PHRED: An Agentic Framework for Pharmacy-Counseling by Health Retrieval & Evidence-Driven Dialogue

Prashani Jayasingha Arachchige[1,*], Bogdan Iancu[1], Malin Andtfolk[2], Sara Rosenberg[3], Mattias Wingren[3] and Johan Lilius[1]

[1]*Department of Engineering and Information Technology, Faculty of Science and Engineering, Åbo Akademi University, Turku, 20500, Finland.*

[2]*Department of Natural and Health Sciences, Faculty of Science and Engineering, Åbo Akademi University, Vaasa, 65200, Finland.*

[3]*Experience Lab, Faculty of Human and Social Sciences, Åbo Akademi University, Vaasa, 65200, Finland.*

Abstract

Medication counseling requires not only accurate information delivery but also contextual awareness, safety validation, and clear explanation of risks capabilities that remain challenging for current conversational AI systems. Although large language models (LLMs) have made indisputable progress in language generation and understanding, they remain prone to hallucinations, opaque reasoning and safety risks in clinical settings.

To address these limitations, we present PHRED (Pharmacy-counseling by Health Retrieval & Evidence-Driven Dialogue), a modular, goal-driven multi-agent framework that improves the accuracy, interpretability, and safety of AI-assisted medication counseling. PHRED integrates Retrieval-Augmented Generation (RAG) for evidence grounding, Knowledge Graph (KG) reasoning for structured inference, and rule-based mechanisms for clinical safety validation. The framework operates through four stages: query decomposition, goal-oriented decision-making, response synthesis, and session validation. User queries are decomposed into sub-questions and classified into clinical intents, after which an orchestration layer coordinates specialized agents (RAG, KG, and rule-based) to retrieve and validate information. A Hierarchical Task Analysis (HTA)-based follow-up module further ensures conversational completeness by generating targeted prompts for follow-up.

As an initial pre-test case, the framework is applied to a sensitive and regulated medication counseling scenario, demonstrating preliminary results that support its potential for safe and transparent communication. Our framework achieved a semantic similarity of 0.89 by employing GPT-4.0 as the LLM assembler, outperforming MedDialog GPT answers. Furthermore, it attained a faithfulness score of 0.90 utilizing Mistral as the LLM assembler and a context precision of 0.89 employing GPT-4.0 as the LLM assembler. Overall, PHRED promotes a scalable agentic pipeline allowing additional agents to handle specialized tasks when required. It also promotes an interpretable architecture that enables traceability of the provided answers by combining symbolic reasoning, structured knowledge, and RAG within a transparent, human-in-the-loop framework.

Keywords

Multi-Agentic Framework, Agentic AI, Symbolic Integration, Conversational AI, medication counseling

1. Introduction

Conversational AI systems that employ LLMs exhibit notable fluency and versatility in their responses. However, these systems continue to face challenges related to factual grounding, reasoning transparency, and safety [1]. Although their responses may appear informed, LLM-based systems often function as black-box mechanisms, generating hallucinations and producing unverifiable assertions because they lack structured knowledge and robust reasoning capabilities [2]. Recent frameworks that integrate LLMs with KGs have shown promise in addressing LLMs' limitations in factuality and reasoning. Approaches such as KG-RAG [3] and WeKnow-RAG [4] demonstrate that combining structured knowledge with retrieval mechanisms can significantly improve the factual grounding and transparency of AI reasoning.

TKTP 2026: Annual Doctoral Symposium of Computer Science, 8.—9.6.2026 Turku, Finland

*Corresponding author.

prashani.jayasinghaarachchige@abo.fi (P. Jayasingha Arachchige); bogdan.iancu@abo.fi (B. Iancu); malin.andtfolk@abo.fi (M. Andtfolk); sara.rosenberg@abo.fi (S. Rosenberg); mattias.wingren@abo.fi (M. Wingren); johan.lilius@abo.fi (J. Lilius)

0000-0003-4168-9732 (P. Jayasingha Arachchige); 0000-0001-8801-7250 (B. Iancu); 0000-0001-9380-3619 (M. Andtfolk); 0000-0003-4269-5093 (M. Wingren); 0000-0002-9176-2881 (J. Lilius)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-02

CEUR-WS.org/Vol-4309/paper08.pdf

The integration of LLMs with KGs is reshaping clinical decision support and question answering in healthcare settings by enabling more accurate multi-hop reasoning and improving factual consistency [5]. Conventional RAG approaches often overlook the explicit semantic relationships between symptoms, medicines, and treatments, which are crucial for robust medical reasoning [6].

Collaboration mechanisms are often present in LLM-based multi-agent systems (MAS) and represent a variety of forms of collaborative behaviors [7]:

- cooperative behavior, where agents collaboratively search for information to provide some form of joint solution to a problem;
- competitive behavior, e.g. debating with one another prior to agreeing on an output;
- coopetitive behavior, i.e. competitive behavior among agents working together toward a common goal.

Cooperative settings facilitate alignment by sharing knowledge of a particular domain and/or problem space. Competitive processes, such as inter-agent debate, help identify and eliminate contradictions, thereby improving the quality of the reasons provided before a commitment is made [7, 8].

