ACCSCIENCE PUBLISHING

Artificial Intelligence and Health

ORIGINAL RESEARCH ARTICLE

Reducing cognitive load in polypharmacy: A prototype clinical decision support system

Insiya Fatma[1], Mary A. Dolansky[2], Evelyn G. Duffy[2], and Colin K. Drummond[1]*

[1]Department of Biomedical Engineering, School of Engineering, Case Western Reserve University, Cleveland, Ohio, United States of America

[2]Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, Ohio, United States of America

Abstract

Polypharmacy, defined as the concurrent use of five or more medications, presents a significant health risk for older adults, particularly when managing complex "as-needed" (PRN) regimens. While healthcare providers establish treatment plans, the operational complexity of managing daily regimens often exceeds the support structures available to the patient, increasing the risk of unintended medication errors. This paper describes the development of a prototype clinical decision support tool designed to enable exploration of automated systems with the potential to reduce this cognitive load by integrating natural language processing with dynamic prescription transcription. The prototype utilizes a Retrieval-Augmented Generation (RAG) architecture, merging targeted information retrieval from a clinical knowledge base with the generative capabilities of Large Language Models (LLMs). This allows the system to interpret natural language queries and provide contextually relevant safety data. At this stage of development, the prototype does not test cognitive load explicitly, rather provides preliminary technical feasibility through retrieval performance, guardrail behavior, and prototype feasibility. Preliminary results are encouraging; the prototype effectively processed complex inquiries regarding NSAID usage, accurately identifying drug-drug interactions (*e.g.*, with lithium) and flagging gastrointestinal risks. Furthermore, the system demonstrated essential safety guardrails by identifying its own data limitations and directing users to professional consultation, but the system has not been clinically validated for such use. These initial findings suggest that RAG-based prototypes offer a feasible pathway toward reducing the cognitive demands of polypharmacy, providing a foundation for future systems that integrate patient-specific health data to further minimize clinical ambiguity.

Keywords: Polypharmacy; Retrieval-Augmented Generation; Medication adherence; Large Language Models; Clinical decision support systems; Geriatric health informatics

***Corresponding author:**
Colin K. Drummond
(colin.drummond@case.edu)

Citation: Fatma I, Dolansky MA, Duffy EG, Drummond CK. Reducing cognitive load in polypharmacy: A prototype clinical decision support system. *Artif Intell Health*. 2026;3(4):026170039. doi: 10.36922/AIH026170039

Received: April 23, 2026

Revised: May 29, 2026

Accepted: June 12, 2026

Published online: June 23, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

In the United States, non-adherence to prescribed medication regimens has reached epidemic proportions,[1-3] representing a critical public health crisis. A recent report indicates that up to 50% of patients fail to take their medications as prescribed.[4] As the fourth leading cause of death, mismanagement accounts for approximately 125,000

fatalities and between $100 billion and $300 billion in excess or avoidable medical expenditures annually.[5-8] Beyond the financial toll, this crisis drives over 700,000 emergency department visits and nursing home admissions,[9,10] underscoring a fundamental breakdown in the management of complex therapy.

Managing these regimens presents a significant cognitive and logistical burden, which is acutely exacerbated for polypharmacy patients—those taking five or more medications—and individuals living in social isolation. While statistics indicate that 61% of adults over 65 manage multiple medications,[11,12] the demand for robust adherence support is a universal concern that transcends age. Across diverse clinical conditions, sub-optimal treatment adherence contributes to as many as half of all treatment failures.[9] For patients managing chronic co-morbidities, the complexity of coordinating fixed-schedule prescriptions with pro re nata (PRN or "as needed") doses introduces a high-stakes cognitive load.[13] Whether the user is an older adult navigating chronic care, a busy professional, or a full-time parent, the risk of a missed dose or an adverse health event (AHE) remains a primary driver of preventable hospital readmissions.

Treatment adherence is highly dependent on the clarity and resonance of health communication, which often influences patient outcomes more significantly than demographic variables.[14] However, significant operational gaps emerge in home-care settings that rely on informal caregivers; approximately 60% of family caregivers maintain full-time employment, limiting their ability to provide continuous oversight,[15] which naturally limits the window for continuous medication oversight. While families may utilize care assistants for additional support, the technical intricacies of pharmacology and precise timing can impose an overwhelming cognitive load on providers who are only present for limited durations.[16,17] To support these fragmented care environments, there is an opportunity to implement interactive AI-based systems that manage complex schedules and tracks contraindications,[18-20] thereby streamlining communication across the patient's entire care network. By applying the Norman and Skinner *Lily Model* [20-22] to align clinical data with user capabilities, the system integrates diverse information domains to ensure clinical data are always intuitive and actionable. Ultimately, effective polypharmacy management requires clinical instructions to be harmonized with the user's daily environment and communication preferences, transforming a complex medical task into a manageable daily routine. Table 1 indicates how linking a clinical information system to a user's unique communication strength facilitates more intuitive and confident medication management. This transition is particularly beneficial for individuals managing complex medication regimens independently.

A clinical decision support system effectively bridges the gap between complex medical protocols and a user's daily, lived reality. Recent frameworks emphasize that next-generation clinical tools must move away from static data repositories toward active, intelligent resilience frameworks.[23] By focusing heavily on user-centered design principles that feature intuitive layouts and personalized user control, systems drastically reduce the user's cognitive load and lower behavioral barriers. The fundamental idea is to align the communication styles with routines already familiar to the user. Recent implementations of home-care clinical decision support system platforms demonstrate that adapting interfaces to the user's familiar environments and daily routines significantly elevates usability while

Table 1. Norman and Skinner[22] for computer decision support adaptive medical output

Feature	Standard CDS output (static)	Literacy-aware output (adaptive)	Norman–Skinner model
Drug identification	"Take 20 mg of atorvastatin."	"Take the small white round pill for your cholesterol."	Traditional/Health
PRN instructions	"Take ibuprofen for breakthrough pain; do not exceed 1200 mg/24 h."	"You can take this for your back pain, but no more than three times today because it might hurt your stomach."	Scientific/Health
Interaction alert	"Warning: Potential contraindication between warfarin and vitamin K."	"I see you're about to take your blood thinner. Avoid eating spinach or kale today, as it changes how your medicine works. "	Information/Science
Dosing logic	"Administer 5 mL every 6 h"	"Use the orange plastic syringe to pull the liquid up to the number 5 line. "	Computer/Health
Tone & complexity	Clinical/Instructional (12th grade level).	Conversational/Scaffolded (5th–8th grade level).	Media/Traditional

promoting safer, independent self-management.[24] In the absence of on-site clinical support to clarify instructions, an adaptive information system better aligns health information with the user's familiar communication styles and digital interaction patterns.[25] Rather than utilizing a static interface, an adaptive context-responsive system dynamically calibrates the delivery of information to ensure it remains intuitive and actionable for the individual.[26] For instance, when managing as-needed (PRN) medications, the system provides interactive prompts that match the user's current level of comfort with medical concepts, effectively translating complex clinical protocols into confident home administration. This transition from abstract data to practical guidance might involve utilizing concrete visual cues—such as a medication's specific color and shape—or providing plain-language explanations for contraindications. By offering this personalized scientific context, the system fosters greater patient independence, ensuring the user feels informed and empowered rather than overwhelmed by the intricacies of their care.

While "smart speaker" technology (*e.g.*, Amazon Echo/Alexa) has achieved significant market penetration—with one in five Americans interacting with a digital assistant—their application in healthcare remains largely recreational or lacks the precision required for complex care.[27,28] Despite the transition of many platforms toward U.S. Health Insurance Portability and Accountability Act (HIPAA) compliance and improved system trust,[29] legal liabilities,[30,31] and ethical issues persist.[32,33] Existing commercial medication reminders often fail to address the specific needs of polypharmacy management due to several systemic limitations [34]:

(i) Static logic: Standard systems typically operate on fixed schedules and cannot dynamically adjust to medication changes or provide real-time alerts for context-specific PRN contraindications.
(ii) Proprietary constraints: Many management algorithms are "black boxes," which prevents collaborative improvement and verification by healthcare researchers and clinicians.
(iii) Communication rigidity: Existing tools often assume a one-size-fits-all approach to information delivery. They lack the adaptive, multi-modal interfaces necessary to support users with varying cognitive loads, physical limitations, or unique communication preferences.

We recognize that the "digital divide" remains a significant hurdle to practical implementation of digital tools for polypharmacy management; not all families possess the financial or infrastructural resources to acquire and maintain decision support technology. Acknowledging that technology does not yet reach every household, we suggest that to at least begin exploring these frameworks provides an opportunity to pilot advance future clinical decision support systems in real-world settings. Such research paves the way for future broad-scale adoption, transforming sophisticated medication management from a specialized tool into a universally accessible public health resource.

