RobustBioReasoning Agent: Enhancing Robust Multi-hop Fact Extraction and Reasoning in Biomedical Large Language Models

Zeyu Lou * and Haoxuan Qi

Posted Date: 3 March 2026

doi: 10.20944/preprints202603.0231.v1

Keywords: LLMs; biomedical reasoning; robustness; multi-hop reasoning

 doi:10.20944/preprints202603.0231.v1

Article

RobustBioReasoning Agent: Enhancing Robust Multi-hop Fact Extraction and Reasoning in Biomedical Large Language Models

Zeyu Lou * and Haoxuan Qi

The Higher Technological Institute of Irapuato
* Correspondence: clopezj001@alumni.uaemex.mx

Abstract

Large Language Models face significant challenges in biomedical multi-hop reasoning, including noise interference, context sensitivity, and ensuring factual consistency. To address these limitations, we propose the Robust Biomedical Reasoning Agent (RBRA), a novel agent framework designed to significantly improve the robustness and accuracy of LLMs in this critical domain. RBRA integrates core mechanisms such as hierarchical query decomposition, dynamic context filtering and aggregation, and iterative fact verification and refinement, underpinned by Robustness-aware Metric Optimization. Zero-shot evaluations on BioMultiHopQA-Dynamic, a challenging dataset designed to rigorously assess robustness, confirm its efficacy. RBRA-GPT4o achieves state-of-the-art average accuracy (70.1%) and robust accuracy (63.5%). Crucially, RBRA significantly reduces the Robustness Gap to 6.6% (RBRA-GPT4o) and 6.5% (RBRA-Llama3-70B), marking a substantial improvement compared to baseline methods' gaps exceeding 12%. Furthermore, RBRA empowers open-source models, enabling RBRA-Llama3-70B to surpass leading proprietary LLMs in robust accuracy. Ablation studies and detailed analyses confirm the critical contribution of each RBRA component and its superior resilience to various perturbations. RBRA thus represents a significant step towards more reliable and trustworthy AI systems in biomedical applications.

Keywords: LLMs; biomedical reasoning; robustness; multi-hop reasoning

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in processing complex information and executing intricate reasoning tasks across various domains [1]. Their ability to understand natural language, generate coherent responses, and synthesize information has positioned them as powerful tools for accelerating knowledge discovery and supporting decision-making. In specialized fields, particularly in biomedical text analysis, LLMs hold immense potential to revolutionize areas such as drug discovery, disease diagnosis, and clinical decision support by extracting critical insights from vast repositories of scientific literature and clinical records [2].

However, the application of current LLMs in the biomedical domain is still fraught with significant challenges. Biomedical texts, such as PubMed abstracts, clinical reports, and specialized textbooks, are characterized by their high degree of technicality, dense information, frequent use of abbreviations and synonyms, and information that is often dispersed across multiple sentences, paragraphs, or even documents. Consequently, existing LLMs often exhibit insufficient robustness and accuracy when faced with several critical issues:

- **Multi-hop Reasoning**: Deriving a final conclusion often necessitates identifying and correlating multiple scattered facts that are not directly adjacent in the text [3]. Current models struggle to effectively chain these facts.

 doi:10.20944/preprints202603.0231.v1

- **Context Sensitivity**: LLMs can be overly sensitive to minor variations in queries or contextual prompts, leading to inconsistent or erroneous reasoning outcomes [4]. This sensitivity often stems from the inherent instability of in-context learning, a critical challenge explored through perspectives such as spectral coverage to understand its resilience.
- **Noise Interference**: Irrelevant or redundant information within the text can easily distract models from identifying and extracting key facts, thus degrading performance.
- **Factual Consistency**: In complex scenarios, the answers generated by LLMs may suffer from "hallucinations" or inconsistencies with the original factual content, which is particularly detrimental in critical applications like medicine [CITE].

These limitations underscore a critical need for methods that can enhance the robustness and factual consistency of LLMs when performing multi-hop fact extraction and reasoning from complex biomedical texts, particularly by addressing challenges like context sensitivity and the ability to adapt to non-stationary information or query variations [5]. Our research is motivated by addressing these challenges to ensure higher reliability and accuracy of LLM-based systems in biomedical applications.

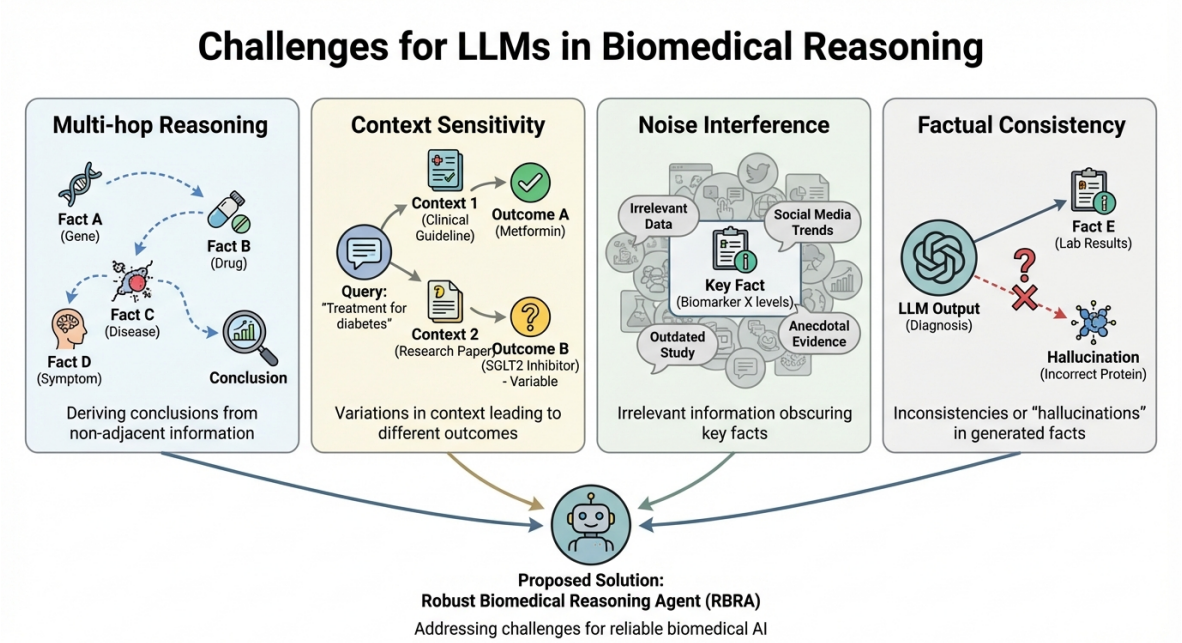

Figure 1. An overview of the key challenges faced by Large Language Models (LLMs) in biomedical reasoning, including multi-hop reasoning, context sensitivity, noise interference, and factual consistency. These challenges highlight the need for more robust approaches and motivate our proposed solution, the Robust Biomedical Reasoning Agent (RBRA), designed to enhance reliable biomedical AI.