MAS design represents a significant improvement over other design approaches because it supports the simultaneous achievement of three key factors in the development of healthcare applications: accuracy, safety, and traceability. Conversational systems in healthcare can often produce fluent responses that cannot be verified by clinicians or patients [7, 9]. Multi-agent architectures enable the decomposition of the tasks involved in generating responses into several independent components. For example, evidence retrieval, constraint verification, and the generation of patient explanations can be separated and logged independently [7, 10]. Separation of functions provides assurances against the propagation of errors, e.g., false positives for potential drug interactions and omissions of critical contraindications [7, 9, 11]. The built-in logging capabilities in MAS systems, along with the natural feedback loops among agents, enable the incorporation of human supervision at various stages of response production. This feature is necessary for legal and ethical reasons when developing healthcare applications [7, 12].

To address these challenges, we propose a multi-agent framework: Pharmacy-counseling by Health Retrieval & Evidence-Driven Dialogue (PHRED), which integrates RAG, KG, and rule-based reasoning. This architecture enables agents to collaboratively verify facts, retrieve structured information, and apply explicit reasoning rules to produce accurate and interpretable outputs. Incorporating multiple layers of knowledge grounding and explanation generation promoting transparency, accountability, and user trust in AI systems. Beyond technical implementation, PHRED offers an alternative solution to operational challenges in pharmacy settings. Pharmacist availability is declining due to burnout and staffing shortages caused by extended occupational stress and intensified work demands [13]. PHRED provides a confidential channel for patients to discuss sensitive topics, for example, emergency contraceptives (ECPs). Although information delivery to patients is expedited through agentic design, PHRED's role is not to replace pharmacists but to help prospective patients with medication counseling. The framework is designed to support human oversight at essential decision-making points. With a human-in-the-loop approach, we ensure that a final evaluation remains under professional supervision.

Contributions:

- A modular multi-agent pipeline combining RAG, KG reasoning, and rule-based logic.
- A transparent reasoning interface that exposes intermediate decision steps for explainability.
- Multi-level knowledge grounding enables users to trace answers back to specific data sources, rules, or graph entities.
- Early stage empirical evidence compared to baseline LLM systems.

The remainder of this paper is organized as follows. Section 2 presents a brief overview of AI-enabled medical counseling frameworks. In Section 2 we introduce the framework design. Section 4 outlines the evaluation and summary of the framework performance. The conclusions are described in Section 5.

2. Related Work

MAS have a longstanding presence in artificial intelligence. The integration of large language models (LLMs) as the reasoning core of individual agents has enabled novel approaches to natural language-driven coordination, planning, and decision-making [7, 8, 9]. In LLM-based MAS, agents are typically assigned specific roles or domains of expertise and operate within structured communication frameworks. This organization supports task decomposition, parallel information retrieval, and iterative refinement of outputs through inter-agent feedback. MAS can solve problems that would otherwise be intractable for a single agent. Agents interact with one another and the environment, thereby enabling continuous context acquisition by individual agents and new possibilities for action [9, 12]. The collective knowledge acquired is then utilized when making future decisions to achieve the desired goals. As such, this type of flexibility has made MAS applicable to many different fields, including in software development, and healthcare alike [7, 8, 9, 14].

The introduction of LLMs has prompted a shift towards identifying the significant impact of transformer-based architectures on tasks such as symptom assessment, clinical documentation, and decision support. Persistent issues, including hallucinations, algorithmic bias, data privacy, and governance, have also been identified, underscoring the need for transparent and explainable frameworks, as well as sustained safety oversight in healthcare [15].

Hybrid frameworks have been developed to address these limitations by integrating Retrieval-Augmented Generation (RAG) with structured knowledge sources. Systems such as MedRAG [16] and KRAGEN [17] combine LLMs with KGs to enhance factual grounding and domain-specific reasoning capabilities. For example, MedRAG improves medical question answering by incorporating diagnostic KGs into an RAG pipeline, whereas KRAGEN employs *graph-of-thought* prompting and biomedical KGs for subproblem decomposition. A recent survey [18] highlights the emergence of agentic AI systems that integrate autonomous agents capable of retrieving domain-specific knowledge and performing complex NLP-driven tasks, including roles such as diagnostic assistants, clinical scribes, health educators, and therapeutic counselors.

Our proposed framework advances the state of the art by introducing a modular multi-agent architecture that explicitly separates reasoning, retrieval, and safety validation into specialized agents. In contrast to previous systems, it features a goal-oriented orchestration layer that dynamically assigns tasks based on intent classification and incorporates a rule-based agent to enforce clinical safety constraints. Additionally, applying insights from Hierarchical Task Analysis (HTA) to assess dialogue completeness and a dual-stage query decomposition pipeline improve both interpretability and user alignment. By integrating symbolic reasoning, structured knowledge, and LLMs within a transparent and traceable pipeline, our framework aims to provide a more robust, explainable, and clinically reliable system that complements existing approaches.