Enhancing medication decision-making requires exploring new methods for linking intuitive user interfaces directly with authoritative clinical databases, ensuring the delivery of precise and unambiguous information. Recently, the Graph-Enhanced Retrieval-Augmented Generation ("Graph-Enhanced RAG") architecture provides this foundation by ensuring that adaptive communication is built on verified, context-specific knowledge.[35-37] Standard Large Language Models (LLMs) may generate unsupported or vague content (sometimes known as "hallucinations"). To address this situation, Graph-Enhanced RAG forces the artificial intelligence (AI) to pull facts from a verified medical database instead of relying on its memory. This involves "AI agents" (digital assistants) that can plan, make decisions, and fix their own mistakes without a human guiding each step.[38] The agentic pipeline functions as a sophisticated linguistic bridge: once the hybrid retrieval mechanism identifies a relevant medication protocol or drug-drug interaction, the agentic layer dynamically calibrates this information to align with the user's communication preferences and existing conceptual familiarity. For example, in a home-care environment, when a patient or caregiver queries the system regarding a specific treatment, the AI retrieves the pertinent clinical data and translates it into an intuitive, actionable format. This process ensures that the guidance remains clinically rigorous yet transitionally accessible, empowering the user to follow medication protocols with confidence and clarity, thereby reducing the likelihood of administration errors

The current work proposes a prototype AI-driven framework that seeks to prioritize clinical accuracy for all users, ensuring that the complexities of daily life—whether professional demands or caregiving responsibilities—do not compromise patient safety. To achieve this, we developed a query-based decision-support system designed to adapt medical information to each patient's unique situational needs; however, please note that this system has not yet undergone clinical validation trials.

Central to this platform is the Python language implementation of a Graph-Enhanced RAG blueprint, which grounds LLMs in curated, authoritative medical data. By anchoring the system in verified knowledge,

Graph-Enhanced RAG significantly mitigates the risk of "hallucinations" and provides a reliable foundation for clinical decision support. Building upon this, we introduce a Graph-Enhanced RAG architecture tailored specifically for pharmaceutical support. Instead of just searching through text, the system uses a "knowledge graph" to understand how different drugs and conditions are connected.[39-41] While standard RAG models rely on flat data retrieval, our system integrates heterogeneous data sources through a more interconnected, multi-dimensional structure.[42] This architecture is defined by three core components:

(i) Master medical library: Rather than looking at one spreadsheet, this multi-modal knowledge base connects multiple sources into one map. This map includes details on specific drugs, how drugs interact with each other, and which medications are used for which illnesses.
(ii) Hybrid retrieval mechanism: As a "smart" search engine, the system does not just look for keywords; it understands context. It uses a two-step check: first, it finds the most relevant medical information (semantic search), and then it double-checks those results against a network of known drug interactions to ensure safety.
(iii) Agentic pipeline: This agentic architecture serves as the primary reasoning and orchestration engine of the system. An orchestration layer listens to the user's question, decides which data to look up, verifies the facts, and then writes a clear, helpful response, all while running on a secure, local server to keep data private.

By leveraging this architecture, this project aims to demonstrate at a prototype level that a Graph-Enhanced RAG architecture can navigate the complexities of polypharmacy using validated reference datasets, even though the system itself has not undergone clinical validation. In this exploratory phase, the goal is not to deliver a deployable clinical tool (no patient data has been used), but to test whether graph-structured retrieval can meaningfully organize and surface medication-interaction knowledge that possibly helps lay groundwork for future, rigorously evaluated systems.

2. Methodology

A prototype Graph-Enhanced RAG architecture was designed to integrate heterogeneous pharmacological data with LLM-based reasoning for medication decision support. The method of development is divided into three stages described below: curated data source integration, dual-path entity recognition, and graph-driven safety verification. This configuration combines structured relational knowledge with semantic search to then explore polypharmacy scenarios.

2.1. Graph-Enhanced RAG Architecture

Our system architecture consists of three primary components, as illustrated in Figure 1 and detailed in the subsequent sections:

(i) Data sources (section 2.2): We leverage a multi-modal pharmacology knowledge base that integrates structured drug databases, complex drug-drug interaction networks, and precise medication-indication mappings.
(ii) Entity recognition and query processing (section 2.3): To ensure accuracy, the system employs a dual-path retrieval mechanism that combines vector-based semantic search for nuanced context with graph-based interaction checking for relational integrity.
(iii) Safety verification through graph lookup (section 2.4): The final stage features an autonomous pipeline that manages query interpretation, multi-source retrieval, and response synthesis, all powered by a locally hosted LLM to maintain data privacy and clinical relevance. However, while local execution prevents external data exposure, it does not inherently guarantee end-to-end HIPAA compliance, which relies on broader organizational and infrastructure safeguards.

The architecture (a) integrates structured pharmacologic data so it is clear where the data came from, (b) with semantic retrieval methods that can be logged and reviewed, and (c) an agent-driven reasoning process that exposes the sequence of operations rather than producing a black-box answer. Overall, this method enables a transparent and verifiable foundation for addressing complex medication-related questions

It is important to note in Figure 1 that while the Knowledge Base stores raw pharmacological data, the system transitions from a data archive into an active mediator via the Hybrid Retrieval Mechanism. This mechanism operationalizes mediation by executing parallel vector-based semantic searches and graph-based network checks to contextually reconcile disjointed data sources.

The adaptive capacity of this system is driven by the agentic reasoning loop. Rather than returning static database outputs, the local LLM dynamically adapts to the clinical context of the user query through a continuous loop of intent deconstruction, retrieval coordination, and iterative synthesis.

2.2. Data sources

To support robust medication reasoning, our system integrates three complementary data sources that

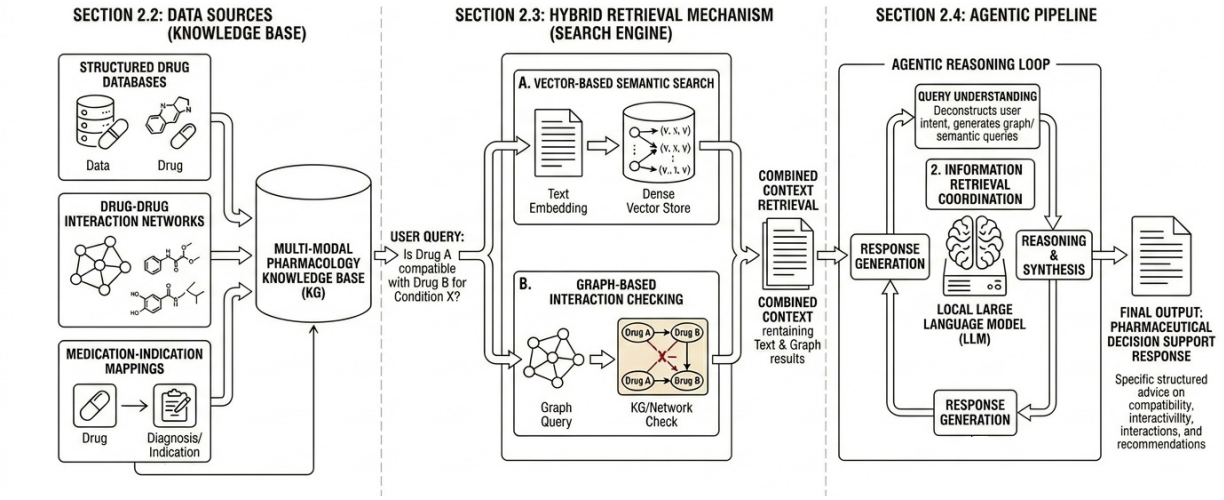

Figure 1. Graph-Enhanced Retrieval-Augmented Generation (Graph-Enhanced RAG) system for pharmaceutical decision support

collectively capture drug properties, clinical indications, and real-world interaction patterns. Each source contributes a distinct layer of pharmacologic and clinical context: a structured baseline drug compendium, a large-scale polypharmacy interaction network, and a comprehensive medication-indication mapping resource. Together, these datasets provide the multimodal foundation required for accurate retrieval, cross-referencing, and inference across diverse medication-related queries.

(i) The baseline drug database—adapted from the USP Drug Classification[43] system—offers standardized pharmaceutical attributes essential for drug-level grounding.
(ii) The Decagon polypharmacy network contributes population-scale evidence of drug-drug interaction risks [44], enabling the system to reason over adverse effect profiles.
(iii) The MEDI-C resource adds clinically validated indication mappings that align medications with diagnostic terminology.

By integrating these heterogeneous but complementary datasets, the system can support more complete, clinically relevant medication guidance across a wide range of use cases.

2.2.1. Polypharmacy interaction graph

The foundation of our safety system is the publicly available Decagon polypharmacy side effects dataset [45] containing over 4.6 million drug-drug interaction pairs derived from clinical data. While other systems such as the U.S. Food and Drug Administration (FDA) Adverse Event Monitoring System (AEMS)[46] or DrugBank[47] can be used, the Decagon system Python software was highly amenable to our prototype development effort. The Decagon dataset represents a comprehensive multimodal graph of drug interactions where each edge between two drugs indicates a specific polypharmacy side effect. The Decagon dataset models 645 unique chemical compounds and 63,473 distinct drug pair combinations, with each interaction annotated with both the polypharmacy side effect code and the corresponding clinical manifestation name. Drug entities in this dataset are represented using an open-source chemical interaction database (STITCH identifiers[48]), which were mapped to standard drug nomenclature for system integration.