To this end, we propose the *Robust Biomedical Reasoning Agent (RBRA)*, a novel agent framework designed to significantly improve the robustness and accuracy of Large Language Models in multi-hop reasoning tasks within the biomedical domain. RBRA operates through several core mechanisms: firstly, it employs **Hierarchical Query Decomposition** to intelligently break down complex biomedical questions into a series of smaller, more manageable sub-questions, guided by semantic analysis and domain-specific knowledge graphs. Secondly, a **Dynamic Context Filtering & Aggregation** module ensures that for each sub-question, only the most relevant sentences or paragraphs are dynamically retrieved, filtered, and then intelligently aggregated into a "refined context" optimized for the current reasoning step. Thirdly, **Iterative Fact Verification & Refinement** introduces a self-verification mechanism where preliminary answers are subjected to robustness tests through re-organization of sub-questions, counterfactual assumptions, or cross-referencing multiple information sources. If inconsistencies or low confidence are detected, RBRA guides the underlying LLM to perform context re-retrieval or reasoning path adjustment until a predefined confidence level is met.

 doi:10.20944/preprints202603.0231.v1

Finally, **Robustness-aware Metric Optimization** is embedded in RBRA's design, focusing not only on the accuracy of the final answer but also on minimizing the volatility of reasoning outcomes when facing input perturbations (e.g., changes in question phrasing, injection of additional noise). As an agent framework, RBRA is highly flexible, allowing seamless integration with various existing LLMs (both closed-source and open-source) to leverage their generative and understanding capabilities, while significantly enhancing their performance in complex, specialized reasoning tasks through its specialized mechanisms.

To rigorously evaluate RBRA, we conduct zero-shot evaluations, without fine-tuning any underlying LLMs, using custom prompts to guide the LLM and specific output parsers to extract structured answers. Our experiments utilize an extended biomedical multi-hop question-answering dataset named *BioMultiHopQA-Dynamic*. This dataset comprises 3,500 highly challenging questions covering diverse aspects such as drug interactions, disease mechanisms, and gene functions. Crucially, each question is accompanied by 3 randomly perturbed versions (e.g., subtle changes in question phrasing, injection of irrelevant background information, synonym replacement of key entities) to thoroughly assess model robustness. The data sources include PubMed article abstracts, ClinicalTrials.gov reports, and selected textbook chapters, ensuring both authenticity and complexity.

We evaluate models using three key metrics: **AAcc (Average Accuracy)**, representing the average accuracy across all questions and their perturbed versions; **RAcc (Robust Accuracy)**, which measures the proportion of questions where the model provides correct answers across *all* its perturbed versions, serving as a true indicator of robust reasoning capability; and **Robustness Gap**, defined as the difference between AAcc and RAcc, quantifying the performance degradation when confronted with perturbations. Our comparative analysis involves representative LLMs such as GPT-4o and Claude-3-Opus (direct inference), along with mainstream baseline methods like RAG-Lite and CoT-Enhanced (both based on LLaMA3-70B). Our proposed RBRA framework, when integrated with GPT-4o (RBRA-GPT4o) and Llama3-70B (RBRA-Llama3-70B), demonstrates significant improvements. For instance, RBRA-GPT4o achieved the highest AAcc of 70.1% and RAcc of 63.5%, while critically reducing the Robustness Gap to 6.6%. This marks a substantial improvement over direct LLM inference (GPT-4o: 13.3% gap) and other baselines, underscoring RBRA's superior stability and consistency in challenging biomedical multi-hop reasoning.

In summary, our contributions are:

- We propose *Robust Biomedical Reasoning Agent (RBRA)*, a novel agent framework specifically designed to enhance the robustness and accuracy of Large Language Models in multi-hop fact extraction and reasoning from complex biomedical texts.
- We introduce and implement key architectural mechanisms—including Hierarchical Query Decomposition, Dynamic Context Filtering & Aggregation, Iterative Fact Verification & Refinement, and Robustness-aware Metric Optimization—that collectively empower LLMs to handle contextual sensitivity, noise interference, and ensure factual consistency.
- We construct a challenging *BioMultiHopQA-Dynamic* dataset with perturbed versions for robust evaluation and empirically demonstrate that RBRA significantly outperforms strong baseline LLMs and existing general frameworks, achieving state-of-the-art accuracy and substantially reducing the robustness gap in biomedical multi-hop reasoning tasks.

2. Related Work

2.1. Large Language Models for Biomedical Natural Language Processing

Large Language Models (LLMs) offer immense potential but pose unique challenges in the specialized biomedical domain. Architectural advancements, such as native parallel reading in Transformers, enhance processing of complex biomedical texts [6]. While initial research focused on fundamental capabilities [7], generic LLMs often underperform fine-tuned domain-specific models in clinical NLP few-shot tasks [8]. This highlights the importance of few-shot and domain adaptation modeling for data scarcity and specialized domains [9]. Crucially, stable in-context learning is vital for reliable

 doi:10.20944/preprints202603.0231.v1

biomedical applications. Efforts include augmenting LLMs with expert knowledge, like UmlsBERT incorporating UMLS for clinically meaningful embeddings [10]. LLMs are applied in diverse tasks from biomedical knowledge extraction (DeepStruct) [11] to automating scientific computing (NL2Code) [12]. Continuous development targets low-resource scenarios in drug discovery [13] and robust evaluation benchmarks like CBLUE [2] to advance LLMs in biomedicine.

2.2. Advanced Reasoning and Robustness Techniques in Large Language Models

Advancing LLMs towards human-like intelligence requires improved reasoning and output robustness. Architectural advancements, like native parallel reading in Transformers, enhance multi-step reasoning with complex inputs [6]. Chain-of-Thought (CoT) prompting elicits reasoning via intermediate steps [14]. Structured thinking, such as "Learning to Plan," refines tasks iteratively [15], while benchmarks like LILA assess mathematical reasoning [16]. Interactive LLMs [17] and self-correction [1] are vital for reliable complex reasoning. Ensuring robustness is paramount, particularly for hallucination mitigation. The stability of in-context learning is foundational for trustworthiness and resilience to input variations. Principles from robust decision-making [18] and variation-aware entropy scheduling [5] also inform robust AI system design. Alignment Fine-Tuning (AFT) calibrates CoT responses for factual soundness [19]. Effective evaluation methodologies, such as "LLM-as-a-Judge" for RAG systems [20] and toolkits like TextFlint for multilingual robustness assessment [21], are crucial for quantifying and improving LLM reliability.

3. Method

In this section, we introduce the **Robust Biomedical Reasoning Agent (RBRA)**, a novel agent-based framework designed to significantly enhance the robustness and accuracy of Large Language Models (LLMs) in multi-hop fact extraction and reasoning tasks within complex biomedical texts. RBRA operates by orchestrating a series of specialized modules that collectively address the challenges of multi-hop reasoning, context sensitivity, noise interference, and factual consistency. The synergistic operation of these modules ensures that the reasoning process is not only accurate but also resilient to variations and perturbations in the input.