3. Framework design

This section introduces a collaborative, multi-layered framework for medication counseling. The design of this framework followed the Design Science Research Methodology (DSRM) [19], which includes six stages. The six stages are problem identification, defining objectives for the solution, design and development, demonstration, evaluation and communication which is followed throughout this research and its continuity. The problem identification and objectives for the solution are reflected within the design of the framework. As illustrated in Figure 1, the pipeline comprises four sequential stages: (i) Problem Identification, which decomposes complex user queries into sub-questions; (ii) Goal-Oriented Decision, involving goal classification and agent selection for targeted information retrieval; (iii) Response Synthesis, which integrates retrieved data, rules, and entities to generate context-aware outputs; and (iv) Session Validation, which ensures conversational completeness by assessing the need for further clarification. Each stage comprises one or more agents that operate collaboratively in a structured and interoperable manner. The specific roles and interactions of these agents are elaborated

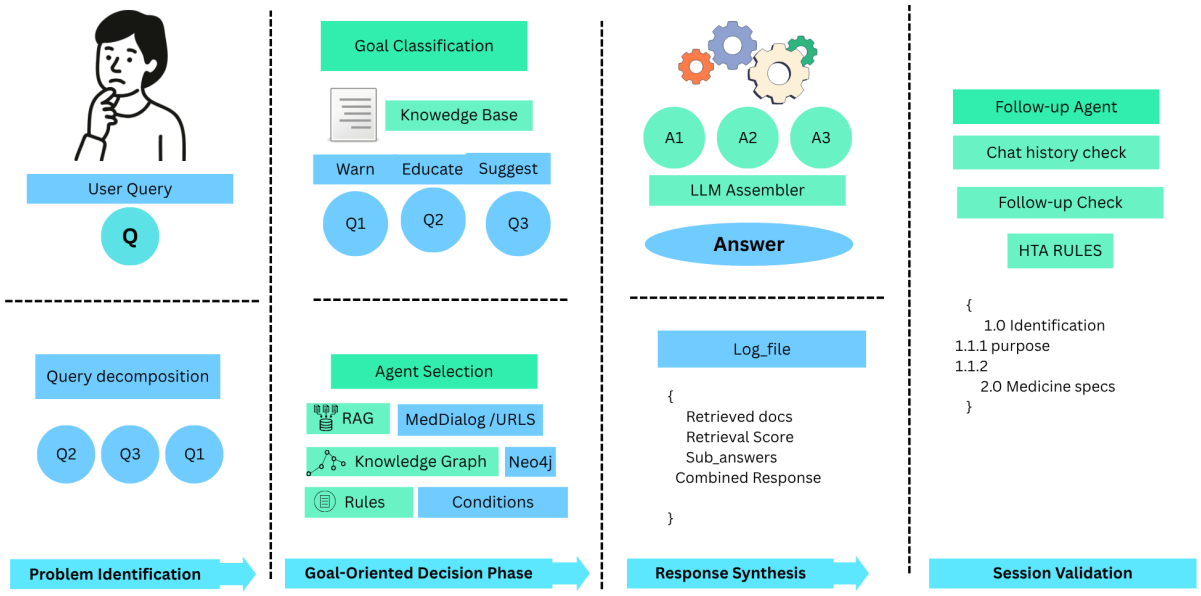


Figure 1: Overall architecture of the proposed goal-driven agentic framework for medication counseling. The system operates in four sequential stages: Problem Identification, Goal-Oriented Decision, Response Synthesis, and Session Validation. User queries are first decomposed into sub-questions and classified into goal categories (educate, warn, or suggest). Based on the goal, the Agent Selector dynamically engages specialized agents, including the RAG, rule-based, and KG agents, to retrieve accurate and context-specific information. The LLM Assembler consolidates the agent responses, removes redundancy, and generates a coherent final answer. Finally, the Follow-Up Agent, guided by Hierarchical Task Analysis (HTA) rules, verifies conversational completeness and proposes follow-up prompts when necessary.

in the subsequent sections. This framework combines large language models (LLMs) with explicit rule-based reasoning and structured knowledge representations. This hybrid design supports coherent, goal-driven interactions while enhancing the explainability through a transparent decision logic and traceable information flow.

3.1. Problem Identification

Several studies in the context of cognitive sciences and Natural Language Processing (NLP) break down complex sentences to reduce cognitive load, identify the core problem, and improve interpretability [20] [21]. User queries in the medication counseling context can vary widely in both type and complexity, as the same factual information may be expressed through diverse linguistic forms or contextual parameters, such as time and situational relevance. Regardless of the complexity, each user query typically reflects an underlying intent or goal within a conversational context. In order to identify the intent of the user, at the first stage of our framework, a decomposition agent is used to simplify complex queries through a two-level process:

- Syntactic and Punctuation-Based Decomposition: Complex sentences are segmented based on grammatical structures and punctuation cues such as conjunctions and delimiters to isolate independent semantic units with a similar principle to chunking and linguistic parsing [22].
- Prompt Engineering Driven Fallback: When explicit syntactic cues are insufficient, a fallback strategy leveraging prompt engineering is employed to reformulate the original query into a set of smaller, answerable questions.

