Ministry of Higher Education and Scientific Research

University of Carthage

Ecole Polytechnique de Tunisie

وزارة التعليم العالي و البحث العلمي

جامعة قرطاج

المدرسة التونسية للتقنيات

Open Science Seminar Project Report

A Deep Dive into Medical AI with Knowledge Graphs: Deep KG-Guided Medical Reasoning

---

15/11/2025 - 24/01/2026

In Collaboration with

***Elaborated by*:**

Hedi KSENTINI

Oumaima WERGHEMMI

Third-Year Engineering Students, SISY

Project Supervised by:

Mr. Kais ZHIOUA CEO, Tanit AI

Mrs. Dhouha BOUCHAALA , Head of SISY department

Academic Year : 2025/2026

Elkhawarezmi BP 743 La Marsa 2078
Phone: 71 774 699 Fax: 71 748 843
web site : www.ept.tn

نهج الخوارزمي ص.ب ٧٤٣ المرسى ٢٠٧٨
الهاتف: 71 774 611 الفاكس: 71 748 843
موقع الواب: www.ept.tn

DEDICATION

To our families, whose unwavering support, patience, and encouragement have been our foundation throughout this challenging research journey. Your belief in us gave us strength during the most demanding phases of this project.

To our parents, siblings, and loved ones.

To healthcare professionals and researchers in reproductive medicine, whose dedication to understanding and treating fertility challenges inspired this work. Your tireless efforts to combine scientific knowledge with compassionate care remind us that medicine is both an art and a science.

To all clinicians, researchers, and medical staff.

To the open-source community and AI pioneers, whose groundbreaking work in knowledge graphs, natural language processing, and machine learning makes innovative research like ours possible. Your commitment to advancing technology for the betterment of humanity lights the path forward.

To the innovators of biomedical AI.

To individuals and couples facing fertility challenges, whose resilience and hope drive the need for better clinical decision support systems. May this work contribute in some small way to making your journey easier and more informed.

To every person navigating reproductive health decisions.

To **all of you**,

we dedicate this work.

Oumaima WERGHEMMI, Hedi KSENTINI

Acknowledgement

We, the engineering students from **Ecole Polytechnique de Tunisie**, extend our heartfelt gratitude to our visionary mentor and guide, **Mr. Kais Zhouia**, whose entrepreneurial insight, strategic direction, and unwavering belief in our mission have been the driving force behind this venture. His expertise in translating complex AI concepts into viable business solutions has been invaluable in shaping Tanit AI's direction and ambition.

We are also deeply grateful to our academic supervisor, **Mrs. Dhouha Bouchaala**, for her exceptional guidance, scholarly rigor, and steadfast support throughout this research-intensive journey. Her deep expertise in artificial intelligence, combined with her commitment to academic excellence and practical innovation, provided the foundational framework that allowed this project to evolve from theoretical exploration to applied implementation.

We acknowledge with gratitude the open-source communities, AI researchers, and medical professionals whose pioneering work in knowledge graphs, natural language processing, and reproductive medicine informed and inspired our technical approach. Your contributions to the fields of biomedical informatics and clinical decision support systems laid the groundwork upon which we have built.

Finally, we acknowledge the resilience and hope of individuals and couples navigating fertility challenges—it is your strength and the dedication of healthcare professionals serving you that motivates our work at Tanit AI. We hope this contribution advances the frontier of AI-assisted clinical reasoning in meaningful and compassionate ways.