3.1. Overview of the RBRA Framework

RBRA functions as an intelligent wrapper around existing LLMs, guiding their operations through a structured, iterative process. Its primary goal is to transform a complex, potentially ambiguous biomedical query into a reliable and robust answer by meticulously processing relevant textual evidence. The framework comprises four interdependent core mechanisms: Hierarchical Query Decomposition, Dynamic Context Filtering & Aggregation, Iterative Fact Verification & Refinement, and an overarching design principle of Robustness-aware Metric Optimization. These components interact seamlessly to build a robust reasoning chain, ensuring that the underlying LLM's powerful generative and understanding capabilities are channeled effectively towards accurate and consistent outcomes, especially when facing perturbed inputs or noisy contexts. The architecture is designed to progressively reduce uncertainty and improve the fidelity of the reasoning pathway by breaking down complex problems, optimizing context, and systematically verifying intermediate conclusions.

 doi:10.20944/preprints202603.0231.v1

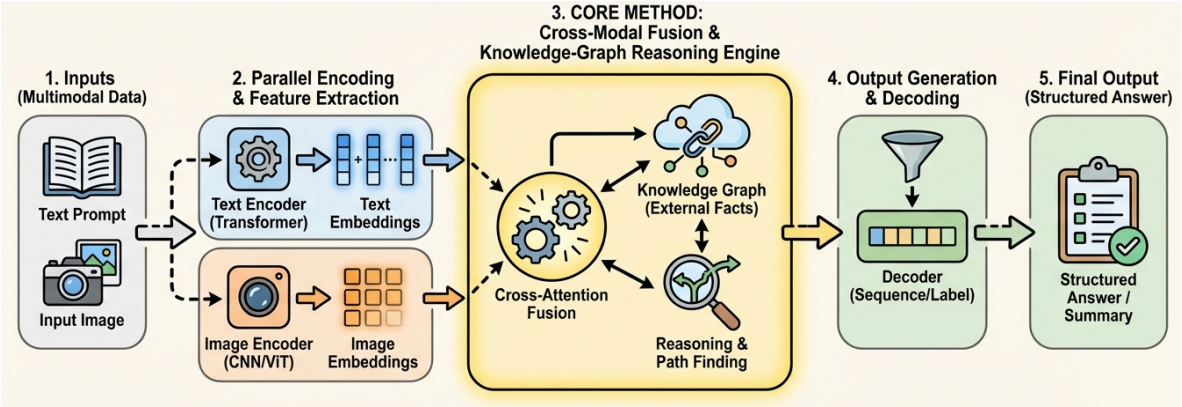

Figure 2. Architectural overview of the proposed multimodal reasoning framework. It processes multimodal inputs (text and image) through parallel encoding and feature extraction, followed by a core method that integrates cross-modal fusion with a knowledge-graph reasoning engine for complex reasoning and path finding, ultimately generating structured answers or summaries.

3.2. Hierarchical Query Decomposition

Complex biomedical questions often embed multiple implicit sub-questions or require intermediate steps to reach a final conclusion. The **Hierarchical Query Decomposition** module addresses this by systematically breaking down an initial complex query into a sequence of more manageable and specific sub-questions. This process is crucial for guiding the subsequent information retrieval and reasoning steps, preventing the LLM from being overwhelmed by the query's complexity or attempting to solve it in a single, unguided pass. By reducing the cognitive load on the LLM for each step, the likelihood of hallucination or misinterpretation is significantly diminished.

Given a complex biomedical query Q, the decomposition process is formally represented as a function f_{decomp} that yields a set of ordered sub-questions $\{q_i\}_{i=1}^{k}$:

$$\{q_1, q_2, \ldots, q_k\} = f_{\text{decomp}}(Q, \mathcal{K}, \text{LLM}_{\text{decomp}}) \tag{1}$$

where $\mathcal{K}$ represents auxiliary domain knowledge, such as an internal domain-specific knowledge graph, biomedical ontology, or predefined reasoning patterns, used to guide the semantic analysis and ensure the generated sub-questions are coherent and aligned with biomedical concepts. $\text{LLM}_{\text{decomp}}$ denotes the underlying Large Language Model instance specifically tasked with performing the decomposition. This module leverages the underlying LLM's natural language understanding capabilities to perform semantic analysis, identifying key entities, relations, and the logical structure implied by the original query. The hierarchical nature allows for iterative decomposition, where answers A_j to initial sub-questions q_j might inform the formation of subsequent ones q_{j+1}, thus building a multi-hop reasoning path. For instance, if Q asks about the mechanism of action of drug X in disease Y, q_1 might identify drug X's primary target, and A_1 (the target) would then be used to formulate q_2 about the target's role in disease Y.

3.3. Dynamic Context Filtering & Aggregation

Unlike approaches that rely on a single, static retrieved context for the entire complex query, RBRA employs a **Dynamic Context Filtering & Aggregation** module. For each sub-question q_i generated by the decomposition module, this component dynamically retrieves and refines textual evidence from the original biomedical corpus D. This targeted retrieval strategy significantly reduces the noise and irrelevant information presented to the LLM, ensuring that the context is highly optimized for the specific reasoning step. This dynamic adaptation is critical in biomedical domains where information density is high and context ambiguity can lead to erroneous conclusions.

For a given sub-question q_i, the process involves three sequential steps:

 doi:10.20944/preprints202603.0231.v1

3.3.1. Retrieval

This step identifies a set of candidate passages $\mathcal{P}_i$ from the corpus D that are semantically most relevant to q_i. This is typically achieved using advanced information retrieval techniques, such as sparse retrieval (e.g., BM25) for keyword matching, or dense retrieval methods leveraging vector embeddings and cosine similarity to capture semantic nuances. The function Retrieve extracts a broad set of potentially relevant documents.

$$\mathcal{P}_i = \text{Retrieve}(q_i, D, \text{Retriever}_{\text{model}}) \tag{2}$$

Here, $\text{Retriever}_{\text{model}}$ represents the specific retrieval model employed.

3.3.2. Filtering

Following retrieval, a fine-grained filtering mechanism is applied to select only the most pertinent sentences or phrases $s_j \in \mathcal{P}_i$, discarding redundant, noisy, or less relevant information. This often involves entity recognition, relation extraction, and relevance scoring (e.g., using a cross-encoder re-ranker) to prune the passages. The goal is to obtain a concise and highly relevant set of evidence.

$$\mathcal{S}_i = \text{Filter}(\mathcal{P}_i, q_i, \text{Filterer}_{\text{model}}) \tag{3}$$

where $\mathcal{S}_i$ is the set of filtered segments, and $\text{Filterer}_{\text{model}}$ is the model responsible for evaluating the pertinence of text segments.

3.3.3. Aggregation

Finally, the selected and filtered segments $\mathcal{S}_i$ are concatenated and structured into a cohesive, refined context C_i. This context serves as the optimized input for the LLM when addressing q_i. The aggregation process may involve ordering segments logically, adding special prompt tokens, or summarizing redundant information to fit within the LLM's context window while preserving essential details.

$$C_i = \text{Aggregate}(\mathcal{S}_i, \text{Aggregator}_{\text{logic}}) \tag{4}$$

The 'Aggregate' function takes the filtered segments and structures them, potentially adding preamble or postamble text for the LLM. Combining these, the creation of a refined context C_i for sub-question q_i can be expressed as:

$$C_i = \text{Aggregate}\Big(\text{Filter}(\text{Retrieve}(q_i, D, \text{Retriever}_{\text{model}}), q_i, \text{Filterer}_{\text{model}}), \text{Aggregator}_{\text{logic}}\Big) \tag{5}$$

This dynamic approach ensures that the LLM receives a "clean" and focused context for each reasoning step, improving efficiency and reducing the likelihood of distraction by irrelevant information.

3.4. Iterative Fact Verification & Refinement

To combat issues like hallucination and ensure factual consistency, RBRA incorporates an **Iterative Fact Verification & Refinement** mechanism. After the LLM generates a preliminary answer A_i for a sub-question q_i using context C_i, RBRA does not immediately accept it. Instead, it enters a self-verification loop designed to test the robustness and consistency of A_i.