This dual-stage query analysis, as shown in Figure 2, facilitates the robust separation of user queries into smaller sub-queries, thereby highlighting modular reasoning as a foundation for the subsequent stage in the framework.

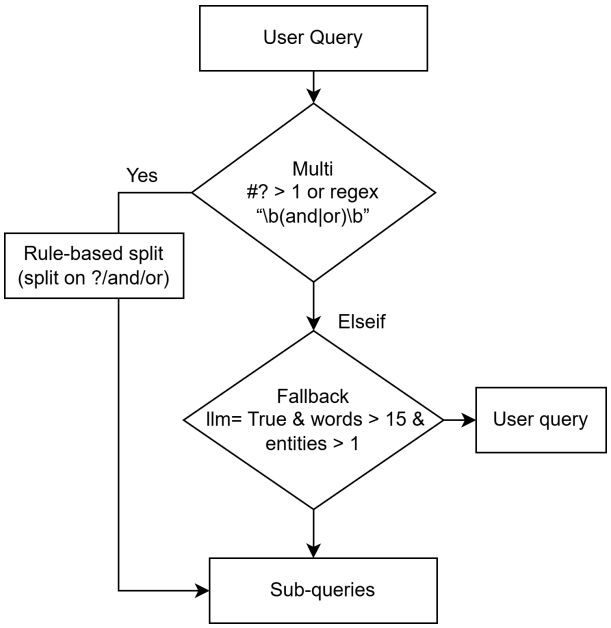


Figure 2: Dual-stage decomposition reflecting implementation: rule-based split on obvious multi-question cues; otherwise semantic LLM fallback only if use_llm & words > 15 & entities> 1; else pass-through.

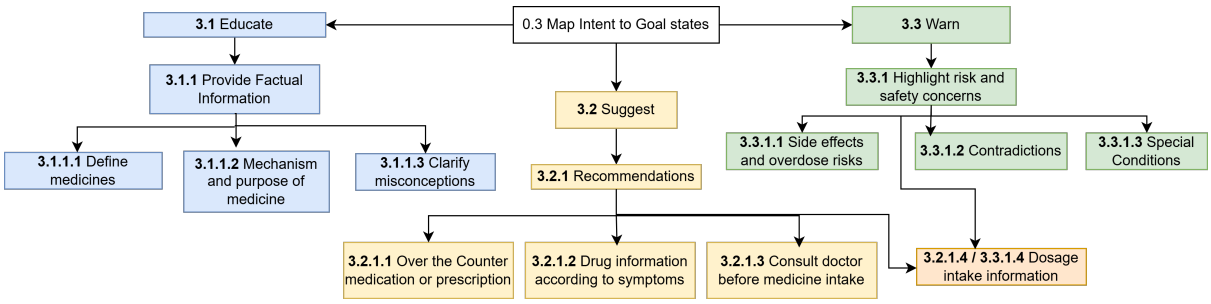


Figure 3: Hierarchical intent–goal mapping based on HTA example for medication counseling, showing how educate, warn, and suggest intents decompose into goal states guiding the response generation.

3.2. Goal-Oriented Decision Phase

As discussed in [23], the proposed system adopts a goal-based architectural paradigm designed to manage complex objectives in agentic environments. Following this paradigm, the next stage of the framework comprises two core processes: (i) a goal classification phase, which identifies the underlying intent of each query, and (ii) an information retrieval stage, which employs multiple agents to gather evidence in accordance with the identified goal.

3.2.1. Goal Classification

Given the medication counseling context of this framework, user intents are categorized into three primary communicative goals based on the Hierarchical Task Analysis (HTA) example for a medication counseling dialogue [24] to prepare a knowledge base with example questions. The three categories and their divisions are shown in Figure 3. Therefore, intent labeling is performed through sentence-embedding similarity between the subqueries and the knowledge base using the SentenceTransformer all-MiniLM-L6-v2 model. During inference, the highest-scoring embedding match determines the assigned intent of the input. Once the goal is identified, it is passed to the subsequent phase for intent-aligned retrieval and response generation.

3.2.2. Goal-Oriented Agents

The objective of goal-oriented agents is to retrieve the information necessary to answer each query in a trustworthy and accurate manner. In this framework, trustworthiness is operationalized through four dimensions: accuracy, completeness, safety, and transparency [25]. To satisfy these dimensions, the overall task is delegated to specialized helper agents. The retrieval (RAG) helper agent provides accurate, evidence-grounded information; the Knowledge-Graph helper agent supplies structured, verifiable domain facts to ensure completeness; and the rule-based safety helper agent identifies critical safety-related contextual elements. These outputs are integrated through a transparent, rule-driven orchestrator, ensuring that the reasoning process is not a black box and is aligned with established trustworthiness criteria. As the framework is designed to support sensitive medication-counseling queries, the current implementation focuses on emergency contraceptives, with retrieval constrained to this domain, which has been previously examined through Hierarchical Task Analysis (HTA) and empirical studies reported in the literature. The methods of the helper agents are as follows:

a. The RAG Agent [26] uses a hybrid retrieval pipeline that integrates a structured knowledge base built from MedDialog entries [27] question–answer pairs and trusted external medical sources with dense vector embeddings. The pipeline consists of:

1. KB construction, where MedDialog entries and verified URLs are stored in JSON format with support for single, concatenated, and multi-answer modes.
2. embedding generation using the multi-qa-mpnet-base-dot-v1 model to encode queries and answers into 768-dimensional vectors
3. vector retrieval via ChromaDB [28], which computes cosine similarity (distance $= 1 - \cos(\theta)$) [29] and returns the top-k semantically closest entries.