TABLE OF CONTENTS

LIST OF FIGURES **vi**

LIST OF TABLES **vii**

GENERAL INTRODUCTION **ix**

1 General Context of the Project **2**

1.1 Presentation of the Strategic Collaboration . . . 3

1.1.1 Tanit Healthcare Technologies . . . 3

1.1.2 The OpenSeminar Program Initiative . . . 4

1.2 Project Context . . . 4

1.2.1 Problem Statement . . . 4

1.2.2 Objectives . . . 5

1.2.3 Work Methodology . . . 5

1.2.4 Tasks Planification . . . 6

2 State of the Art **9**

2.1 Knowledge Graphs in Biomedical Applications . . . 10

2.1.1 Biomedical Knowledge Representation . . . 10

2.1.2 Clinical Decision Support with Knowledge Graphs . . . 10

2.2 Large Language Models in Medicine . . . 11

2.2.1 Biomedical Language Models . . . 11

2.2.2 Limitations of Current LLM Approaches . . . 12

2.3 Integrated Approaches : Combining KGs & LLMs . . . 12

2.3.1 Retrieval-Augmented Generation . . . 12

2.3.2 Chain-of-Thought Reasoning with Knowledge Grounding . . . 12

2.4 Specialized Applications in Reproductive Medicine . . . 13

2.4.1 AI in Fertility Treatment Prediction . . . 13

2.4.2 Drug–Disease Relationship Discovery in Reproductive Health . . . 13

2.5 The MedReason Framework and Related Work . . . 14

2.6 Gaps in Current Research 14
2.6.1 Integration Depth 14
2.6.2 Domain Specialization 14
2.6.3 Explainability and Evidence Tracing 15
2.6.4 Robustness to Knowledge Gaps 15
2.7 Contributions of Our Research 15
2.7.1 Technical Innovations 15
2.7.2 Clinical Relevance 15
2.7.3 Methodological Advances 16

3 Model Conception & Integration **17**
3.1 System Architecture 19
3.1.1 Overall Pipeline Design 19
3.1.2 Component Integration Strategy 19
3.2 Knowledge Graph Design and Implementation 21
3.2.1 Biomedical Node Schema 21
3.2.2 Graph Loading and Property Enhancement 22
3.3 Entity Extraction Module 22
3.3.1 Fertility-Specific Entity Recognition 22
3.3.2 Fallback Pattern Matching 23
3.4 Property-Enhanced Knowledge Graph Mapping 24
3.4.1 Three-Stage Matching Strategy 24
3.4.2 Property Similarity Calculation 25
3.5 Biomedical Path Discovery and Analysis 26
3.5.1 Graph Traversal for Biological Relationships 26
3.5.1.1 Path Discovery Algorithm 27
3.5.1.2 Biomedical Context-Aware Node Representation 27
3.5.1.3 Path Quality Assessment and Selection 27
3.5.1.4 LLM-Enhanced Path Pruning 28
3.5.1.5 Path Representation and Natural Language Generation 28
3.5.1.6 Performance Optimization and Error Handling 29
3.5.1.7 Clinical Integration 29
3.6 Chain-of-Thought Reasoning Generation 30
3.6.1 Biomedical Reasoning Framework 30
3.7 Integration and Orchestration 33
3.7.1 Main Pipeline Orchestrator 33

3.7.2 Configuration Management 34
3.8 System Performance and Robustness Features 35
3.8.1 Error Handling and Fallback Mechanisms 35
3.8.2 Clinical Validation Features 35

4 Deployment & Evaluation **37**
4.1 System Implementation Architecture 38
4.1.1 Development Environment and Tools 38
4.1.2 Code Structure and Organization 39
4.2 Graphical User Interface Development 40
4.2.1 Streamlit-Based Interface Design 40
4.2.2 Interface Components and Layout 40
4.2.3 Interactive Features 41
4.3 Experimental Evaluation Framework 42
4.3.1 Evaluation Dataset 42
4.3.2 Evaluation Metrics 42
4.4 Experimental Results 43
4.4.1 Entity Extraction Performance 43
4.4.2 Entity Mapping Results 44
4.4.3 Path Discovery and Selection Performance 44
4.4.4 Biomedical Pathway Analysis and clinical reasoning chain 44
4.4.5 Final CoT grounded Answer Generation 45
4.5 Technical Limitations 46
4.6 Comparative Analysis 46
4.6.1 Comparison with Baseline Systems 46