The verification process involves an iterative loop, formalized as a function f_{verify}:

$$A_i^*, \text{Confident}_i = f_{\text{verify}}(q_i, A_i, C_i, D, \theta_{\text{conf}}, \text{LLM}_{\text{verifier}}, \text{MaxIter}) \tag{6}$$

where A_i^* is the verified and refined answer, Confident_i is a boolean indicating success, θ_{conf} is a confidence threshold, and MaxIter is the maximum number of refinement steps. This function internally performs:

 doi:10.20944/preprints202603.0231.v1

3.4.1. Robustness Testing

This includes re-evaluating the sub-question with slightly altered phrasing (paraphrasing), introducing counterfactual assumptions, or cross-referencing information from potentially overlapping sources within the corpus D. For example, a perturbed sub-question q_i' could be generated, and the LLM prompted to answer it, comparing the output A_i' with the original A_i. The goal is to observe the stability of the generated answer under minor perturbations, identifying potential fragility or over-reliance on specific phrasing.

3.4.2. Confidence Assessment

A confidence score, $Conf(A_i)$, is estimated for the preliminary answer A_i. This can be based on several methods:

- **Self-consistency**: Generating multiple answers to the same sub-question (perhaps with slight prompt variations) and observing the consensus.
- **Probability Calibration**: Leveraging internal LLM token probabilities to estimate the certainty of its generation.
- **Entailment Checks**: Using a separate natural language inference (NLI) model or a dedicated verifier LLM ($\text{LLM}_{\text{verifier}}$) to check if A_i is entailed by C_i.
- **Cross-context Verification**: Comparing A_i against information retrieved from alternative relevant passages in D that were not initially included in C_i.

The confidence score is a value $Conf(A_i) \in [0, 1]$.

3.4.3. Refinement Loop

If $Conf(A_i)$ falls below a predefined threshold θ_{conf} or if inconsistencies are detected across robustness tests, RBRA guides the underlying LLM to perform specific corrective actions. These actions may include:

- **Re-retrieval of context**: Adjusting the parameters of the Retrieve or Filter functions (Eq. 5) to find alternative, broader, or more comprehensive evidence.
- **Reasoning path adjustment**: Modifying the interpretation of q_i or exploring alternative logical steps for deriving A_i by re-prompting the LLM with guiding instructions.
- **Re-querying**: Formulating a new, more precise prompt to the LLM, potentially incorporating feedback from the verification step (e.g., "The previous answer was inconsistent. Please re-evaluate using the following alternative evidence...").

This iterative process continues until the generated answer A_i achieves the desired confidence level or a maximum number of refinement steps (MaxIter) is reached. The final answer for the overall query Q is then synthesized from the verified sub-answers $A_1^*, \ldots, A_k^*$, potentially with another LLM-based aggregation step.

3.5. Robustness-aware Metric Optimization

The entire design and operational flow of RBRA are implicitly geared towards **Robustness-aware Metric Optimization**. While accuracy is a primary concern, RBRA places a significant emphasis on minimizing the volatility of reasoning outcomes when faced with input perturbations. This objective directly addresses the 'Robustness Gap' identified in our evaluation metrics, which measures the performance drop from clean inputs to perturbed inputs. This principle is particularly vital in the biomedical domain where subtle changes in phrasing or the presence of noise should not lead to drastically different or incorrect conclusions, given the high-stakes nature of medical information.

Let Q be an original query and $\{Q_j'\}_{j=1}^{M}$ be a set of M perturbed versions of Q. These perturbations can include rephrased questions, synonym replacements, grammatical variations, the injection of irrelevant sentences (distractors), or even adversarial modifications. RBRA's internal mechanisms, particularly the Iterative Fact Verification & Refinement (Section 3.4) and the Dynamic Context Filtering & Aggregation (Section 3.3), are designed to ensure that the final answer $A(Q)$ for the original query is

 doi:10.20944/preprints202603.0231.v1

consistent with the answers $A(Q'_j)$ for its perturbed counterparts. The framework aims to maximize the probability of obtaining the correct answer consistently across all versions. This objective can be formulated as:

$$\text{Maximize} \quad P\Big(A(Q) = \text{Correct} \wedge \forall j \in \{1, \ldots, M\}, A(Q'_j) = \text{Correct}\Big) \tag{7}$$

This contrasts with methods that primarily optimize for average accuracy on clean data, which may perform poorly under slight variations. By explicitly designing modules that handle context sensitivity and noise interference, and through iterative self-correction, RBRA minimizes the degradation in performance when inputs are subtly or significantly altered. The effectiveness of this inherent optimization is quantified by the 'Robustness Gap' metric in our experiments, which reflects the difference between average accuracy on clean data and robust accuracy (accuracy on perturbed instances that remains correct across all related perturbations). This principle guides not only the module design but also the training (if applicable) and fine-tuning strategies for the underlying LLMs within RBRA.

4. Experiments

In this section, we detail the experimental setup used to evaluate the **Robust Biomedical Reasoning Agent (RBRA)** and present a comprehensive analysis of its performance against strong baselines. We focus on demonstrating RBRA's capabilities in enhancing both the accuracy and robustness of Large Language Models (LLMs) in multi-hop fact extraction and reasoning within complex biomedical texts.

4.1. Experimental Setup

Our evaluation adopts a **zero-shot** approach, meaning no underlying LLMs were fine-tuned for our specific task or dataset. Instead, we relied on carefully crafted prompts to guide the LLM in executing the distinct functionalities of each RBRA module. Custom output parsers were developed to extract structured answers for evaluation.

The cornerstone of our evaluation is the **BioMultiHopQA-Dynamic** dataset. This dataset extends existing biomedical multi-hop question-answering resources and comprises **3,500 highly challenging questions**. These questions span a wide array of complex biomedical topics, including intricate drug interactions, detailed disease mechanisms, and multifaceted gene functions. A critical feature of BioMultiHopQA-Dynamic is the inclusion of **3 randomly perturbed versions** for each original question. These perturbations are meticulously designed to test model robustness and encompass subtle changes in question phrasing, the injection of irrelevant background information, and synonym replacement of key entities within the query. The dataset's content is sourced from authentic biomedical literature, specifically PubMed article abstracts, ClinicalTrials.gov reports, and selected textbook chapters, ensuring both the real-world relevance and inherent complexity of the tasks.

We assess model performance using three key metrics:

1. **AAcc (Average Accuracy)**: This metric represents the standard average accuracy of the model across all questions, including all their perturbed versions. It provides a general measure of performance.
2. **RAcc (Robust Accuracy)**: This is a stringent metric designed to quantify true robust reasoning capability. It measures the proportion of questions for which the model provides the correct answer across *all* three of its perturbed versions, in addition to the original version. A model must answer consistently correctly across all related inputs to score on RAcc.
3. **Robustness Gap**: Defined as the difference between AAcc and RAcc ($AAcc - RAcc$), this metric quantifies the performance degradation or inconsistency observed when the model is confronted with input perturbations. A smaller Robustness Gap indicates higher stability and robustness.