A relevance threshold filters out low-scoring results, and fallback logic ensures safe, non-speculative responses.

Listing 1: Relation templates used by the Knowledge Graph Agent.

```
REL_TEMPLATES = {
    "CONTAINS":"{h} contains {t}",
    "USES":"{h} uses {t}",
    "WORKS_BY":"{h} works by {t}",
    "EFFECTIVE":"{h} is effective {t}",
    "HAS_SIDE_EFFECT" "{h} has side effect {t}",
    "NOT_RECOMMENDED_FOR":"{h} is not recommended for {t}",
    "SPECIAL_CASE":"{h} special case: {t}",
    "LESS_EFFECTIVE_IN":"{h} is less effective in {t}",
    "EFFECTIVENESS_UNCHANGED_IN":"{h} effectiveness unchanged in {t
        }",
    "INEFFECTIVE_IF":"{h} is ineffective if {t}",
}
```

b. The KG agent retrieves factual, structured information about medications directly from the Neo4j database. Unlike generative agents, it outputs machine-interpretable triplets that capture relationships among medicines in simple, single-text form rather than free-form text. The KG was built for the test case of the emergency contraceptive pill.

The operation of the KG Agent can be summarized as follows:

- Entity Resolution: User queries are parsed using fuzzy string-matching techniques implemented through RapidFuzz [30] to identify the most relevant entity in the KG. Entities were matched using a similarity threshold of 0.65, enabling recognition even when the queried term did not exactly match the stored entity label. Case-insensitive normalization (toLower()) was also established (e.g., “Copper IUD” and “copper iud”).

- Fact Extraction: Once the target entity is identified, the agent expands its neighborhood within the graph to a depth of two hops [31], thereby providing additional information about the relationships in the nodes.
- Fact Formatting: Retrieved triples are converted into templated factual statements using predefined REL-TEMPLATES *Listing* 1. [32]

The KG Agent delivers structured, fact-based responses by retrieving and presenting relevant knowledge graph triples through probabilistic entity matching and controlled natural-language generation.

c. The Rule-based Agent performs deterministic evaluation of user inputs to identify potential clinical risks, safety alerts, and vulnerable conditions. Its primary function is to ensure that the responses generated by other agents (e.g., LLMs or retrieval modules) adhere to safety constraints, thereby minimizing the possibility of unsafe or harmful recommendations. The agent utilizes a series of pattern-matching rules constructed around three principal dimensions of clinical risk:

- Emergency Terms (e.g., chest pain, difficulty in breathing).
- Vulnerable populations (e.g., children, pregnancy).
- Time-critical or misuse indicators (e.g., missed dosage).

User queries are scored, producing a cumulative risk level: high (≥ 3) triggers a high emergency alert, medium (=2) adds clinical caution, and low (≤ 1) continues through the standard pipeline.

The system aims to achieve balanced integration of transparency, interpretability, and clinical reliability by embedding deterministic safety logic alongside probabilistic reasoning modules (e.g., LLMs and retrieval agents).

3.3. Response synthesis

The information retrieved from each helper agent varies in modality and may overlap. To synthesize these into a single coherent response, the system employs constrained prompting methods: self-consistency and chain of thought prompting [33] that restrict model behavior through explicit task-level constraints rather than open-ended generation. The prompt enforces three policies:

- Responses must rely solely on information retrieved by the helper agents: principles that reduce hallucination by grounding outputs in fact-based knowledge
- The structure and tone of the response must match the goal identified for the query, reflecting task-oriented dialogue design strategies [34]
- Medical terminology must be preserved to minimize semantic distortion and remain consistent.

This procedure consolidates overlapping content, supports factual alignment, and produces a coherent user-appropriate response while maintaining alignment with safety-critical requirements.

3.4. Session Validation

To ensure conversational completeness, the framework applies an HTA-based rule set (Fig. 3) through a Follow-Up Agent that monitors which goal-specific subtopics have been addressed in the dialogue. Dialogue turns, and topic labels are embedded using a SentenceTransformer model, and topics are marked as covered when the threshold is exceeded. If unaddressed topics remain, the agent prompts an LLM to generate a single follow-up question; otherwise, it terminates. The procedure is summarized in Algorithm 1. To support transparency, a comprehensive log file of the dialogue is maintained and subsequently summarized for human-in-the-loop assessment. The log file includes detailed records of retrieved documents from the RAG as well as outputs from the KG and rule-based agents. Additionally, it captures retrieval scores, intermediate sub-answers corresponding to each sub-query, and the final aggregated response (Fig. 1). This structured logging enables traceability of the information sources and reasoning steps underlying the system's responses, thereby facilitating validation and analysis.