We constructed the interaction network as an undirected graph G = (V, E) using the NetworkX library, where vertices V correspond to individual drugs and edges E encode documented adverse interactions. Each edge carries attributes specifying both the polypharmacy side effect code and the clinical manifestation (*e.g.*, "increased bleeding risk" or "serotonin syndrome"). The graph structure enables constant-time lookup of interactions between any two drugs, providing the deterministic safety checking that forms the core of our system's reliability.

A critical implementation detail involved handling STITCH identifier variants. Some drugs in the Decagon dataset have multiple ID numbers because the database encodes different stereo-isomer versions of the same molecule using big numerical shifts. That means two IDs that look very different may actually refer to the same drug (*e.g.*, CID100003033 and CID000003033 represent the

same base molecule). To resolve this, we convert all such IDs to a single shared base number through a modulo-based normalization process, ensuring that the system can always find the right drug regardless of which specific STITCH encoding appears in user queries.

2.2.2. Baseline drug database

For detailed pharmaceutical information beyond interaction data, we integrated a nursing-focused curated database of more than 2,000 medications, organized according to the USP Drug Classification system.[43] Each record included standard pharmaceutical identifiers, therapeutic indications, mechanisms of action, and documented adverse effects, curated to support downstream retrieval tasks. This database serves two purposes: (i) it supplies contextual information that enriches system responses beyond simple safety warnings, and (ii) it provides the drug name standardization layer that bridges between patient-facing terminology and the technical identifiers used in the interaction graph.

2.2.3. Medication indication mappings

To enable treatment recommendation queries, we incorporated the MEDI-C (MEDication Indication Combined) resource,[26] an ensemble dataset aggregating medication-indication relationships from authoritative sources including RxNorm, MedlinePlus, SIDER, Mayo Clinic, and WebMD. MEDI-C contains 223,620 medication-indication pairs spanning 3,837 unique medications and 9,626 clinical conditions, with each indication mapped to standardized ICD-9-CM and ICD-10-CM diagnostic codes.

MEDI-C provides both a complete dataset and a high-precision subset (HPS) derived from consensus across multiple source databases. The HPS achieves 94% precision while maintaining 95% recall, offering a quality-filtered subset that prioritizes reliability for safety-critical applications. Our system preserves this distinction, enabling preferential weighting of high-confidence recommendations in treatment suggestion contexts.

2.3. Entity recognition and query processing

To facilitate robust drug-drug interaction analysis and semantic information retrieval, we developed a multi-layered knowledge base integrating relational, graph-based, and vector data. This construction process focuses on bridging the gap between raw clinical datasets and computational models by transforming heterogeneous drug information into standardized, searchable representations. The architecture comprises three primary pillars:

(i) Structural graph modeling: Mapping the complex web of polypharmacy interactions using graph theory to represent chemical compounds and their clinical manifestations.
(ii) Semantic vectorization: Utilizing transformer-based embeddings to capture the nuanced textual descriptions of drug indications and mechanisms.
(iii) Entity resolution: Implementing cross-database mapping to ensure unified interoperability between disparate identifier systems like STITCH and RXCUI.

By combining these approaches, the knowledge base provides a comprehensive framework for identifying potential adverse interactions while maintaining the contextual richness of clinical drug data.

2.3.1. Drug name extraction

Accurate entity extraction is essential for mapping patient queries to the structured identifiers used in our knowledge base. We implemented a hybrid approach that balances coverage with precision. The system first performs word-level matching against a lowercase drug name dictionary, identifying single-word drug mentions that can be directly resolved. This handles the most common cases efficiently.

However, many pharmaceutical compounds use multi-word names (*e.g.*, "Metformin Hydrochloride," "Calcium Carbonate") that naïve tokenization would miss. To address this, we augment word-level matching with a pattern-based search that targets multi-word drugs using regular expressions with word-boundary constraints. This two-stage approach operates on a longest-match-first basis: the system first attempts to match multi-word drug names, then falls back to single-word matching for any unresolved tokens. This prevents "Metformin Hydrochloride" from being incorrectly split into separate "Metformin" and "Hydrochloride" entities.

Salt variants present a particular challenge in pharmaceutical nomenclature. A single active ingredient may appear under multiple names differing only in their salt form (*e.g.*, "Metformin," "Metformin Hydrochloride," "Metformin ER"). Our entity resolution layer treats these as equivalent for interaction checking, implementing a prefix-matching heuristic: any drug name that begins with a recognized base name is mapped to the same underlying compound identifier. This design choice prioritizes safety by ensuring that interaction warnings trigger regardless of which specific salt formulation a user mentions.

2.3.2. Medical condition recognition

For queries about treatment options, the system extracts medical conditions using a substring matching approach. User queries are normalized to lowercase and searched against a condition index derived from MEDI-

C's standardized indication terminology. To maximize specificity, conditions are sorted by length in descending order before matching, ensuring that "type 2 diabetes" is recognized as a complete entity rather than incorrectly matching only to "diabetes."

The choice to utilize a length-sorted, verbatim string-matching index anchored to standardized resources like MEDI-C (which aggregates RxNorm data) and the Decagon graph was an intentional design trade-off prioritizing high precision over recall: the system will only identify conditions that appear verbatim in its knowledge base. While utilizing advanced biomedical NLP models (such as neural NER pipelines or dedicated RxNorm/DrugBank concept linkers) would broaden coverage, our current design strictly prioritizes the mitigation of false-positive entity resolutions that could otherwise synthesize inappropriate medication suggestions in a safety-critical workflow. The integration of contextual biomedical NER models remains a primary objective for future architectural iterations.

2.4. Safety verification through graph lookup

When a user query contains multiple drug entities, the system performs exhaustive pairwise interaction checking against the Decagon graph. For each pair of drugs (d_1, d_2), we query the graph for the existence of an edge and retrieve any associated polypharmacy side-effect annotations. This lookup is deterministic: if an edge exists in the graph, the interaction warning will always be surfaced, regardless of how the query is phrased.

The graph-based approach provides several advantages over keyword-based safety checking. First, it operates on standardized chemical identifiers rather than free-text drug names, eliminating ambiguity from synonyms or alternative nomenclatures. Second, the interaction data are grounded in clinically observed polypharmacy effects rather than theoretical mechanisms, providing empirically validated warnings. Third, the graph structure enables efficient scaling: checking N drugs for interactions requires $O(N^2)$ graph lookups, which remains tractable even for complex polypharmacy regimens.

When interactions are detected, the system extracts the specific polypharmacy side effect annotation from the graph edge and includes it in a structured safety warning. This ensures that users receive not just a generic "these drugs interact" message, but specific information about the nature of the risk (*e.g.*, "increased risk of gastrointestinal bleeding").

2.5. Response generation and safety guardrails

We generate the system's responses using a small Microsoft language model called Phi-3-mini. It was chosen because it is good at medical reasoning but still light enough to run efficiently on local hardware. We load the model in 16-bit floating-point precision using the Hugging Face transformers library, requiring approximately 7.6GB of GPU memory. When producing text, we set the randomness very low (temperature 0.1) so the model gives stable, consistent answers while still sounding natural.

The language model receives a stru"ture' prompt that clearly delineates retrieved safety information from general pharmaceutical knowledge. This separation is critical as it helps distinguish between retrieved medical facts and LLM-generated responses: rather than asking the LLM to "know" drug interactions, we provide interaction warnings as factual context that the model must incorporate into its response. The prompt structure includes three sections: (i) drug-drug interaction warnings retrieved from the graph, formatted as explicit safety alerts; (ii) relevant drug information from the clinical database; and (iii) treatment options from MEDI-C with their indication mappings and confidence levels.

A critical architectural decision is to implement a hard safety guardrail that operates outside the language model. When graph lookup detects any interaction, the system automatically prepends a formatted safety warning to the model's response. This warning is generated deterministically from the retrieved graph data and cannot be suppressed or softened by the LLM. This design ensures that safety alerts reach the user even if the language model generates an incomplete or incorrect response, providing defense-in-depth against potential model failures.

The guardrail message follows a fixed template: "*** CRITICAL SAFETY WARNING *** The safety database indicates a specific interaction risk: [interaction details]." By making this warning visually distinct and explicitly framing it as derived from database retrieval rather than model inference, we communicate both the source and certainty of the safety information to users, ensuring a strict, inviolable boundary between retrieved medical facts and LLM-generated text. In other words, if a drug-drug interaction is found in the Decagon graph, the system automatically creates the message that cannot be suppressed or softened by the LLM.

2.6. System evaluation framework

We evaluated system performance along two dimensions critical for clinical decision support: safety interaction detection and treatment recommendation accuracy. For each dimension, we designed test sets that reflect realistic usage patterns while enabling systematic performance measurement.

2.6.1. Safety interaction evaluation

The safety evaluation assessed whether the system correctly identifies and flags documented drug-drug interactions. We constructed test cases by randomly sampling edges from the Decagon interaction graph, converting the STITCH identifiers at each edge endpoint back to standard drug names, and formulating natural language queries of the form "Can I take [drug A] and [drug B]?"