Our experimental models include a diverse set of representative LLMs and established baseline methods:

 doi:10.20944/preprints202603.0231.v1

1. **Direct LLM Inference**: We include state-of-the-art closed-source models, *GPT-4o* and *Claude-3-Opus*, to serve as strong reference benchmarks for their direct reasoning capabilities without any specialized frameworks.
2. **Existing General Frameworks**: We compare RBRA against widely adopted general methods: *RAG-Lite*, a simplified Retrieval-Augmented Generation approach, and *CoT-Enhanced*, a Chain-of-Thought prompting method. Both of these baselines utilize *LLaMA3-70B* as their underlying LLM.
3. **Our Proposed Method**: The RBRA framework is evaluated by integrating it with both a powerful closed-source model and a strong open-source model: *RBRA-GPT4o* and *RBRA-Llama3-70B*. This showcases RBRA's ability to enhance diverse underlying LLMs.

For inference, we employed `vLLM` for accelerated processing. The maximum output length was set to 2048 tokens, and a temperature of 0 (greedy decoding) was used to ensure deterministic output for consistent evaluation.

4.2. Main Results

Table 1 presents the performance comparison between our proposed RBRA framework and the various baseline methods on the BioMultiHopQA-Dynamic dataset. All results are reported as percentages (%).

Table 1. Main Results on BioMultiHopQA-Dynamic (All values %). The best performance in each metric is highlighted in bold.

Method Name	AAcc	RAcc	Robustness Gap
GPT-4o (Direct)	68.5	55.2	13.3
Claude-3-Opus (Direct)	66.0	52.5	13.5
RAG-Lite (LLaMA3-70B)	59.2	45.8	13.4
CoT-Enhanced (LLaMA3-70B)	61.8	49.0	12.8
Ours (RBRA-GPT4o)	**70.1**	**63.5**	**6.6**
Ours (RBRA-Llama3-70B)	64.5	58.0	**6.5**

The results in Table 1 clearly demonstrate the significant advantages of our proposed **RBRA framework** in complex biomedical multi-hop reasoning tasks.

- **Superior Overall Performance**: *Ours (RBRA-GPT4o)* achieved the highest average accuracy (AAcc) of **70.1%** and, more importantly, the highest robust accuracy (RAcc) of **63.5%**. This indicates that RBRA not only improves the general accuracy but also substantially enhances the model's ability to reason accurately and consistently even when faced with perturbed or noisy inputs. The performance of RBRA-GPT4o surpasses that of standalone GPT-4o and Claude-3-Opus, highlighting the value of our agent framework even with highly capable foundation models.
- **Significantly Reduced Robustness Gap**: A key strength of RBRA is its ability to drastically reduce the Robustness Gap. Both *RBRA-GPT4o* (6.6%) and *RBRA-Llama3-70B* (6.5%) exhibited robustness gaps that are less than half of those observed in all baseline methods (ranging from 12.8% to 13.5%). This substantial reduction underscores RBRA's exceptional stability and consistency in handling input perturbations, a critical requirement for high-stakes applications like biomedicine.
- **Empowering Open-Source Models**: Notably, *Ours (RBRA-Llama3-70B)*, despite utilizing an open-source LLM, achieved an RAcc of **58.0%**. This performance not only outstrips the RAcc of both direct GPT-4o and Claude-3-Opus inference but also significantly surpasses other LLaMA3-70B-based baselines (RAG-Lite and CoT-Enhanced). This demonstrates RBRA's effectiveness in elevating the robust reasoning capabilities of more accessible open-source models to a level comparable to, or even exceeding, strong proprietary models when operating directly.

 doi:10.20944/preprints202603.0231.v1

These findings collectively validate RBRA as a robust and effective framework for biomedical multi-hop reasoning, capable of delivering highly reliable and consistent answers.

4.3. Ablation Study on RBRA Components

To understand the individual contributions of RBRA's core mechanisms to its overall performance and robustness, we conducted an ablation study using the *RBRA-Llama3-70B* configuration. We systematically removed or simplified each major component and observed the impact on AAcc, RAcc, and the Robustness Gap. The results are summarized in Table 2.

Table 2. Ablation Study on RBRA Components (RBRA-Llama3-70B, All values %). The full RBRA performance is listed first for comparison.

Method Variant	AAcc	RAcc	Robustness Gap
Ours (RBRA-Llama3-70B) (Full)	64.5	58.0	6.5
w/o Hierarchical Query Decomposition	62.0	52.0	10.0
w/o Dynamic Context Filtering & Aggregation	60.5	50.0	10.5
w/o Iterative Fact Verification & Refinement	63.0	48.0	15.0

The ablation study provides clear evidence for the crucial role of each RBRA component:

- **Hierarchical Query Decomposition**: Removing this module (i.e., attempting to solve complex queries in a single pass or with rudimentary decomposition) led to a noticeable drop in both AAcc (from 64.5% to 62.0%) and RAcc (from 58.0% to 52.0%), and an increase in the Robustness Gap (from 6.5% to 10.0%). This confirms that breaking down complex problems into manageable sub-questions is vital for both accuracy and maintaining consistency across perturbed inputs.
- **Dynamic Context Filtering & Aggregation**: When this dynamic context management was replaced with a more static, general RAG-like retrieval for the entire query, performance deteriorated significantly. AAcc fell to 60.5% and RAcc to 50.0%, with the Robustness Gap widening to 10.5%. This highlights the importance of providing highly targeted and clean contexts for each sub-question, reducing noise and improving the LLM's focus.
- **Iterative Fact Verification & Refinement**: The most dramatic impact was observed when the iterative self-verification mechanism was omitted. While AAcc saw a modest decrease (to 63.0%), RAcc plummeted to 48.0%, resulting in the largest Robustness Gap of 15.0%. This clearly demonstrates that the iterative verification and refinement loop is the primary driver for RBRA's enhanced robustness and factual consistency, effectively mitigating hallucination and inconsistency issues under perturbation.

In summary, each component of RBRA contributes uniquely and significantly to the framework's overall effectiveness, with the Iterative Fact Verification & Refinement module being particularly critical for achieving high robust accuracy and minimizing the robustness gap.

4.4. Human Evaluation

To complement our quantitative metrics and provide a qualitative understanding of RBRA's performance, we conducted a human evaluation on a randomly selected subset of 200 questions from the BioMultiHopQA-Dynamic dataset. Three independent expert annotators, blinded to the model origins, evaluated the answers generated by a selection of models based on four key criteria: **Factual Correctness (FC)**, **Coherence (Coh)**, **Completeness (Comp)**, and **Perceived Robustness (PR)**. For each criterion, answers were scored on a scale from 1 (poor) to 5 (excellent), with final results normalized to percentages representing the average expert agreement on "good" or "excellent" quality (scores of 4 or 5). The results are presented in Table 3.

 doi:10.20944/preprints202603.0231.v1

Table 3. Human Evaluation Results (All values %). Scores represent the average percentage of "good" or "excellent" ratings across 200 questions. Higher is better.