GENERAL CONCLUSION **48**

BIBLIOGRAPHY **49**

LIST OF FIGURES

1.1 Logo of the Company Tanit Healthcare Technologies 4

2.1 Clinical AI Application 13

3.1 Clinical reasoning pipeline architecture showing the process from entity extraction to answer generation 19

3.2 Knowledge graph node schema and relationships 21

3.3 Loading data from Neo4j 22

3.4 Three-stage matching strategy with progressive refinement 25

3.5 Three-stage matching strategy with progressive refinement 26

3.6 Orchestrator workflow with error handling and recovery paths 33

4.1 GUI Interface Layout with Key Components 40

4.2 GUI Interface Layout with Key Components 41

4.3 Entity Extraction Example 43

4.4 Entity Mapping Example 44

4.5 Path Discovery and selection 44

4.6 Biomedical Pathway Analysis and clinical reasoning chain 45

4.7 Reasoning Steps(example) 45

4.8 Final answer (example) 45

LIST OF TABLES

3.1 Key System Components and Their Functions 18
3.2 Technology Stack Implementation Details 20
3.3 Fertility Entity Categories and Examples 23
3.4 Matching Score Weights by Property Type 25
3.5 Path Scoring Criteria and Weights 28
3.6 Configuration Categories and Examples 34

4.1 Development Environment Specifications 39
4.2 Comparative Performance Analysis 46

LIST OF ABBREVIATIONS

API Application Programming Interface

BERT Bidirectional Encoder Representations from Transformers

COT Chain-of-Thought

ERNIE Enhanced Representation through Knowledge Integration

KG Knowledge Graph

KGE Knowledge Graph Embedding

LLM Large Language Model

NER Named Entity Recognition

NLP Natural Language Processing

RAG Retrieval-Augmented Generation

UMLS Unified Medical Language System

GENERAL INTRODUCTION

The field of reproductive medicine represents one of the most complex and emotionally significant domains of healthcare, where clinical decision-making requires the synthesis of vast amounts of biomedical knowledge, patient-specific factors, and evolving scientific evidence. Fertility challenges affect millions worldwide, yet the path from diagnosis to effective treatment remains fraught with uncertainty, requiring clinicians to navigate intricate biological relationships, pharmacological interactions, and personalized therapeutic strategies.

In recent years, artificial intelligence has emerged as a transformative force across healthcare, offering unprecedented capabilities in pattern recognition, knowledge synthesis, and predictive modeling. However, the application of AI in specialized clinical domains like reproductive medicine presents unique challenges : the need for domain-specific knowledge representation, the requirement for explainable clinical reasoning, and the ethical imperative to augment rather than replace human clinical judgment.

Current approaches to clinical decision support in fertility medicine often operate in isolation

—structured knowledge bases lack reasoning capabilities, while general-purpose language models lack domain-specific biomedical grounding. This gap creates a significant opportunity : to develop systems that can understand clinical questions in context, map them to structured biomedical knowledge, trace evidence-based relationships, and generate clinically relevant, explainable reasoning.

The central challenge addressed in this project is the development of an integrated AI pipeline that connects natural language understanding of clinical questions with structured biomedical knowledge representation and generates evidence-based, explainable clinical reasoning specifically tailored to fertility and reproductive medicine.

This project embodies Tanit AI's broader vision : to develop AI systems that enhance, rather than replace, human clinical expertise—providing clinicians with intelligent tools that synthesize complex biomedical information, trace evidence-based relationships, and generate explainable reasoning while preserving human judgment and patient-centered care.

By focusing specifically on reproductive medicine, we address a domain where emotional sensitivity, personalized care, and complex decision-making intersect, creating AI systems that are not only technically sophisticated but also clinically meaningful and ethically grounded.

The following chapters detail our technical implementation, experimental validation, and the broader implications of our work for both AI research and clinical practice in reproductive medicine.

Chapter 1

General Context of the Project

Sommaire

1.1 Presentation of the Strategic Collaboration 3
1.1.1 Tanit Healthcare Technologies 3
1.1.2 The OpenSeminar Program Initiative 4
1.2 Project Context 4
1.2.1 Problem Statement 4
1.2.2 Objectives 5
1.2.3 Work Methodology 5
1.2.4 Tasks Planification 6

Introduction

This project emerges from a strategic collaboration between an industry leader in medical artificial intelligence and a prestigious academic institution recognized for engineering excellence. The partnership aims to bridge the gap between cutting-edge industrial research and advanced academic training, with a particular focus on addressing one of the most critical challenges in medical AI : ensuring factual accuracy and reliability.

In high-stakes clinical environments, artificial intelligence systems must not only perform well but also provide transparent, explainable, and verifiably correct outputs. This project is therefore positioned at the intersection of applied research and real-world clinical needs, targeting the development of trustworthy AI systems for medical reasoning.