This methodology ensures that every test case corresponds to a known interaction that should trigger a safety warning. We evaluated system performance on 100 such cases across three independent runs with different random seeds to assess consistency. For each test query, we measured two metrics: retrieval recall (did the system detect that an interaction exists?) and guardrail compliance (did the safety warning appear in the final response to the user?).

The distinction between these metrics is important. Retrieval recall measures whether the underlying graph lookup successfully identified the interaction—a test of our entity extraction and graph querying logic. Guardrail compliance measures whether that detected interaction resulted in a visible warning to the user—a test of our end-to-end pipeline including the safety prepending mechanism. In a properly functioning system, these metrics should be identical, since detection should always result in a warning being generated.

2.6.2. Treatment recommendation evaluation

The treatment evaluation assessed whether the system can correctly map medical conditions to appropriate medication options. We sampled 100 conditions from MEDI-C's indication vocabulary and formulated queries asking "What is the treatment for [condition]?" For each query, we evaluated whether the system: (1) correctly identified the condition entity in the query, and (2) retrieved at least one medication actually associated with that condition in the MEDI-C database.

This evaluation is more lenient than requiring perfect retrieval of all treatments. We count a response as successful if it includes any valid treatment option, reflecting the practical requirement that the system surface at least one reasonable therapeutic approach. This design acknowledges that for many conditions, multiple treatment options exist and suggests that any evidence-based option provides value even if it is not the complete set of possibilities.

2.7. Computational infrastructure

Software development and all experiments were conducted on an NVIDIA A100 GPU (40 GB) with 16 CPU cores and 128 GB RAM. Our AI model uses the Microsoft Phi-3-mini model, an open-source small language model (SLM) designed for high performance for mobile phones and resource-constrained environments. We stored everything locally so the AI does not need the internet to function. This ensures we meet strict privacy standards—a critical requirement for deployment in clinical environments where patient privacy regulations may restrict network access. With a response time of 3–5 s, the system is fast enough for smooth use, balancing quick conversation with thorough safety checks. End-to-end query processing latency ranged from 3 to 5 s, which is acceptable for conversational interaction.

3. Results

Preliminary results for pilot software runs provide insight on drug-drug interaction detection, treatment recommendation accuracy, and qualitative behavior at knowledge boundaries. Results show high reliability in surfacing documented interaction risks and more variable performance in condition-based treatment retrieval, offering a clear view of current strengths and remaining limitations. In addition to quantitative performance metrics, this section emphasizes qualitative system behavior, as these examples most clearly illustrate how the platform operates in real clinical-style interactions.

The qualitative cases In section 3.3 reveal how safety warnings are triggered, how the system handles uncertainty, and where its design principles succeed or strain in practice. These examples ground the numerical results in concrete, interpretable system behavior.

3.1. Safety interaction detection performance

Our primary evaluation focused on the system's ability to correctly identify and surface drug-drug interactions, as this is the platform's core safety-critical function. Across 100 test cases involving documented drug interaction pairs, the system achieved a safety recall of 95.0% ± 2.4% over three independent runs (Table 2). This indicates that when two drugs with a known interaction are mentioned in a query, the system successfully detects and warns about the interaction in 95 of 100 cases.

The 5% failure rate warrants examination. Analysis of failure cases revealed two primary causes. First, approximately 3% of failures were due to drug name variants that our entity extraction did not map to the canonical forms used in the interaction graph. For example, certain brand names or uncommon generic formulations were not present in our drug name dictionary, leading to those mentions being missed during entity recognition. Second,

Table 2. System performance on safety and treatment recommendation tasks

Evaluation task	Samples per run	Runs	Performance metric	Score
Safety interaction detection	100	3	Retrieval recall	95.0% ± 2.4%
Safety interaction detection	100	3	Guardrail compliance	95.0% ± 2.4%
Treatment recommendation	100	3	Accuracy	75.0% ± 2.2%

approximately 2% of failures occurred when multi-word drug names were partially matched, for instance, with "Calcium Carbonate" being recognized only as "Calcium" and thus not mapping to the correct graph node.

Importantly, we observed no false negatives due to graph lookup failures or guardrail implementation errors. In every case where the entity extraction successfully identified both drugs, the graph correctly returned the interaction if one existed, and the safety warning was properly prepended to the response. This confirms that the core retrieval and safety enforcement mechanisms operate deterministically as designed.

Guardrail compliance matched retrieval recall at 95.0% ± 2.4%, confirming that detected interactions consistently translated into visible user warnings. The variance across runs (±2.4%) reflects the stochastic sampling of test pairs rather than system inconsistency; when tested on identical queries, the system produced identical outputs deterministically.

3.2. Treatment recommendation performance

For treatment recommendation queries, the system achieved 75.0% ± 2.2% accuracy over 100 test conditions (Table 2). This metric measures whether the system correctly identified the medical condition in the query and retrieved at least one valid treatment option associated with that condition in MEDI-C.

The 25% failure rate breaks down into two categories. Approximately 15% of failures occurred when the system could not identify the condition entity at all, primarily cases where the user's phrasing did not contain an exact substring match to MEDI-C's standardized condition terminology. For example, a query about "stomach pain" would not match the formal indication "dyspepsia" without semantic understanding beyond our current substring matching. The remaining 10% of failures occurred when condition extraction succeeded but no treatments were retrieved, typically due to edge cases in the MEDI-C data where certain conditions had unusual or incomplete mappings.

The 75% success rate—although lower than the safety-recall benchmark—is still a promising result for a system that depends on exact string matching rather than advanced language understanding. It shows that people tend to describe medical conditions using terms close to clinical language, which allows our simple substring-matching method to capture a large share of real-world queries. The small variation across runs (±2.2%) also suggests that the system behaves consistently and that its limitations come mainly from the entity-recognition method itself, not from unstable retrieval. Unstable retrieval means a system returns different results for the same input due to randomness or inconsistent indexing. The low ±2.2% variance shows that this system is stable, as its limits are due to simple entity-matching rather than unpredictable retrieval.

3.3. Qualitative system examples

To illustrate the system's real-world performance, we designed an evaluation framework using three distinct test cases, each tailored to a specific system test:

(i) Safety interaction detection: Assessing the system's guardrails and risk mitigation.
(ii) Treatment recommendations: Evaluating clinical accuracy and logic.
(iii) Knowledge boundaries: Testing system behavior and constraints when data are unavailable.

Driven by actual system outputs, these scenarios provide a transparent look at both current capabilities and inherent limitations of the prototype system.

3.3.1. Successful safety interaction detection

Consider a patient being treated for two different co-existing illnesses such as high blood pressure paired with depression. Clinical data shows that roughly 20% to 30% of patients with hypertension also suffer from comorbid depression.[49] This physiological link is well-documented: chronic stress and depression elevate sympathetic nervous system activity (the "fight or flight" response) which directly drives up blood pressure. Consequently, managing a patient who is concurrently being treated for both high blood pressure and depression is a daily reality in clinical practice. Our sample query regarding irbesartan and

mirtazapine perfectly illustrates this common intersection, evaluating the potential interaction risks between a first-line angiotensin receptor blocker (ARB) and a widely prescribed tetracyclic antidepressant. While it is safe to take these drugs under medical supervision, combining them may cause a temporary drop in blood pressure, leading to dizziness or a headache when standing up too quickly.

Example 1. Interaction risk test

Query: "Can I take irbesartan and mirtazapine?"

System Response:
*** CRITICAL SAFETY WARNING ***
The safety database indicates a specific interaction risk:
[RISK] irbesartan + mirtazapine -> bruxism ----------
As a clinical assistant, I must inform you that there is a safety alert regarding the combination of irbesartan and mirtazapine. The combination of these two medications has been associated with an increased risk of bruxism, which is a condition where you grind, gnash, or clench your teeth. This can lead to dental problems and jaw pain. It's important to consult with your healthcare provider before taking these medications together to discuss the potential risks and alternative treatment options.

This example demonstrates the core safety architecture functioning as designed. The graph lookup successfully identified the documented bruxism risk between irbesartan and mirtazapine triggering the automatic safety warning prepended to the response. The language model then synthesized a contextual explanation of the specific risk without diminishing the warning's prominence. In this case the model picks up the chronic antidepressant-linked bruxism but did not emphasize the risk of temporary orthostatic hypotension, a possible fall risk if the patient is not under observation. Cross-referencing this case with the Drugs.com interaction tool [50] revealed a discrepancy: the reference database highlights a primary risk of synergistic blood pressure reduction, an indication that was absent from our syste"s generated output. This omission highlights a specific area for future optimization in the system's retrieval or synthesis capabilities.