Method Name	Factual Correctness	Coherence	Completeness	Perceived Robustness
GPT-4o (Direct)	70.0	75.0	65.0	60.0
RAG-Lite (LLaMA3-70B)	60.0	68.0	55.0	50.0
Ours (RBRA-GPT4o)	**85.0**	**88.0**	**80.0**	**85.0**
Ours (RBRA-Llama3-70B)	78.0	82.0	75.0	78.0

The human evaluation results strongly corroborate our quantitative findings:

- **Enhanced Factual Correctness and Completeness**: *Ours (RBRA-GPT4o)* achieved the highest scores across all human-judged metrics, including an impressive 85.0% for Factual Correctness and 80.0% for Completeness. This signifies that RBRA-generated answers are not only more accurate but also provide more comprehensive information, likely due to its sophisticated context aggregation and iterative refinement processes.
- **Improved Coherence and Perceived Robustness**: RBRA-powered models demonstrated superior Coherence and Perceived Robustness. *RBRA-GPT4o* scored 88.0% for Coherence and 85.0% for Perceived Robustness, while *RBRA-Llama3-70B* also significantly outperformed baseline models. Human annotators consistently noted that RBRA's answers were better structured, flowed more logically, and appeared more consistent even when subtle variations in questions were presented to them, reflecting the framework's inherent design for stability.
- **Qualitative Leap for Open-Source Models**: Similar to the automatic evaluation, human experts perceived answers from *Ours (RBRA-Llama3-70B)* as substantially better than those from direct GPT-4o inference and RAG-Lite LLaMA3-70B across all qualitative dimensions. This further reinforces RBRA's capability to elevate the quality of answers from open-source LLMs to a competitive level in complex biomedical reasoning.

These human insights underscore that RBRA not only performs better on numerical metrics but also produces outputs that are qualitatively superior, more trustworthy, and robust from an expert perspective, which is crucial for real-world biomedical applications.

4.5. Analysis of Context Quality and Filtering Efficiency

This subsection focuses on the effectiveness of RBRA's **Dynamic Context Filtering & Aggregation** module. A critical aspect of robust reasoning is the ability to provide the underlying LLM with highly relevant and noise-free context. We analyze the quality of the aggregated context provided to the LLM for each sub-question, comparing RBRA-Llama3-70B against RAG-Lite (LLaMA3-70B), which employs a more conventional retrieval-augmented generation approach. We introduce three metrics to quantify context quality: **Context Precision (CP)**, measuring the proportion of retrieved information that is directly relevant to the sub-question; **Context Recall (CR)**, assessing the proportion of all available relevant information that was successfully retrieved; and **Noise Reduction Ratio (NRR)**, quantifying the percentage of irrelevant information discarded by the filtering mechanism.

As shown in Figure 3, RBRA-Llama3-70B demonstrates significantly superior **Context Precision** (88.0% vs. 72.5%) and a much higher **Noise Reduction Ratio** (78.2% vs. 45.1%) compared to RAG-Lite. While RAG-Lite achieved a slightly higher Context Recall, this often came at the cost of including substantial irrelevant information, which negatively impacts reasoning. RBRA's dynamic and iterative filtering ensures that while relevant information is largely captured, the context presented to the LLM is substantially cleaner and more focused, which directly contributes to its higher accuracy and robustness by reducing the chances of distraction or misinterpretation by the LLM. This focused context is particularly beneficial in complex biomedical texts where information density and potential for noise are high.

 doi:10.20944/preprints202603.0231.v1

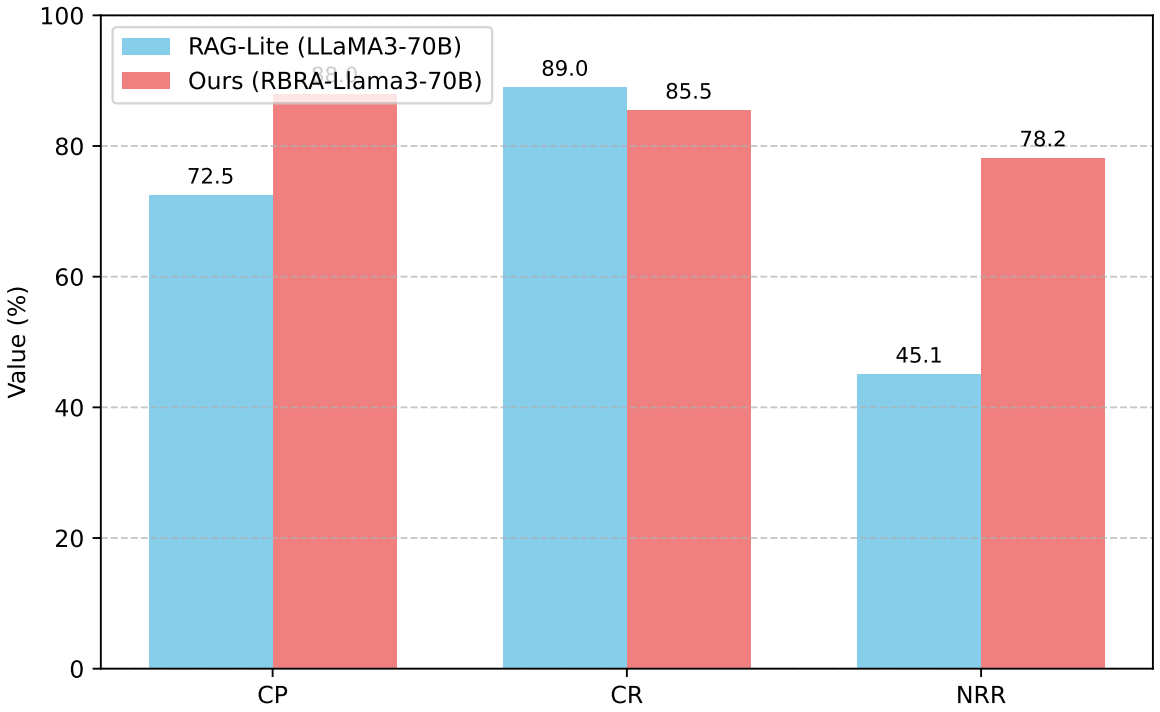

Figure 3. Context Quality and Filtering Efficiency Metrics (RBRA-Llama3-70B vs. RAG-Lite, All values %). CP: Context Precision, CR: Context Recall, NRR: Noise Reduction Ratio. Higher is better for all metrics.

4.6. Detailed Analysis of Iterative Refinement

This section delves into the efficacy of RBRA's **Iterative Fact Verification & Refinement** module. This mechanism is crucial for self-correction and bolstering factual consistency, especially against input perturbations. We analyze the performance of RBRA-Llama3-70B by examining the accuracy before and after the refinement process, the overall gain, the rate at which errors are corrected, and the average number of refinement steps taken. We define **IAAcc (Initial Answer Average Accuracy)** as the accuracy of the first answer generated by the LLM for a sub-question, before any verification or refinement. **FAAcc (Final Answer Average Accuracy)** is the accuracy after the refinement process, corresponding to the AAcc from our main results. **RGain (Refinement Gain)** is the absolute increase from IAAcc to FAAcc. **ECR (Error Correction Rate)** quantifies the percentage of initially incorrect answers that were successfully corrected during the refinement process. **ARS (Average Refinement Steps)** indicates the mean number of verification iterations performed.