1.1 Presentation of the Strategic Collaboration

1.1.1 Tanit Healthcare Technologies

Tanit Healthcare Technologies is a pioneering company specializing in AI-driven solutions for reproductive medicine and fertility care. Positioned at the forefront of medical AI innovation, Tanit's core mission is to develop trustworthy, evidence-based clinical decision support systems.

To mitigate the limitations of general-purpose language models, particularly factual hallucinations, Tanit grounds AI reasoning in structured biomedical knowledge. Within this project, the company provides domain expertise, its proprietary Reproductive Medicine Knowledge Graph, benchmark datasets, and real-world clinical problem formulations.

The official logo of the company is shown in Figure 1.1.

FIGURE 1.1 – Logo of the Company Tanit Healthcare Technologies

1.1.2 The OpenSeminar Program Initiative

The OpenSeminar program is a collaborative initiative jointly launched by Tanit Healthcare Technologies and École Polytechnique de Tunisie, with the primary goal of providing students with immersive, hands-on experience in applied medical artificial intelligence. The program is designed to bridge academic learning and industrial practice by engaging participants in carefully scoped, industry-driven projects that address real-world research and development challenges.

A strong emphasis is placed on data-centric AI development, reproducibility, and rigorous evaluation methodologies, reflecting the constraints and standards of professional AI engineering environments. Within this framework, our project, entitled *"Deep KG-Guided Medical Reasoning (MedReason-Style)"*, specifically targets the enhancement of factual accuracy and reasoning transparency in medical AI systems through the integration of structured knowledge graphs into large language model reasoning pipelines.

1.2 Project Context

1.2.1 Problem Statement

While Large Language Models (LLMs) have achieved impressive results in general domains, their application in medical decision-making remains limited by hallucinations and lack of verifiability. In clinical contexts, even minor factual errors can have serious consequences.

This project addresses this limitation by grounding language model reasoning in structured medical knowledge, using Knowledge Graphs (KGs) to ensure transparency, traceability, and factual correctness.

1.2.2 Objectives

The main objective of Project B is to design and evaluate a deep KG-guided medical reasoning system capable of producing accurate and explainable answers. The specific objectives include :

— Implementing a robust entity linking pipeline between natural language queries and KG entities.

— Extracting multi-hop reasoning paths using deep path-finding algorithms such as k-Shortest Paths.

— Adapting or fine-tuning a base LLM to reason over KG-derived subgraphs.

— Benchmarking the system against medical QA datasets using rigorous evaluation metrics.

— Exploring innovative methodological alternatives within the project framework.

1.2.3 Work Methodology

To successfully conduct this project, we adopted an agile-inspired workflow grounded in the principles of the Scrum framework. This methodological choice enabled development in short, iterative cycles, providing the flexibility required to adapt efficiently to evolving technical challenges, which are characteristic of artificial intelligence research and experimental development.

Scrum [?] structures work into fixed-duration iterations known as *sprints*, typically spanning two weeks. Each sprint is oriented toward clearly defined objectives and the delivery of tangible, measurable outcomes. Such a structure promotes adaptive project management by allowing the continuous incorporation of new technical insights, experimental results, and methodological refinements throughout the project lifecycle.

Within this framework, a regular meeting routine was established, including sprint reviews and technical discussions with both academic supervisors and industry engineers. These sessions served to evaluate progress, identify obstacles, and realign priorities when necessary.

The combination of iterative development, structured planning, and frequent cross-disciplinary communication fostered an efficient and responsive workflow. This approach ensured steady progress, facilitated risk mitigation, and supported continuous improvement across all phases of the project.

1.2.4 Tasks Planification

The project was structured over a two-month period, spanning from mid-November 2025 to January 2026, and organized into a set of well-defined phases corresponding to the major milestones of the project. Each phase comprised clearly identified tasks with explicit interdependencies, ensuring a coherent, incremental, and traceable progression from initial research to final delivery.