The next query simulates polypharmacy confusion and accidental overconsumption across both OTC and prescription medication boundaries. The clinical underpinnings represent two real-world scenarios: an acute setting where a patient stacks multiple OTC antipyretics to break a severe fever, and a chronic care setting where a patient treats back pain with daily OTC ibuprofen and acetaminophen while simultaneously taking a cardiologist-prescribed daily low-dose aspirin.[51–53] We are quick to note that our system does not flag a clinically important situation of a severe cumulative gastrointestinal bleeding risk and the competitive inhibition mechanism where ibuprofen negates aspirin's cardioprotective efficacy. Indeed, the query successfully flags ibuprofen and acetaminophen risks but for the wrong reasons. Another key aspect of this query highlights an important and intentional design boundary: because Decagon encodes only binary drug–drug relationships, the system correctly reports that it cannot evaluate higher-order combinations or emergent risks arising from three-way or multi-drug regimens. This behavior reflects the underlying graph topology rather than a computational shortcoming, ensuring that the system remains transparent about what its evidence base can and cannot support.

Example 2. Interaction test leading to a referral

Query: "Is it safe to take aspirin, ibuprofen, and acetaminophen at the same time?"

System Response:
*** CRITICAL SAFETY WARNING ***
The safety database indicates a specific interaction risk:
[RISK] ibuprofen + acetaminophen -> hypernatremia
[RISK] ibuprofen + acetaminophen -> fistula
[RISK] ibuprofen + acetaminophen -> patent ductus arteriosis
[RISK] ibuprofen + acetaminophen -> head ache ----------
As a clinical assistant, I must inform you that combining ibuprofen and acetaminophen can lead to serious health risks, including hypernatremia, fistula, patent ductus arteriosus, and headaches. While aspirin is a different medication, it's important to consult with a healthcare provider before taking any combination of these drugs to ensure safety and avoid potential complications.

Again, this example reveals both a capability and a limitation. The system successfully detected multiple documented interaction risks between ibuprofen and acetaminophen, retrieving four distinct polypharmacy side effects from the graph. However, the current implementation only checks pairwise interactions—aspirin was mentioned in the query but not evaluated against the other drugs. While the system defers to professional guidance for the three-drug combination, it simultaneously generates critical clinical errors. Specifically, the inclusion of 'fistula' and 'patent ductus arteriosus' as drug side effects is fundamentally incorrect. These erroneous outputs provide a clear, unambiguous indication that default datasets require rigorous, expert curation before they can be safely deployed.

3.3.2. Knowledge boundary recognition

In this polypharmacy case, we evaluate the system's ability to navigate data gaps during medication screening. By introducing unrecognized drug names, we test whether the model prioritizes safety via explicit warnings or risks patient health through silent failure and hallucination. This test of our system involves what might be considered a remote drug combination, based on sevoflurane (used to induce and maintain general anesthesia during surgical procedures) and a patient's existing use of linezolid (a broad-spectrum antibiotic) in the perioperative setting. While alternative antibiotics are more common, this query effectively models a true "needle-in-a-haystack" safety event that might be overlooked during a hurried preoperative chart review. Still, the question during a pre-op evaluation for a patient currently on linezolid therapy remains: Are there any contraindications or specific risks associated with using sevoflurane for general anesthesia?

Example 3. Boundary recognition handling

Query: "Can I take sevoflurane and linezolid together?"

System Response:
*** CRITICAL SAFETY WARNING ***
The safety database indicates a specific interaction risk:
[WARNING] UNKNOWN DRUG: Cannot check linezolid or sevoflurane (Not in database).

As a clinical assistant, I must inform you that there are safety warnings regarding the use of linezolid and sevoflurane. Linezolid is an antibiotic, and sevoflurane is an inhalational anesthetic agent. They are not typically used together in a clinical setting. Sevoflurane is used for anesthesia during surgical procedures, while linezolid is used to treat certain bacterial infections. It is crucial to consult with a healthcare professional before combining any medications, as they can interact in ways that may be harmful. The database indicates that these drugs are not in the database, which suggests that they are not commonly used together, and there may not be established safety information available.

This example demonstrates appropriate uncertainty handling. When the entity extraction failed to map the drug names to identifiers in the interaction graph, the system explicitly flagged this as an unknown drug warning rather than proceeding silently or fabricating information. The language model provided general context about the medications while appropriately emphasizing the need for professional consultation. This behavior is confirmed with Drugs.com database[54] and reflects the current system's design principle of failing safely—when definitive safety data cannot be retrieved, the response defers to human expertise rather than offering potentially incorrect reassurance.

Our next case examines the system's response to "mixed-domain" queries, where a recognized medication is paired with a fictional substance. We explore whether the safety architecture can detect nonsensical inputs or if it defaults to standard language model generation, potentially bypassing the rigorous interaction checks required for clinical decision support systems.

Example 4. Preventing creative hallucinations

Query: "Can I take aspirin and unobtainium together?"

System Response: As a clinical assistant, I must inform you that taking aspirin and unobtainium together is not recommended. Aspirin is a common medication used to reduce pain, fever, and inflammation. It can also be used in low doses to prevent blood clots. Unobtainium, on the other hand, is not a recognized substance in medical literature or pharmacology. It appears to be a fictional or hypothetical element. Given that unobtainium is not a known substance, there are no established safety warnings or interactions to consider. However, it is crucial to consult with a healthcare professional before combining any medications or substances, even if they are not recognized in current medical databases.

This case reveals an interesting system behavior: when one drug is recognized (aspirin) but the other is completely outside the pharmaceutical domain (a fictional element), no safety warning is triggered because the graph lookup returns no results. The language model then generates a response based on its training data, correctly identifying "unobtainium" as fictional. While this response is reasonable, it highlights that the system's safety guardrails only activate when both drugs are identifiable entities—queries mixing real and nonsensical drug names fall back to unconstrained language model generation. This outcome provides a roadmap for refinement due to a hidden structural vulnerability. If a user types a real, obscure drug that the system mistakes for nonsense, the safety guardrails will completely fail, leaving the user entirely unprotected from LLM hallucinations. It suggests that the system needs a more robust upstream Intent Classifier that recognizes a "medical safety query" regardless of whether the entities exist in the local graph. In addition, this illustrates that a""reasonabl"" output (identifying fiction) can still stem from a "flawed" process (bypassing safety checks) and underscores the need for guardrails that govern the reasoning path, not just the final answer.

3.3.3. Treatment recommendation responses

Of interest is the system's ability to integrate structured

database retrieval with narrative synthesis during treatment exploration. The following query might be a general concern for a newly diagnosed type 2 diabetes patient interested in a breakdown of therapy options or even a medical resident looking for a quick reference on guidelines. The query tests whether the model prioritizes specific MEDI-C mappings or reverts to generalized training data, testing the balance between retrieved clinical facts and broad medical knowledge.

Example 5. Treatment recommendation

Query: "What medications can I take for type 2 diabetes?"

System Response: As a clinical assistant, I would respond to the user's query by first acknowledging their condition and then providing information based on the database facts. Thank you for reaching out with your concern about managing type 2 diabetes. It's important to consult with your healthcare provider for personalized advice, but I can share some general information about medications commonly used for this condition. Metformin is often the first medication prescribed for type 2 diabetes. It helps lower blood sugar levels by improving the way your body handles insulin. However, it's crucial to monitor your kidney function while on this medication, as it can affect renal health. If metformin alone isn't sufficient, your doctor might consider adding a sulfonylurea, which stimulates the pancreas to release more insulin. Remember, these medications can increase the risk of hypoglycemia, so it's essential to eat regular meals and monitor your blood sugar levels. Another option could be a DPP-4 inhibitor, which helps your body produce more insulin and decreases the amount of glucose your liver makes. These medications are generally well-tolerated but can cause mild side effects like headaches or nausea. [Response continues with SGLT2 inhibitors and GLP-1 receptor agonistI.]

This response illustrates the system's behavior when the MEDI-C condition lookup succeeds but the retrieved medication list is not prominently displayed in the structured format. The language model generated a comprehensive overview of diabetes medication classes with their mechanisms and safety considerations, drawing from its training data rather than exclusively from the retrieved MEDI-C mappings. While medically accurate and helpful, this response pattern differs from the intended design where specific medications from the database would be explicitly listed before the LLM's synthesis. This represents an area where stricter prompt engineering could better enforce the separation between retrieved facts and general medical knowledge.

Example 6. Treatment recommendation

Query: "What should I take for high blood pressure?"

System Response: As a clinical assistant, I would advise that managing high blood pressure often involves lifestyle changes and may include medication. However, it's crucial to consult with a healthcare provider for personalized advice. Common medications for high blood pressure include ACE inhibitors, ARBs, beta-blockers, calcium channel blockers, and diuretics. Safety warnings include potential side effects such as dizziness, electrolyte imbalances, and kidney function changes. Always follow your doctor's prescription and report any adverse effects.

This more concise response demonstrates the system's handling of a general treatment query. The language model provided medication classes commonly used for hypertension without retrieving specific drug names from MEDI-C. The response appropriately emphasizes consultation with healthcare providers and mentions general safety considerations, though it lacks the specific medication-indication mappings that structured retrieval would provide.