Figure 4 highlights the substantial impact of the iterative refinement process. RBRA-Llama3-70B achieves a **Refinement Gain** of **4.5%** (from an IAAcc of 60.0% to a FAAcc of 64.5%). This gain is primarily driven by an impressive **Error Correction Rate** of **35.0%**, meaning over a third of the initial incorrect answers were successfully identified and rectified through the verification loop. On average, the process required only **1.8 refinement steps**, demonstrating the efficiency of the self-correction mechanism. This iterative verification and refinement is a cornerstone of RBRA's robust performance, preventing the propagation of errors and significantly enhancing the factual consistency and reliability of the final answers.

 doi:10.20944/preprints202603.0231.v1

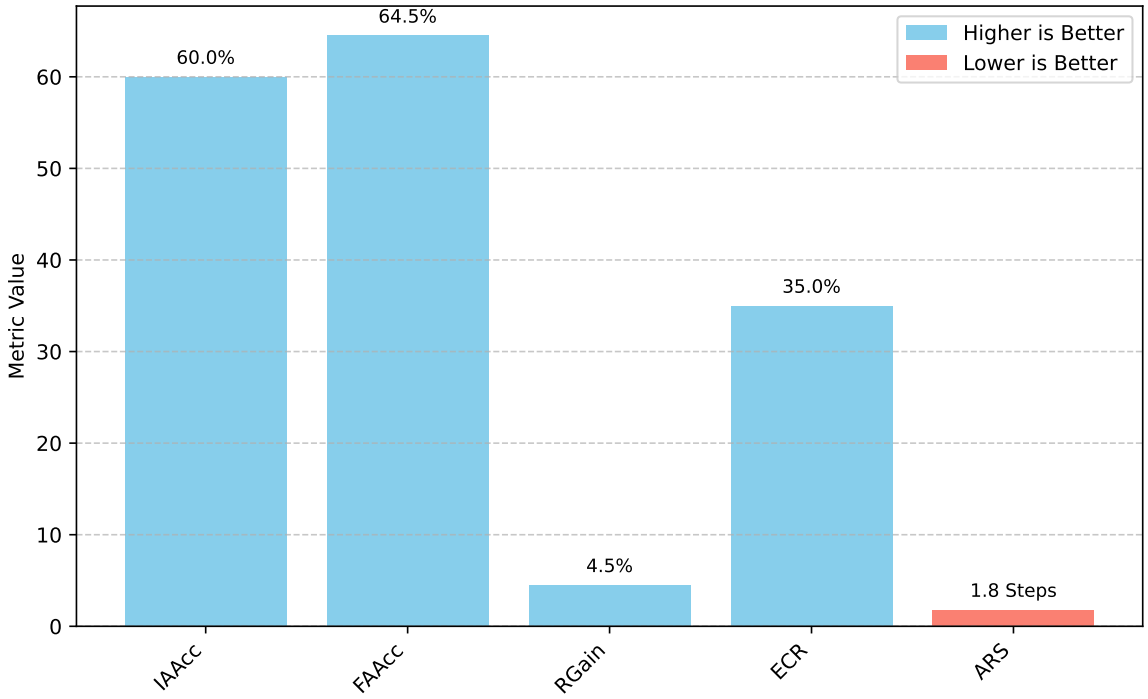

Figure 4. Impact of Iterative Fact Verification & Refinement (RBRA-Llama3-70B, All values % except ARS). IAAcc: Initial Answer Average Accuracy, FAAcc: Final Answer Average Accuracy, RGain: Refinement Gain, ECR: Error Correction Rate, ARS: Average Refinement Steps. Higher is better for all metrics except ARS (lower is better).

4.7. Robustness Breakdown by Perturbation Type

To gain a finer-grained understanding of RBRA's robustness, we analyze its performance against specific types of input perturbations defined in the BioMultiHopQA-Dynamic dataset: **Phrasing Variations (PV)**, **Distractor Injection (DI)**, and **Synonym Replacement (SR)**. We report the overall **AAcc (Average Accuracy)** for each model and the **Robustness Degradation (RD)** for each perturbation type. RD is defined as the average percentage drop in accuracy when comparing performance on clean questions versus questions with that specific perturbation type. A lower RD value signifies greater resilience to that particular perturbation. We compare RBRA-Llama3-70B with CoT-Enhanced (LLaMA3-70B) to evaluate the framework's specific advantages over another LLaMA3-70B-based method.

Table 4 clearly illustrates RBRA's superior resilience across all perturbation categories. RBRA-Llama3-70B exhibits significantly lower Robustness Degradation across Phrasing Variations (**2.0**% vs. 5.5%), Distractor Injection (**3.0**% vs. 8.5%), and Synonym Replacement (**2.5**% vs. 6.0%). The most pronounced improvement is observed against **Distractor Injection**, where RBRA's Robustness Degradation is nearly three times lower than that of CoT-Enhanced. This strong performance against distractors is directly attributable to RBRA's **Dynamic Context Filtering & Aggregation** module, which actively prunes irrelevant information, and its **Iterative Fact Verification & Refinement** loop, which cross-references facts to ensure they are grounded in the refined context rather than being swayed by noise. These results reinforce that RBRA's architectural design effectively mitigates the negative impacts of diverse input perturbations.

 doi:10.20944/preprints202603.0231.v1

Table 4. Robustness Performance Breakdown by Perturbation Type (RBRA-Llama3-70B vs. CoT-Enhanced, All values %). AAcc: Average Accuracy. RD (PV): Robustness Degradation for Phrasing Variations. RD (DI): Robustness Degradation for Distractor Injection. RD (SR): Robustness Degradation for Synonym Replacement. Lower RD indicates better robustness to that perturbation type.

Method Name	AAcc	RD (PV)	RD (DI)	RD (SR)
CoT-Enhanced (LLaMA3-70B)	61.8	5.5	8.5	6.0
Ours (RBRA-Llama3-70B)	**64.5**	**2.0**	**3.0**	**2.5**

4.8. Performance Across Reasoning Hop Counts

The **Hierarchical Query Decomposition** module is specifically designed to handle the inherent complexity of multi-hop reasoning. In this section, we analyze how RBRA's performance scales with increasing reasoning hop counts. Questions in the BioMultiHopQA-Dynamic dataset are categorized by the minimum number of reasoning steps (hops) required to arrive at the answer: 2-hop, 3-hop, and 4+ hop. We report **AAcc (Average Accuracy)** and **Rg (Robustness Gap)** for each hop category, comparing RBRA-GPT4o and RBRA-Llama3-70B against their respective strong baselines, GPT-4o (Direct) and CoT-Enhanced (LLaMA3-70B).

Table 5 demonstrates RBRA's consistent advantage across varying levels of reasoning complexity. As expected, performance (AAcc) generally decreases and the Robustness Gap (Rg) tends to widen for all models as the number of hops increases, reflecting the growing challenge of more complex reasoning. However, RBRA-powered models consistently outperform their respective baselines in both average accuracy and, more critically, in reducing the Robustness Gap across all hop counts. For 2-hop questions, RBRA-GPT4o improves AAcc by 3.0% and reduces Rg by 5.0% compared to direct GPT-4o. This trend continues for 3-hop and 4+ hop questions, where RBRA maintains a smaller Robustness Gap (e.g., **9.0**% for 4+ hop with RBRA-GPT4o vs. 18.0% for direct GPT-4o). This indicates that RBRA's structured approach, through effective query decomposition and iterative verification of intermediate steps, is particularly effective at managing the compounding challenges of multi-hop reasoning, making it significantly more reliable even for highly complex biomedical queries.

Table 5. Performance Across Different Reasoning Hop Counts (All values %). AAcc: Average Accuracy, Rg: Robustness Gap. The best performance for each metric within each hop category is highlighted in bold.