Algorithm 1 HTA-Based Follow-Up Agent

```
Input: Dialogue history H, Topic set by HTA T
for each topic t_i ∈ T do
    Compute cosine similarity to H
    if similarity ≥ 0.75 then
        Mark t_i as covered
    end if
end for
if all topics covered then
    return null
else
    Generate follow-up question
end if
```

4. Results & Preliminary Evaluation

The early stage evaluation of the proposed agentic framework focuses on three key dimensions: (i) grounding and factual alignment of generated responses, (ii) semantic alignment between the retrieved content and user intent, and (iii) distribution of communicative goals following query decomposition. This evaluation constitutes a preliminary, first-stage test, designed to assess the feasibility and internal consistency of the framework's core components. The framework is currently applied to a constrained domain and evaluated using a medically oriented but non-pharmacy-specific dataset (MedDialog) [27], with a focus on ECPs. The choice of ECP-related scenarios is motivated by the need for timely, precise, and context-aware information delivery, in which user queries often involve implicit constraints and require sensitive clarifications. This makes the domain well-suited for evaluating the framework's ability to support safe decision-making and personalized, context-aware response generation. While the present results are restricted to a specific set of scenarios, the framework is designed to generalize to a broader range of medications and clinical contexts. This initial evaluation, therefore, establishes a baseline for subsequent, more extensive testing across diverse datasets and interaction settings.

4.1. Grounded Response Quality

To assess the extent to which model responses remain anchored in retrieved evidence rather than parametric memory, we applied the RAG Assessment Score (RAGAS) [35]. We report two metrics relevant to our retrieval pipeline:

1. Faithfulness, which measures factual consistency with supporting documents. Let $S(a) = \{s_1, \dots, s_n\}$ denote the set of atomic statements extracted from the generated answer a, and let $C = \{c_1, \dots, c_m\}$ be the retrieved context C. For each statement s_i, we define a support indicator

$$\text{supp}(s_i, C) = \begin{cases} 1, & \text{if } s_i \text{ is supported by at least one } c_j \in C, \\ 0, & \text{otherwise.} \end{cases}$$

The faithfulness score for (a, C) is then given by

$$\text{Faithfulness}(a, C) = \frac{1}{|S(a)|} \sum_{i=1}^{|S(a)|} \text{supp}(s_i, C).$$

2. Context Precision, capturing the relevance and correctness of the retrieved knowledge. Given the retrieved context set $C = \{c_1, \dots, c_m\}$, we define a relevance indicator

$$\text{rel}(c_j) = \begin{cases} 1, & \text{if } c_j \text{ is relevant to answering the query,} \\ 0, & \text{otherwise.} \end{cases}$$

The context precision is defined as

$$\text{ContextPrecision}(C) = \frac{\sum_{j=1}^{|C|} \text{rel}(c_j)}{|C|} = \frac{|C_{\text{rel}}|}{|C|}.$$

The two retrieval measurements, faithfulness and context precision, across different LLM models using RAGAS are listed in Table 1.

Table 1
Semantic Similarity and RAGAS evaluation across Model Variants: 1. Semantic similarity was measured using multiple prompting techniques and GPT-generated responses from the MedDialog dataset. 2. RAGAS evaluation assessing faithfulness and context precision.

Metric	MedDialog	GPT-4.0	GPT-4 Mini	Mistral
Semantic Similarity				
Chain of Thought	0.62	0.63	0.58	0.62
Few-Shot	—	0.58	0.57	0.60
Self-Consistent	—	0.77	0.73	0.67
Max Score	0.84	0.89	0.82	0.81
RAGAS				
Faithfulness		0.82	0.81	0.90
Context Precision		0.89	0.83	0.67

4.2. Semantic Similarity and Goal Alignment

We evaluated semantic similarity using cosine similarity between reference answers and model-generated responses, which were computed across multiple prompting techniques. The MedDialog dataset includes GPT-generated responses, referred to as MedDialog in Table 1. The maximum semantic scores are also reported in Table 1, indicating that constrained self-consistent prompting improves answer relevance in this study.

Following query decomposition, we analyzed the distribution of communicative goals assigned by the goal classifier. Let N_{total} denote the total number of subqueries produced, and N_g represent the number assigned to goal $g \in \{\text{educate}, \text{warn}, \text{suggest}\}$. We define the goal proportion as:

$$P(g) = \frac{N_g}{N_{\text{total}}}.$$

As shown in Table 2, the classifier predominantly yields *warn*-aligned intents ($P(\text{warn}) = 0.78$), reflecting the safety-critical nature of emergency contraception queries. This behavior is consistent with task-oriented dialogue systems in medical domains, where system responses are expected to prioritize risk mitigation and safety.