The identified phases are summarized below :

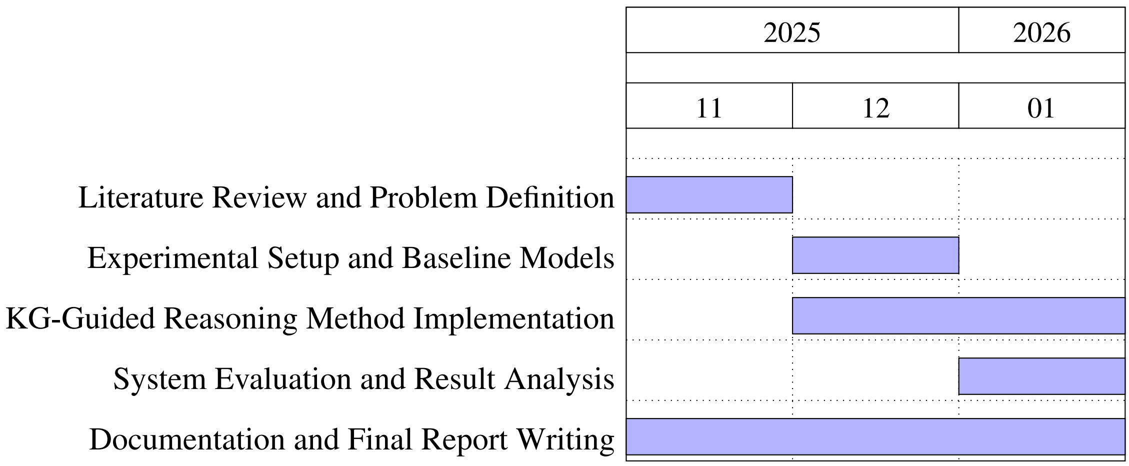
— **Phase 1 – Literature Review and Problem Definition (November) :** This initial phase focused on conducting a comprehensive review of the state of the art in knowledge-enhanced medical reasoning and Large Language Model (LLM) architectures. Key research works such as KG-augmented reasoning frameworks, medical question-answering systems, and hallucination mitigation techniques were analyzed. In parallel, the project scope and problem statement were refined through discussions with academic supervisors and industry mentors, leading to a clear definition of objectives, constraints, and evaluation criteria.

— **Phase 2 – System Design and Data Preparation (December) :** This phase centered on the technical design and implementation of the proposed system. Core components, including the entity linking pipeline, knowledge graph traversal and path-finding mechanisms, and data preprocessing workflows, were developed and validated. Additionally, benchmark datasets and training resources were curated, cleaned, and formatted to ensure compatibility with the evaluation pipelines.

— **Phase 3 – Model Evaluation and Finalization (January) :** The final phase was dedicated to testing the final architecture and integrating it with the knowledge graph reasoning components. Extensive experimental evaluations were conducted using predefined metrics to assess factual accuracy, reasoning quality, and robustness. The phase concluded with the consolidation of results, documentation of findings, and preparation of final deliverables, including the technical report and reproducible code with user interface.

This phased task plan ensured a structured workflow, facilitated risk management, and allowed for continuous validation of intermediate results, ultimately contributing to the successful completion of the project.

The following Gantt Chart outlines this complete planning, including the dependencies between tasks :



Gantt Chart

Conclusion

This chapter outlined the general context of the project, highlighting the collaboration between Tanit Healthcare Technologies and École Polytechnique de Tunisie and their complementary expertise in medical artificial intelligence and knowledge-based systems.

It introduced the main challenge addressed by the project : improving the reliability of Large Language Models in medical applications by reducing factual hallucinations and enhancing explainable reasoning through the use of knowledge graphs.

The project objectives and development methodology were also presented. An agile-inspired workflow based on the Scrum framework was adopted to support iterative progress, regular evaluation, and effective collaboration between academic and industry stakeholders. The work was structured into successive phases covering research, system development, evaluation, and documentation.

This context provides a clear foundation for the technical chapters that follow, which detail the proposed approach, its implementation, and the experimental results.