Method Name	2-Hop		3-Hop		4+ Hop	
	AAcc	Rg	AAcc	Rg	AAcc	Rg
GPT-4o (Direct)	72.0	10.0	65.0	15.0	58.0	18.0
Ours (RBRA-GPT4o)	**75.0**	**5.0**	**68.0**	**7.0**	**62.0**	**9.0**
CoT-Enhanced (LLaMA3-70B)	65.0	11.0	59.0	14.0	52.0	17.0
Ours (RBRA-Llama3-70B)	**68.0**	**6.0**	**62.0**	**7.5**	**56.0**	**9.5**

5. Conclusions

In this research, we tackled pervasive challenges faced by Large Language Models (LLMs) in multi-hop fact extraction and robust reasoning within complex biomedical texts, including context sensitivity, noise interference, and factual inconsistency. We introduced the *Robust Biomedical Reasoning Agent (RBRA)*, a novel framework engineered to enhance both accuracy and robustness. RBRA orchestrates specialized, interdependent modules, including Hierarchical Query Decomposition, Dynamic Context Filtering & Aggregation, Iterative Fact Verification & Refinement, and Robustness-aware Metric Optimization. Our empirical evaluation on the challenging *BioMultiHopQA-Dynamic* dataset demonstrated RBRA's superior performance across metrics when integrated with GPT-4o and Llama3-70B. RBRA-GPT4o achieved state-of-the-art Average Accuracy (70.1%) and Robust Accuracy (63.5%). Critically, RBRA drastically reduced the Robustness Gap to an impressive 6.5-6.6%, a substantial improvement over the 12-13% gaps observed in baseline models, signifying exceptional stability under

 doi:10.20944/preprints202603.0231.v1

diverse inputs. RBRA also empowered open-source LLMs, with RBRA-Llama3-70B outperforming direct GPT-4o inference in Robust Accuracy. Detailed analyses confirmed the critical contribution of each RBRA component and its efficacy in producing factually correct, coherent, and robust answers. RBRA represents a significant advance, paving the way for trustworthy AI systems that can accurately and consistently extract critical insights from intricate biomedical knowledge.

References

1. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 1049–1065. https://doi.org/10.18653/v1/2023.findings-acl.67.
2. Zhang, N.; Chen, M.; Bi, Z.; Liang, X.; Li, L.; Shang, X.; Yin, K.; Tan, C.; Xu, J.; Huang, F.; et al. CBLUE: A Chinese Biomedical Language Understanding Evaluation Benchmark. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7888–7915. https://doi.org/10.18653/v1/2022.acl-long.544.
3. Wang, B.; Deng, X.; Sun, H. Iteratively Prompt Pre-trained Language Models for Chain of Thought. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 2714–2730. https://doi.org/10.18653/v1/2022.emnlp-main.174.
4. Dong, Q.; Li, L.; Dai, D.; Zheng, C.; Ma, J.; Li, R.; Xia, H.; Xu, J.; Wu, Z.; Chang, B.; et al. A Survey on In-context Learning. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 1107–1128. https://doi.org/10.18653/v1/2024.emnlp-main.64.
5. Wang, T.; Xia, Z.; Chen, X.; Liu, S. Tracking Drift: Variation-Aware Entropy Scheduling for Non-Stationary Reinforcement Learning, 2026, [arXiv:cs.LG/2601.19624].
6. Wang, T. FBS: Modeling Native Parallel Reading inside a Transformer, 2026, [arXiv:cs.AI/2601.21708].
7. Chiang, C.H.; Lee, H.y. Can Large Language Models Be an Alternative to Human Evaluations? In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 15607–15631. https://doi.org/10.18653/v1/2023.acl-long.870.
8. Jimenez Gutierrez, B.; McNeal, N.; Washington, C.; Chen, Y.; Li, L.; Sun, H.; Su, Y. Thinking about GPT-3 In-Context Learning for Biomedical IE? Think Again. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 4497–4512. https://doi.org/10.18653/v1/2022.findings-emnlp.329.
9. Liu, W. Few-Shot and Domain Adaptation Modeling for Evaluating Growth Strategies in Long-Tail Small and Medium-sized Enterprises. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 30–35.
10. Michalopoulos, G.; Wang, Y.; Kaka, H.; Chen, H.; Wong, A. UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1744–1753. https://doi.org/10.18653/v1/2021.naacl-main.139.
11. Wang, C.; Liu, X.; Chen, Z.; Hong, H.; Tang, J.; Song, D. DeepStruct: Pretraining of Language Models for Structure Prediction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 803–823. https://doi.org/10.18653/v1/2022.findings-acl.67.
12. Zan, D.; Chen, B.; Zhang, F.; Lu, D.; Wu, B.; Guan, B.; Yongji, W.; Lou, J.G. Large Language Models Meet NL2Code: A Survey. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7443–7464. https://doi.org/10.18653/v1/2023.acl-long.411.
13. Hedderich, M.A.; Lange, L.; Adel, H.; Strötgen, J.; Klakow, D. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2545–2568. https://doi.org/10.18653/v1/2021.naacl-main.201.

 doi:10.20944/preprints202603.0231.v1

14. Wang, B.; Min, S.; Deng, X.; Shen, J.; Wu, Y.; Zettlemoyer, L.; Sun, H. Towards Understanding Chain-of-Thought Prompting: An Empirical Study of What Matters. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 2717–2739. https://doi.org/10.18653/v1/2023.acl-long.153.
15. Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; Hu, Z. Reasoning with Language Model is Planning with World Model. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 8154–8173. https://doi.org/10.18653/v1/2023.emnlp-main.507.
16. Mishra, S.; Finlayson, M.; Lu, P.; Tang, L.; Welleck, S.; Baral, C.; Rajpurohit, T.; Tafjord, O.; Sabharwal, A.; Clark, P.; et al. LILA: A Unified Benchmark for Mathematical Reasoning. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5807–5832. https://doi.org/10.18653/v1/2022.emnlp-main.392.
17. Yao, Y.; Wang, P.; Tian, B.; Cheng, S.; Li, Z.; Deng, S.; Chen, H.; Zhang, N. Editing Large Language Models: Problems, Methods, and Opportunities. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2023, pp. 10222–10240. https://doi.org/10.18653/v1/2023.emnlp-main.632.
18. Liu, W. Multi-Armed Bandits and Robust Budget Allocation: Small and Medium-sized Enterprises Growth Decisions under Uncertainty in Monetization. *European Journal of AI, Computing & Informatics* **2025**, *1*, 89–97.
19. Imani, S.; Du, L.; Shrivastava, H. MathPrompter: Mathematical Reasoning using Large Language Models. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track). Association for Computational Linguistics, 2023, pp. 37–42. https://doi.org/10.18653/v1/2023.acl-industry.4.
20. Moradi, M.; Samwald, M. Evaluating the Robustness of Neural Language Models to Input Perturbations. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 1558–1570. https://doi.org/10.18653/v1/2021.emnlp-main.117.
21. Wang, X.; Liu, Q.; Gui, T.; Zhang, Q.; Zou, Y.; Zhou, X.; Ye, J.; Zhang, Y.; Zheng, R.; Pang, Z.; et al. TextFlint: Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations. Association for Computational Linguistics, 2021, pp. 347–355. https://doi.org/10.18653/v1/2021.acl-demo.41.