3.3.4. Observations on system behavior

These examples reveal several patterns in system performance. Safety interaction detection functions reliably when both drugs are successfully extracted and mapped to graph identifiers: the deterministic graph lookup and automatic warning prepending ensure that documented interactions consistently surface. However, entity extraction remains a critical dependency: drugs with non-standard names, special characters, or chemical nomenclature may fail to map correctly, resulting in "unknown drug" warnings even when interactions might exist under alternative names. Treatment recommendation responses show greater variability, with the language model often drawing on its training data to provide medically sound general guidance even when structured database retrieval does not execute as designed. While this fallback behavior maintains response helpfulness, it obscures the distinction between retrieved facts and model knowledge, a transparency issue that could be addressed through stricter prompt constraints or response templating in future iterations.

3.4. Performance consistency across evaluation runs

We conducted three independent evaluation runs with different random test case samples to assess system stability. Table 2 presents results aggregated across runs, with standard deviations indicating performance variation. The low standard deviations—2.4% for safety metrics and 2.2% for treatment—demonstrate that the system's performance is consistent rather than dependent on specific test cases.

This consistency is a desirable property for clinical decision support systems, where reliability matters as much as average-case performance. The deterministic nature of our graph-based retrieval and rule-based entity extraction means that identical queries produce identical results, and performance variation across test sets reflects the difficulty distribution of queries rather than system randomness.

3.5. Failure mode analysis

To understand the practical limitations of our system, we manually reviewed failed test cases to categorize error types. For safety interaction detection, the dominant failure mode was incomplete entity extraction, cases where drug mentioned in queries were not successfully mapped to knowledge base entries. This occurred most frequently with:

(i) Brand names do not present in our drug name dictionary.
(ii) Informal drug references (*e.g.*, "blood thinner" instead of "warfarin").
(iii) Multi-word drugs where only part of the name matched (*e.g.*, "Magnesium" from "Magnesium Oxide").
(iv) Misspellings or non-standard capitalizations that broke exact string matching.

Notably, we did not identify cases where the system incorrectly claimed two drugs were safe to combine when a documented interaction existed. The system generated false negatives but zero false positives (it never falsely claimed a dangerous combination was safe). This asymmetry demonstrates that the architecture successfully prioritizes conservative clinical safety, preferring a defensive "unknown drug" warning over a potentially catastrophic, hallucinated reassurance. Missing an interaction due to entity extraction failure is problematic, but incorrectly reassuring a user about a dangerous combination would be far worse. Our architecture achieves this by surfacing warnings whenever the graph contains an interaction but remaining silent (deferring to LLM-generated general guidance) when drugs cannot be definitively identified.

For treatment recommendations, the primary failure mode was mismatch between user terminology and standardized medical condition names in MEDI-C. Patients often describe symptoms ("my joints hurt") rather than diagnoses ("osteoarthritis"), and our substring matching approach cannot bridge this semantic gap without more sophisticated natural language understanding. This limitation is acceptable for a pilot system but represents a clear target for improvement in future iterations.

3.6. Comparison to design objectives

Our evaluation demonstrates that the system meets its core design objectives with respect to safety interaction detection. The 95% recall on known interaction pairs indicates that the graph-based retrieval approach successfully identifies documented polypharmacy risks in a majority of realistic queries. The deterministic guardrail mechanism ensures that identified interactions always result in explicit user warnings, providing a reliable safety layer independent of language model behavior.

The 75% accuracy on treatment recommendations, while respectable for a rule-based entity extraction approach, highlights an area where more sophisticated natural language processing could provide substantial value. However, this lower performance is less critical from a safety perspective; failing to suggest a treatment option is less consequential than failing to warn about a dangerous drug interaction. The system's design appropriately prioritizes reliability on safety-critical tasks while accepting more limited performance on recommendation queries.

4. Discussion

Overall, our six case examples support success with the primary technical objective of this pilot effort to develop a Graph-Enhanced RAG system for safe polypharmacy decision support. It seems deterministic graph-based safety mechanisms can mitigate known LLM risks in clinical contexts; this suggests the system has the potential to be involved in future research to study cognitive load in more detail.

In this model, when specific data is missing, AI falls back on its own training to provide general medical advice. This keeps conversations helpful but makes it hard to tell if information is a proven fact or just a guess. This lack of clear sourcing creates safety risks in clinical settings.

Examples 1 and 2 show the system correctly triggers warnings and avoids guessing about complex drug mixes it does not understand. This proves basic safety checks work, but it also highlights that the system still struggles with complicated, multi-drug interaction maps.

When the AI encounters an unknown drug, it issues an explicit warning instead of guessing. By avoiding silent failures or fake reassurances, the system maintains clinical trust. This "honest" failure mode is essential for ensuring providers can rely on the AI's boundaries during medical decision-making.

Mixing real and fictional data can trick safety filters, allowing the AI to bypass standard protections. Even if the final answer is correct, the logic used to get there may be flawed. This proves that safety systems must monitor the AI's reasoning process, not just the final result.

The system ensures safety by separating roles: the knowledge graph acts as the factual authority, while the AI merely synthesizes that data. This structured approach, combined with low-variance outputs, provides the determinism required for regulated clinical environments. Unlike unpredictable, standard AI pipelines, this architecture supports more reproducible, reliable results.

Three implications are of interest from the research outcomes:

(i) We illustrate a paradigm shift from "retrieval" to "mediation" whereby passive repositories become active mediators. Standard clinical systems require humans to look up a drug network, look up a text guideline, and do the synthesis in their heads. We believe the work suggests that a Graph-Enhanced RAG system has the potential synthesizing disjointed multi-modal data into a unified form. We did not validate this clinically, but did identify areas where open-source data repositories have limitations.

(ii) By utilizing a local LLM, we demonstrate that advanced clinical reasoning no longer requires a reliance on cloud-based commercial systems. This is a major implication for healthcare infrastructure.

(iii) Pure LLMs are notorious for hallucinating medical facts. By feeding the LLM a context combined from both text embeddings and explicit graph-based network checks (section 2.3), the prototype implies that the solution to LLM unreliability in medicine can possibly be mitigated by grounding models in strict topological structures (knowledge graphs).

4.1. Limitations

The system's core strength lies in its safety-first approach, prioritizing recall to catch potential risks through a structured knowledge graph. However, significant weaknesses remain. The system relies on rigid string matching, making it unable to handle synonyms, misspellings, or colloquialisms. Furthermore, its scope is restricted to simple pairwise interactions, ignoring crucial factors like dosage, timing, and complex multi-drug effects. Finally, because testing used synthetic queries rather than live clinical workflows, its real-world performance remains unproven.

The Decagon polypharmacy dataset, while large (4.6 million interaction pairs across 645 compounds), was originally published in 2018 and may not reflect the current pharmaceutical landscape. New drugs approved since then—including novel biologics, gene therapies, and recently introduced small molecules—are absent from the interaction graph entirely. This creates a systematic blind spot: the system can only surface interactions for drugs it "knows about," and the failure mode is silent rather than flagged. Importantly, the dataset models only 645 compounds, a small fraction of the thousands of approved medications in clinical use today. This coverage gap is particularly relevant for polypharmacy patients, who are disproportionately likely to be on newer agents for complex chronic conditions.

While the Decagon dataset serves as a vital benchmark for architectural validation, its reliance on 645 compounds does not fully capture the structural diversity of modern therapeutics. Future work will focus on scaling our approach to more expansive repositories—such as the TWOSIDES dataset, the expanded DrugBank database, or raw clinical data from the FDA's Adverse Event Reporting System (FAERS). Utilizing these continuously updated databases will allow us to evaluate how our model handles thousands of active compounds and millions of potential synergistic interactions.

Extending the Decagon framework from pairwise drug-drug interactions to multi-drug scenarios (>2 drugs) is theoretically possible but highly challenging and outside the scope of our exploratory prototype effort. Fundamentally, Decagon's entire architecture is explicitly hardcoded for binary relationships.

The three integrated sources vary considerably in methodological rigor and intended clinical use, which has implications for the reliability of the system's outputs. The MEDI-C resource aggregates indication data from sources ranging from peer-reviewed clinical vocabularies (RxNorm, ICD-coded mappings) to consumer-facing websites (WebMD, Mayo Clinic), which differ substantially in their evidence standards. Although the High-Precision Subset (HPS) mitigates this concern by requiring cross-source consensus (94% precision, 95% recall), the system does not consistently distinguish to the user whether a recommendation derives from the HPS or the broader, lower-confidence dataset. Similarly, the nursing-focused baseline drug database, while clinically curated, is not explicitly linked to an evidence tier or update cycle. Transparently communicating source provenance to the end user would be an important step toward meeting the auditability standards expected in clinical decision support tools.

