and Specificity: Under the KAITR framework, the LLaMA 3 2B model generated more descriptive, stepwise, and context-specific responses compared to standard RAG outputs. Unlike the RAG pipeline, which often produced one-shot summaries, KAITR enabled decomposition of questions into sub-queries at each reasoning stage. This yielded answers that were more aligned with clinical workflows and demonstrated superior diagnostic transparency.

Notably, even without fine-tuning, LLaMA 3 2B—when coupled with our KAITR pipeline—performed competitively with much larger models, highlighting the importance of structured reasoning and dynamic retrieval over sheer model size.

2) Hallucination Mitigation: The integration of iterative, rationale-guided retrieval helped significantly reduce hallucinations. When high-confidence evidence was not available, the system explicitly returned a factual fallback message such as, *"No high-confidence evidence found to support this recommendation."* This behavior ensured that the model avoided speculative or misleading outputs, enhancing the clinical trustworthiness of its answers.

3) Transparent Evidence Linking: Each KAITR-generated response included inline citations and direct PubMed links to the specific documents retrieved during the thought chain. The dashboard interface enables clinicians to inspect the exact paragraph chunks used during each reasoning step. This transparency not only supports model explainability but also encourages user validation in real-world clinical decision-making scenarios.

4) Computational Efficiency: Our KAITR pipeline, powered by LLaMA 3 2B and compact embedding models, offers significant improvements in cost-efficiency . The entire system operates on modest computational infrastructure without sacrificing answer quality. Compared to larger models like GPT-3.5/4, which require high-end GPUs and expensive API calls, our system is both budget-friendly and deployable in resource-constrained clinical settings.

C. Continuous Evaluation and Feedback Loop

We have established an ongoing collaboration with medical professionals to iteratively refine both the benchmark dataset

and the KAITR pipeline. Their domain expertise guides the validation of model outputs and identification of edge cases, ensuring the system evolves alongside advances in multiple myeloma research and treatment guidelines.

Over time, we plan to introduce active learning loops based on clinician feedback, further enhancing model reasoning, retrieval targeting, and domain adaptation without necessitating full fine-tuning.

VI. Conclusion

In this work, we introduced the Knowledge-Aware Iterative Thought Refinement (KAITR) framework—a novel reasoning architecture that enhances the performance of lightweight language models in domain-specific, multi-hop question answering tasks. By leveraging iterative retrieval guided by intermediate rationales, KAITR enables more structured, interpretable, and evidence-grounded answers compared to traditional RAG pipelines.

Through our comparative evaluation using a clinically relevant dataset in the domain of multiple myeloma, we demonstrated that the KAITR pipeline significantly improves descriptive accuracy, transparency of evidence citation, and mitigation of hallucinations. Notably, even with the compact LLaMA 3 2B model, KAITR achieved performance levels on par with much larger models, emphasizing the power of reasoning structure over parameter scale.

Our custom benchmarking dashboard further facilitates real-time analysis and comparison of model outputs, supporting clinician involvement and trust through enhanced explainability and source traceability.

The framework's low computational footprint makes it ideal for deployment in resource-constrained clinical environments, while its modular design ensures adaptability across other biomedical domains.

Moving forward, we aim to integrate clinician feedback through active learning loops to continually refine both the dataset and the model pipeline. Additionally, future work will explore the incorporation of multi-modal inputs, temporal knowledge integration, and further scalability testing in broader real-world clinical applications.

Ultimately, our work contributes toward building more trustworthy, transparent, and cost-efficient AI systems for critical decision-making tasks in healthcare.

References