Table 2
Distribution of Classified Goals Across Decomposed Sub-Queries

Goal Category	Proportion
Educate	0.08
Warn	0.78
Suggest	0.14

4.3. Discussion

Early-stage evaluations suggest that retrieval grounding maintains alignment with the knowledge base, whereas goal-based orchestration supports structured, medically coherent communication. Current limitations, including the domain-specific focus on ECPs, reliance on static safety rules, the absence of expert human evaluation, and the lack of baseline comparisons, highlight key areas for future improvement. To address these limitations, the next steps in the evaluation phases will incorporate expert-driven human assessment [36] by domain specialists, such as pharmacists, to assess clinical relevance, safety, and provide feedback on the generated responses. This will also include a metrical analysis based on the HTA rules and goal mapping from the framework. Additionally, future work will focus on evaluating the framework across a wider range of medications and more diverse datasets. This includes strengthening safety agents through improved ethical and privacy-aware mechanisms, incorporating a broader set of clinical guidelines and adding international-specific medical regulations to the current regional guidelines and healthcare practices. The framework will also be evaluated through user studies [37], examining usability, trust and personalized interaction quality in realistic scenarios. These steps will support a more comprehensive validation of the framework as it scales to broader medical domains and more dynamic, context-sensitive dialogue systems.

5. Conclusion

This study advances the integration of multidisciplinary approaches through structured knowledge representation and retrieval-augmented reasoning using multi-agent AI systems. Building on recent developments in RAG, KGs and rule-based systems, the proposed framework addresses persistent challenges related to the factual accuracy, interpretability, and explainability of large language models. The primary contribution of this work lies in the design of a modular multi-agent architecture that unifies RAG, KG-based reasoning, and rule-based logic within a coordinated pipeline, enabling complementary reasoning strategies and different task specialization across agents. Preliminary evaluations demonstrate promising performance across multiple dimensions. The framework achieves a semantic similarity score of 0.89 when using GPT-4 as the LLM assembler, outperforming baseline results on the MedDialog dataset. Additionally, a faithfulness score of 0.90 is obtained using Mistral, while a context precision score of 0.89 is achieved with GPT-4. These results indicate the effectiveness of the proposed architecture in maintaining alignment between generated responses and supporting evidence. While these findings are based on a preliminary evaluation, the framework establishes a goal-driven pipeline that enhances traceability and safety in information generation. Future work will extend the evaluation to broader medical scenarios and incorporate expert human assessments to further validate the system's applicability in real-world settings. Overall, this work provides a foundation for developing robust, explainable, and safety-aware multi-agent systems for healthcare-oriented dialogue systems.

Declaration on Generative AI

During the preparation of this manuscript, the authors used Grammarly, an AI-assisted editing tool for language improvement, specifically to improve writing style to enhance readability, check grammar and spelling, and reword sentences to improve clarity. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Chua, Y. Li, S. Yang, C. Wang, L. Yao, Ai safety in generative ai large language models: A survey, arXiv preprint arXiv:2407.18369 (2024).

[2] S. S. Rahman, M. A. Islam, M. M. Alam, M. Zeba, M. A. Rahman, S. S. Chowa, M. A. K. Raiaan,