While this investigation explores the technical feasibility of a Graph-Enhanced RAG architecture, transitioning from a localized prototype to an operational clinical decision support system requires navigating substantial regulatory and ethical boundaries. Because this exploratory study relied exclusively on open-source, de-identified reference datasets without processing actual protected health information (PHI), comprehensive

compliance with HIPAA and European Union General Data Protection Regulation (GDPR) frameworks were deliberately outside the scope of this technical baseline. However, any real-world clinical deployment demands strict data-at-rest and data-in-motion encryption, role-based access controls, and institutional data-sharing agreements. Ethically, the integration of generative LLMs introduces critical risks regarding automation bias (where users might unquestioningly trust automated outputs) and downstream inequities if the training datasets contain hidden clinical or demographic biases. Furthermore, from a regulatory standpoint, deploying this framework as an active mediator of pharmaceutical oversight would classify it as Software as a Medical Device (SaMD). This necessitates rigorous clinical validation trials, risk mitigation audits, and clearance from regulatory bodies such as the FDA or European Medicines Agency (EMA). Future work must bridge this prototype's technical verification with strict compliance engineering, ensuring robust data governance accompanies our structural safety guardrails before integration into live care workflows.

4.2. Future works

Future versions of the system will focus on refining how it understands and classifies medical information. The integration of fuzzy matching, synonym expansion, and specialized clinical NER models would better handle misspellings and informal language. Additionally, a new safety intent classifier could be introduced to detect medical queries even when entity names are unclear, effectively closing gaps where mixed-domain inputs currently bypass safety guardrails.

The implementation of string-matching approach is limited when handling complex typos, varied morphological structures, or phrases outside our current vocabulary indices. Future work involves integration of biomedical NER frameworks (such as scispaCy or BioBERT) and broader ontologies (like DrugBank and RxNorm) in the next phase of our system's evolution.

Beyond simple recognition, the knowledge graph can be expanded to model complex, higher-order drug interactions, including dosage, patient demographics, and timing. To strengthen clinician trust, a key step forward would be to implement stricter provenance enforcement. This involves a clear structural separation in the AI's output, explicitly labeling retrieved facts versus general medical knowledge to ensure full auditability and transparency in every clinical recommendation.

We note that our test of an existing resource such as Drugs.com is effective for retrieving pairwise drug interactions such as the additive hypotensive effect seen with irbesartan and mirtazapine, but that basic lookup structure can be limiting. Our Graph-Enabled RAG system is designed to move beyond static tables, enabling reasoning over multi-drug interactions that may be implied or distributed across the literature. In our case study, the system's failure to flag the primary hypotensive interaction reflects a current gap in how we extract and weight clinical pair data. However, this limitation does not diminish the system's future ability to operate on high-dimensional, patient-specific information in ways that conventional databases cannot.

The practical application of the proposed architecture would need to transition the framework from a standalone text interface into an integrated clinical utility by establishing active linkages to electronic medical record (EMR) systems. In our pilot, focusing strictly on pairwise drug-drug interactions was an intentional design constraint; it provided a clean, structurally sound, and highly validated data environment to test how effectively our Graph-Enhanced RAG pipeline retrieves and reasons over graph structural data. We noted earlier that a comprehensive clinical system must evolve beyond static binary pairs. Integrating a curated repository or a live EMR system represents a vital next step that would allow the architecture to draw higher-order interactions into the decision-making process. By ingesting real-time patient-specific variables such as active medication lists, current laboratory data and active comorbidities, the system can evaluate the contextual, multi-layered risks necessary for true polypharmacy decision support.

5. Conclusion

This study demonstrates that separating structured retrieval from LLM synthesis creates a safer, more deterministic prototype tool. While the system successfully manages uncertainty and prioritizes safety-first warnings, limitations in entity recognition and multi-drug modeling remain. Future development will incorporate advanced fuzzy-matching techniques, tighter provenance enforcement, and formal clinical validation to maintain the AI's transparency and reliability in real-world medical workflows.

Acknowledgments

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.

Funding

Research reported in this publication was internally funded by Case Western Reserve University Department of Biomedical Engineering.

Conflict of interest

The authors declare no conflicts of interest.

Author contributions

Conceptualization: All authors
Formal analysis: Insiya Fatma, Colin K. Drummond
Investigation: All authors
Methodology: Insiya Fatma, Colin K. Drummond
Software: Insiya Fatma
Writing–original draft: All authors
Writing–review & editing: All authors

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data

The datasets presented in this article are publicly available through the citations provided in the manuscript.

References

1. Scarlett W, Young S. Medical Noncompliance: The Most Ignored National Epidemic. *J Am Osteopath Assoc*. 2016;116:554-555.

   doi: 10.7556/jaoa.2016.111

2. Mansoor M, Kumari M, Manalai P. Treatment Nonadherence: An Epidemic Hidden in Plain Sight. *Psychiatric Times*. 2020;37(Issue No. 3). Accessed June 8, 2026. https://www.psychiatrictimes.com/view/treatment-nonadherence-epidemic-hidden-plain-sight

3. Punnapurath S, Vijayakumar P, Platty PL, Krishna S, Thomas T. A study of medication compliance in geriatric patients with chronic illness. *J Family Med Prim Care*. 2021;10(4):1644-1648.

   doi: 10.4103/jfmpc.jfmpc_1302_20

4. Pruitt SD, Khan R, Chaiyakunapruk N, et al. The silent epidemic of non-adherence – insights from the 2024 a:care congress. *BMC Proc*. 2025;19(S10):13.

   doi: 10.1186/s12919-025-00326-4

5. Kleinsinger F. The Unmet Challenge of Medication Nonadherence. *Perm J*. 2018;22:18-033.

   doi: 10.7812/TPP/18-033

6. Watanabe JH, McInnis T, Hirsch JD. Cost of Prescription Drug-Related Morbidity and Mortality. *Ann Pharmacother*. 2018;52(9):829-837.

   doi: 10.1177/1060028018765159

7. Dusetzina SB, Besaw RJ, Whitmore CC, et al. Cost-Related Medication Nonadherence and Desire for Medication Cost Information Among Adults Aged 65 Years and Older in the US in 2022. *JAMA Netw Open*. 2023;6(5):e2314211.

   doi: 10.1001/jamanetworkopen.2023.14211

8. Roy S. 35+ Latest Patient Adherence Statistics for 2025. Dialog Health. February 17, 2025. Accessed May 20, 2026. https://www.dialoghealth.com/post/patient-adherence-statistics

9. Ruiz-Ramos J, Plaza-Diaz A, Roure-i-Nuez C, *et al*. Drug-Related Problems in Elderly Patients Attended to by Emergency Services. *J Clin Med.* 2024;13(1):3.

   doi: 10.3390/jcm13010003

10. BCC Research. *Medication Adherence: Systems, Technologies and Global Markets*. 2021. Accessed January 11, 2026. https://academic-bccresearch-com.eu1.proxy.openathens.net/market-research/pharmaceuticals/medication-adherence-systems-technologies-and-global-markets.html

11. Steinman M. Polypharmacy: time to get beyond numbers. *JAMA Intern Med*. 2016;176(4):482-483.

    doi: 10.1001/jamainternmed.2015.8597

12. Quinn K, Shah N. A dataset quantifying polypharmacy in the United States. *Sci Data*. 2017;4, 170167.

    doi: 10.1038/sdata.2017.167

13. Mardani A, Paal P, Weck C, Jamshed S, Vaismoradi M. Practical Considerations of PRN Medicines Management: An Integrative Systematic Review. *Front Pharmacol*. 2022;13:759998.

    doi: 10.3389/fphar.2022.759998

14. Coughlin SS, Vernon M, Hatzigeorgiou C, George V. Health Literacy, Social Determinants of Health, and Disease Prevention and Control. *J Environ Health Sci*. 2020;6(1):3061.

15. Tija J, Ellington L, Clayton M. Managing Medications During Home Hospice Cancer Care: The Needs of Family Caregivers. *J Pain Symptom Manage*. 2015;50(5):630-641.

    doi: 10.1016/j.jpainsymman.2015.06.005

16. Makowka J, Lau T. AARP Caregivers and Technology: What they need and what they want.". *AARP Project Catalyst, Apr*. Published online 2018. Accessed June 8, 2026. https://imph.org/wp-content/uploads/Caregivers-and-Technology-AARP-2016.pdf

17. Chi NC, Demirir G, Pike KC, Washington K, Oliver DP. Pain Management Concerns From the Hospice Family Caregivers' Perspective. *Am J Hosp Palliat Care*. 2018;35(4):601-611.

    doi: 10.1177/1049909117729477

18. Borkowski AA, Ben-Ari A. Multiagent AI Systems in Health Care: Envisioning Next-Generation Intelligence. *Fed Pract*. 2025;42(5):188-194.

doi: 10.12788/fp.0589

19. Colbert J. Modernizing Chronic Care Management with Artificial Intelligence. Accessed May 20, 2026. https://www.commure.com/blog/modernizing-chronic-care-management-with-artificial-intelligence

20. Xu G, Li X, Chen Y, *et al*. A comprehensive survey of AI agents in healthcare. *J Biomed Inform*. 2026;179:105045.

doi: 10.1016/j.jbi.2026.105045

21. Milanti A, Norman C, Chan DNS, So WKW, Skinner H. eHealth Literacy 3.0: Updating the Norman and Skinner 2006 Model. *J Med Internet Res*. 2025;27:e70112.