Chapter 2

State of the Art

Sommaire

2.1 Knowledge Graphs in Biomedical Applications 10
2.1.1 Biomedical Knowledge Representation 10
2.1.2 Clinical Decision Support with Knowledge Graphs 10
2.2 Large Language Models in Medicine 11
2.2.1 Biomedical Language Models 11
2.2.2 Limitations of Current LLM Approaches 12
2.3 Integrated Approaches : Combining KGs & LLMs 12
2.3.1 Retrieval-Augmented Generation 12
2.3.2 Chain-of-Thought Reasoning with Knowledge Grounding 12
2.4 Specialized Applications in Reproductive Medicine 13
2.4.1 AI in Fertility Treatment Prediction 13
2.4.2 Drug–Disease Relationship Discovery in Reproductive Health . 13
2.5 The MedReason Framework and Related Work 14
2.6 Gaps in Current Research 14
2.6.1 Integration Depth 14
2.6.2 Domain Specialization 14
2.6.3 Explainability and Evidence Tracing 15
2.6.4 Robustness to Knowledge Gaps 15
2.7 Contributions of Our Research 15
2.7.1 Technical Innovations 15
2.7.2 Clinical Relevance 15
2.7.3 Methodological Advances 16

INTRODUCTION

The integration of artificial intelligence with clinical decision support systems represents one of the most promising frontiers in healthcare technology. In the specific domain of reproductive medicine, this intersection has gained significant attention due to the complexity of fertility-related decision-making, the emotional sensitivity of patients, and the rapid evolution of treatment protocols. This chapter examines the current landscape of AI applications in clinical decision support, with particular focus on knowledge graph-based systems, large language models in medicine, and specialized applications in reproductive health.

2.1 Knowledge Graphs in Biomedical Applications

Knowledge graphs (KGs) have emerged as powerful tools for representing and reasoning about biomedical knowledge in a structured, interconnected format. In the context of clinical decision support, KGs offer several advantages over traditional databases and ontologies.

2.1.1 Biomedical Knowledge Representation

Recent research has demonstrated the effectiveness of knowledge graphs for integrating heterogeneous biomedical data. Multiple studies have shown that KGs can integrate information from biomedical literature, clinical trial repositories, drug databases such as DrugBank and PubChem, and disease ontologies including SNOMED CT and Mondo. Graph database technologies such as Neo4j have been widely adopted in biomedical applications due to their efficient traversal mechanisms and expressive query language (Cypher), enabling complex relationship-based reasoning.

2.1.2 Clinical Decision Support with Knowledge Graphs

Several studies have explored KG-based clinical decision support systems. Prior work has demonstrated improved treatment recommendation accuracy in chronic disease management and effective identification of drug repurposing opportunities in emerging diseases. These systems typically follow a pipeline consisting of :

1. Entity extraction from clinical or biomedical text
2. Mapping entities to structured knowledge graph nodes
3. Relationship traversal and subgraph extraction
4. Inference generation based on discovered paths

However, many existing systems lack specialization in reproductive medicine and struggle to robustly map free-text clinical questions to structured KG entities, limiting their practical usability in fertility-related decision support.

2.2 Large Language Models in Medicine

The emergence of large language models (LLMs) has significantly advanced natural language processing capabilities in healthcare, enabling improved understanding and generation of clinical text.

2.2.1 Biomedical Language Models

Specialized biomedical LLMs, including BioBERT, ClinicalBERT, and BioMedLM, have demonstrated superior performance on medical text understanding tasks compared to general-purpose language models. These models are pre-trained on biomedical literature and clinical notes, allowing them to better capture medical terminology and domain-specific semantics.

More recently, large-scale medical models such as Med-PaLM 2 have demonstrated strong performance in medical question answering, achieving results comparable to clinicians on certain standardized medical examinations. Despite their capabilities, these models generally function as opaque systems, offering limited transparency regarding their reasoning processes and evidence sources.

2.2.2 Limitations of Current LLM Approaches

Despite their impressive performance, current LLM-based medical systems exhibit several limitations :

— **Hallucination** : Generation of plausible but factually incorrect medical information

— **Lack of evidence tracing** : Inability to explicitly reference supporting medical knowledge

— **Domain specificity gap** : Reduced performance in specialized domains such as reproductive medicine

— **Interpretability challenges** : Limited visibility into intermediate reasoning steps

These limitations raise significant concerns in safety-critical clinical environments.

2.3 Integrated Approaches : Combining KGs & LLMs

To address the weaknesses of standalone approaches, recent research has explored hybrid systems combining structured knowledge graphs with LLM-based reasoning.

2.3.1 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) frameworks enhance LLM responses by incorporating external knowledge sources during inference. In biomedical contexts, knowledge graphs have been increasingly used as structured retrieval sources, allowing models to ground their outputs in verified medical relationships and entities.