S. Azam, Hallucination to truth: A review of fact-checking and factuality evaluation in large language models, arXiv preprint arXiv:2508.03860 (2025).
[3] D. Sanmartin, Kg-rag: Bridging the gap between knowledge and creativity, arXiv preprint arXiv:2405.12035 (2024).
[4] W. Xie, X. Liang, Y. Liu, K. Ni, H. Cheng, Z. Hu, Weknow-rag: An adaptive approach for retrieval-augmented generation integrating web search and knowledge graphs, arXiv preprint arXiv:2408.07611 (2024).
[5] R. Xu, P. Jiang, L. Luo, C. Xiao, A. Cross, S. Pan, J. Sun, C. Yang, A survey on unifying large language models and knowledge graphs for biomedicine and healthcare, in: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2, 2025, pp. 6195–6205.
[6] Z. Xu, M. J. Cruz, M. Guevara, T. Wang, M. Deshpande, X. Wang, Z. Li, Retrieval-augmented generation with knowledge graphs for customer service question answering, in: Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval, 2024, pp. 2905–2909.
[7] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, H. D. Nguyen, Multi-agent collaboration mechanisms: A survey of llms, arXiv preprint arXiv:2501.06322 (2025).
[8] S. Chen, Y. Liu, W. Han, W. Zhang, T. Liu, A survey on llm-based multi-agent system: Recent advances and new frontiers in application, arXiv preprint arXiv:2412.17481 (2024).
[9] T. Guo, X. Chen, Y. Wang, R. Chang, S. Pei, N. Chawla, O. Wiest, X. Zhang, Large language model based multi-agents: A survey of progress and challenges., in: 33rd International Joint Conference on Artificial Intelligence (IJCAI 2024), IJCAI; Cornell arxiv, 2024.
[10] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, M. Gerstein, Medagents: Large language models as collaborators for zero-shot medical reasoning, in: ICLR 2024 Workshop on Large Language Model (LLM) Agents, 2024.
[11] K. Zuo, Y. Jiang, F. Mo, P. Lio, Kg4diagnosis: A hierarchical multi-agent llm framework with knowledge graph enhancement for medical diagnosis, in: AAAI Bridge Program on AI for Medicine and Healthcare, PMLR, 2025, pp. 195–204.
[12] Y. Yang, Q. Peng, J. Wang, Y. Wen, W. Zhang, Llm-based multi-agent systems: Techniques and business perspectives, arXiv preprint arXiv:2411.14033 (2024).
[13] D. N. patel, A. Andy, Addressing pharmacist burnout and staffing shortages: A review of challenges and solutions, International Journal For Multidisciplinary Research (2024).
[14] S. Han, Q. Zhang, W. Jin, Z. Xu, Llm multi-agent systems: Challenges and open problems, arXiv preprint arXiv:2402.03578 (2024).
[15] J. C. Chow, K. Li, Large language models in medical chatbots: opportunities, challenges, and the need to address ai risks, Information 16 (2025) 549.
[16] X. Zhao, S. Liu, S.-Y. Yang, C. Miao, Medrag: Enhancing retrieval-augmented generation with knowledge graph-elicited reasoning for healthcare copilot, in: Proceedings of the ACM on Web Conference 2025, 2025, pp. 4442–4457.
[17] N. Matsumoto, J. Moran, H. Choi, M. E. Hernandez, M. Venkatesan, P. Wang, J. H. Moore, Kragen: a knowledge graph-enhanced rag framework for biomedical problem solving using large language models, Bioinformatics 40 (2024) btae353.
[18] G. Xu, X. Li, Y. Chen, Y. Duan, S. Wu, A. Yu, C.-H. Chiu, J. Ni, N. Tang, T. J.-J. Li, et al., A comprehensive survey of agentic ai in healthcare (2025).
[19] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, A design science research methodology for information systems research, Journal of management information systems 24 (2007) 45–77.
[20] H. A. Simon, The Sciences of the Artificial, 3 ed., MIT Press, Cambridge, MA, USA, 1996.
[21] D. Jurafsky, J. H. Martin, Speech and language processing, Draft manuscript, Chapter 17: "Semantics", 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/.
[22] M. Collins, Three generative, lexicalised models for statistical parsing, in: Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain, 1997, pp. 16–23.
[23] D. B. Acharya, K. Kuppan, B. Divya, Agentic ai: Autonomous intelligence for complex goals—a comprehensive survey, IEEE Access 13 (2025) 18912–18936.

[24] M. Wingren, S. Andersson, S. Rosenberg, M. Andtfolk, S. Hägglund, P. J. Arachchige, L. Nyholm, Using role-play and hierarchical task analysis for designing human-robot interaction, in: O. Palinko, L. Bodenhagen, J.-J. Cabibihan, K. Fischer, S. Šabanović, K. Winkle, L. Behera, S. S. Ge, D. Chrysostomou, W. Jiang, H. He (Eds.), Social Robotics, Springer Nature Singapore, Singapore, 2025, pp. 319–328.
[25] NIST, Artificial intelligence risk management framework: Generative artificial intelligence profile, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (2024).
[26] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021.
[27] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, P. Xie, Meddialog: Large-scale medical dialogue datasets, 2020.
[28] P. Martra, Large Language Models Projects, Apress, Berkeley, CA, 2024.
[29] M. Farouk, Measuring text similarity based on structure and word embedding, Cognitive Systems Research 63 (2020) 1–10.
[30] M. Bachmann, Rapidfuzz: Fuzzy string matching in python, 2024.
[31] T. Wang, R. Huang, H. Wang, H. Zhi, H. Liu, Multi-hop knowledge graph question answer method based on relation knowledge enhancement, Electronics 12 (2023).
[32] C. Gardent, A. Shimorina, S. Narayan, L. Perez-Beltrachini, The WebNLG challenge: Generating text from RDF data, in: J. M. Alonso, A. Bugarín, E. Reiter (Eds.), Proceedings of the 10th International Conference on Natural Language Generation, Association for Computational Linguistics, Santiago de Compostela, Spain, 2017, pp. 124–133.
[33] J. Liu, F. Liu, C. Wang, S. Liu, Prompt engineering in clinical practice: Tutorial for clinicians, J Med Internet Res 27 (2025) e72644.
[34] Z. Wang, N. Geng, Z. Guo, W. Ma, M. Zhang, Human vs. agent in task-oriented conversations, arXiv preprint arXiv:2509.17619 (2025).
[35] S. Es, J. James, L. E. Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2024, pp. 150–158.
[36] J. Deriu, A. Rodrigo, A. Otegi, G. Echegoyen, S. Rosset, E. Agirre, M. Cieliebak, Survey on evaluation methods for dialogue systems, Artificial Intelligence Review 54 (2020) 755–810. doi:https://doi.org/10.1007/s10462-020-09866-x.
[37] C. Courage, K. Baxter, Understanding your users: a practical guide to user requirements methods, tools, and techniques, Gulf Professional Publishing, 2005.