doi: 10.2196/70112

22. Norman CD, Skinner HA. eHealth Literacy: Essential Skills for Consumer Health in a Networked World. *J Med Internet Res*. 2006;8(2):e506.

doi: 10.2196/jmir.8.2.e9

23. Abernethy A, Afsar N, Anderson B, Barfield W, Bharel M, Brown J, *et al*. Toward a National Health Digital and Data Architecture: Laying the Foundation for Digital Transformation. *NAM Perspectives*. 2026;3.

doi: 10.31478/202603b

24. Méndez A, Aguirre M, Del Pozo A, García-Navarro T, Torralbo M, Carreño J. From Conversation to Clinical Decision Supported Assistance: A Secure, Personalized Platform for Elder Care: In: *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies*. Setúbal, Portugal: SCITEPRESS - Science and Technology Publications; 2026:146-157.

doi: 10.5220/0014249500004070

25. Vamadevan A, Vijayan V, Cole C, Jayasudha K, Yusuf N, Walker LE. A NASSS framework-guided systematic review and exploratory modelling of digital health interventions for polypharmacy management in older adults. *BMC Geriatr*. 2025;25:1032.

doi: 10.1186/s12877-025-06602-4

26. Fitzpatrick PJ. Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners. *Front Digit Health*. 2023;5:1264780.

doi: 10.3389/fdgth.2023.1264780

27. Chambers C. The potential of placing a digital assistant in patients' homes. *Br J Gen Pract*. 2020;70(690):8-9.

doi: 10.3399/bjgp20X707273

28. Langston EM, Hattakitjamroen V, Hernandez M, *et al*. Exploring artificial intelligence-powered virtual assistants to understand their potential to support older adults' search needs. *Hum Factors Healthcare*. 2025;7:100092.

doi: 10.1016/j.hfh.2025.100092

29. Bailoni T, Dragoni M. A Review on Trustworthiness of Digital Assistants for Personal Healthcare. *ACM Comput Surv*. 2025;57(7):164:1-164:34.

doi: 10.1145/3714999

30. Barsky A. Suing Therapeutic AI Systems for Malpractice. Accessed May 20, 2026. https://www.psychologytoday.com/us/blog/agens-scientiam/202602/suing-therapeutic-ai-systems-for-malpractice

31. Everett J, Jagoda J, Pike J, Pullen S, Sullivan, Sean. Your AI Therapist May Need a Lawyer: Pennsylvania Brings Suit Against Chatbot Developer. JDSupra. Accessed May 20, 2026. https://www.jdsupra.com/legalnews/your-ai-therapist-may-need-a-lawyer-3525513/

32. Johnson C. Ethical considerations regarding AI use in healthcare: How much is too much? The DO. August 29, 2025. Accessed May 20, 2026. https://thedo.osteopathic.org/columns/ethical-considerations-regarding-ai-use-in-healthcare-how-much-is-too-much/

33. Weiner EB, Dankwa-Mullan I, Nelson WA, Hassanpour S. Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice. *PLoS Digit Health*. 2025;4(4):e0000810.

doi: 10.1371/journal.pdig.0000810

34. Adeyemi AH, Wiredu B, Okobi OE, Nebuwa P, Ezeani EI, Alozie AS. Challenges and Strategies in Medication Management for Patients with Multiple Comorbidities. *Cureus*. 17(6):e85992.

doi: 10.7759/cureus.85992

35. Abo El-Enen M, Saad S, Nazmy T. A survey on retrieval-augmentation generation (RAG) models for healthcare applications. *Neural Comput & Applic*. 2025;37(33):28191-28267.

doi: 10.1007/s00521-025-11666-9

36. Neha F, Bhati D, Shukla DK. Retrieval-Augmented Generation (RAG) in Healthcare: A Comprehensive Review. *AI*. 2025;6(9):226.

doi: 10.3390/ai6090226

37. Wang S, Yang H, Liu W. Research on the construction and application of retrieval enhanced generation (RAG) model based on knowledge graph. *Sci Rep*. 2025;15:40425.

doi: 10.1038/s41598-025-21222-z

38. Zhou J, Jiang J, Han Z, Wang Z, Gao X. Streamline automated biomedical discoveries with agentic bioinformatics. *Brief Bioinform*. 2025;26(5):bbaf505.

doi: 10.1093/bib/bbaf505

39. Chen X, Lu G, Zhang H, Wan J. Knowledge graph-enhanced deep learning for pharmaceutical demand forecasting. *Sci Rep*. 2026;16(1):4776.

doi: 10.1038/s41598-026-35113-4

40. Chen J, Deng Q, Zhen P, et al. Knowledge-graph-enhanced multi-scale modeling for drug-drug interaction prediction. *Mol Ther Nucleic Acids*. 2026;37(1):102855.

doi: 10.1016/j.omtn.2026.102855

41. As'ad M, Faran N, Joharji H. AI-Supported Shared Decision-Making (AI-SDM): Conceptual Framework. *JMIR AI*. 2025;4(1):e75866.

doi: 10.2196/75866

42. Brown A, Roman M, Devereux B. A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. *Big Data Cogn Comput*. 2025;9(12):320.

doi: 10.3390/bdcc9120320

43. USP. USP Drug Classification. Accessed April 21, 2026. https://www.usp.org/health-quality-safety/usp-drug-classification-system

44. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*. 2018;34(13):i457-i466.

doi: 10.1093/bioinformatics/bty294

45. Leskovec J, Sosič R. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Trans Intell Syst Technol*. 2017;8(1):1-20.

doi: 10.1145/2898361

46. FDA. Adverse Event Monitoring System (AEMS). March 11, 2026. Accessed May 20, 2026. https://www.fda.gov/drugs/surveillance/fda-adverse-event-monitoring-system-aems

47. National Library of Medicine. DRUGBANK Source Information. Accessed May 20, 2026. https://www.nlm.nih.gov/research/umls/rxnorm/sourcereleasedocs/drugbank.html

48. Kuhn M, von Mering C, Campillos M, Jensen LJ, Bork P. STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res*. 2008;36(Database issue):D684-D688.

doi: 10.1093/nar/gkm795

49. Li S, Yang R, Liu X, Li X. Risk factors for comorbid hypertension and depression in the elderly: Evidence from the China health and retirement longitudinal study. *Medicine*. 2025;104(33):e44004.

doi: 10.1097/MD.0000000000044004

50. Drugs.com. Drug Interaction Report: irbesartan, mirtazapine. Accessed May 28, 2026. https://www.drugs.com/drug-interactions/irbesartan-with-mirtazapine-1383-0-1640-0.html

51. Hermann M, Carstens N, Kvinge L, *et al*. Polypharmacy and Potential Drug–Drug Interactions in Home-Dwelling Older People – A Cross-Sectional Study. *J Multidiscip Healthc*. 2021;14:589-597.

doi: 10.2147/JMDH.S297423

52. Gottlieb S. Cardioprotective effects of aspirin compromised by other NSAIDs. *BMJ*. 2003;327(7414):520.

doi: 10.1136/bmj.327.7414.520

53. Barkin RL. Acetaminophen, aspirin, or Ibuprofen in combination analgesic products. *Am J Ther*. 2001;8(6):433-442.

doi: 10.1097/00045391-200111000-00008

54. Drugs.com. Drug Interaction Report: linezolid, sevoflurane. Accessed May 28, 2026. https://www.drugs.com/interactions-check.php?drug_list=1472-0,2059-0

Appendix

Table A1. Overall system performance (n = 40)

Metric Category	Metric	Score
Intent & Extraction Accuracy	Intent classification accuracy	0.85
	Drug extraction accuracy	0.85
	Condition extraction accuracy	1.00
	Retrieval success rate	0.95
Information Retrieval Quality	Precision (macro)	0.63
	Recall (macro)	0.66
	F1-score (macro)	0.62
	Precision (micro)	0.69
	Recall (micro)	0.50
	F1-score (micro)	0.58
Response Characteristics	Response completeness	1.00
	Average response length (tokens)	931
	Response length standard deviation	352

Notes: Scores represent the proportion of correct outcomes (1.00 = perfect). Intent classification measures whether the system correctly identifies query type (e.g., drug information vs. interaction check vs. treatment recommendation). Retrieval success indicates whether the knowledge base returned relevant information. Precision, recall, and F1 quantify the overlap between system-retrieved and ground-truth relevant items. These metrics were computed using a rule-based entity extraction pipeline with exact string matching; ML-based NER would be expected to improve these scores.

Table A2. Performance by query category

Category	Sample Size	Intent Accuracy	Retrieval Success	Precision	Recall	F1-score
Drug information	10	1.00	1.00	0.50	0.50	0.50
Indication search	10	0.80	0.30	0.00	0.00	0.00
Side effects	10	0.60	1.00	0.50	0.50	0.50

Table A3. Mean performance with 95% confidence intervals

Metric	Mean	Lower CI	Upper CI
Intent classification accuracy	0.80	0.65	0.95
Retrieval success rate	0.77	0.61	0.92
Precision (macro)	0.33	0.16	0.50
Recall (macro)	0.33	0.16	0.50
F1-score (macro)	0.33	0.16	0.50