2.3.2 Chain-of-Thought Reasoning with Knowledge Grounding

Chain-of-thought prompting has been shown to improve reasoning accuracy and interpretability by encouraging models to generate intermediate reasoning steps. When combined with structured knowledge graph evidence, this approach enables transparent, evidence-based reasoning, which is particularly critical for clinical decision support systems.

2.4 Specialized Applications in Reproductive Medicine

2.4.1 AI in Fertility Treatment Prediction

Several studies have applied machine learning techniques to fertility treatment outcomes, including prediction of in vitro fertilization success rates and ovarian response to stimulation protocols. These systems typically rely on statistical or machine learning models and do not incorporate explicit reasoning or explainability mechanisms.

2.4.2 Drug–Disease Relationship Discovery in Reproductive Health

Automated extraction of drug–disease relationships from biomedical literature has been explored in the context of reproductive health. While these approaches contribute valuable structured knowledge, they often lack interactive question-answering capabilities and clinical reasoning support.

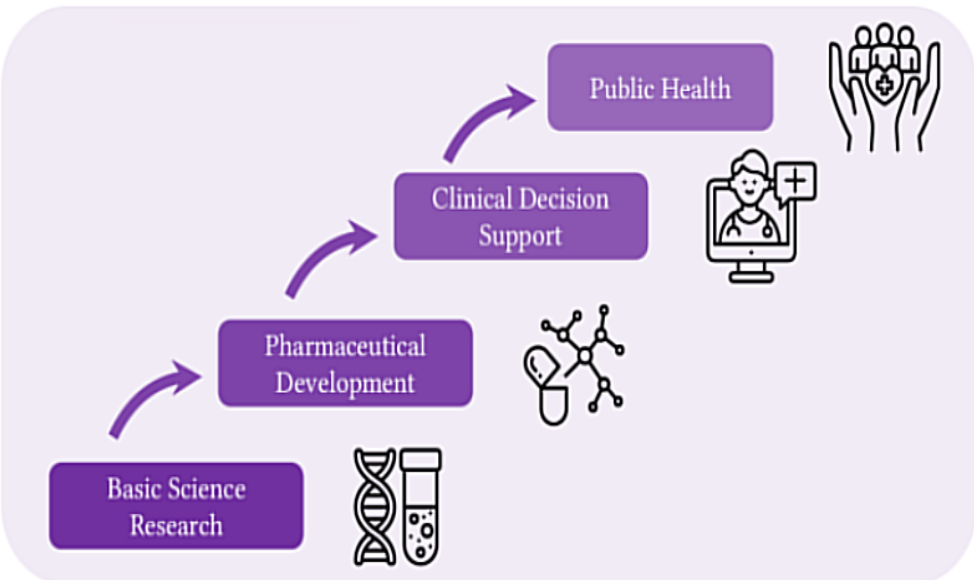


FIGURE 2.1 – Clinical AI Application

2.5 The MedReason Framework and Related Work

The MedReason framework, introduced in recent work on medical reasoning systems, represents a significant advancement in knowledge-grounded clinical reasoning. MedReason integrates :

1. Medical knowledge acquisition from structured and unstructured sources
2. Patient representation through clinical entities and relationships
3. Hybrid reasoning mechanisms combining symbolic paths with neural language models
4. Explanation generation to support clinical decision-making

While MedReason demonstrates strong performance on general medical reasoning benchmarks, it does not specifically target reproductive medicine.

2.6 Gaps in Current Research

Our analysis of the state of the art reveals several key gaps :

2.6.1 Integration Depth

Most existing systems either :

- Use knowledge graphs without advanced natural language interaction
- Use LLMs without structured medical grounding
- Lack tight integration between retrieval, reasoning, and explanation components

2.6.2 Domain Specialization

Few systems are specifically designed for reproductive medicine, lacking :

- Fertility-specific entity schemas and property weighting
- Domain-focused relationship types and reasoning patterns
- Evaluation metrics aligned with reproductive health decision-making

2.6.3 Explainability and Evidence Tracing

Many systems fail to provide :

- Clear mappings from clinical questions to structured evidence
- Step-by-step reasoning grounded in medical knowledge
- Transparency regarding uncertainty and evidence limitations

2.6.4 Robustness to Knowledge Gaps

Existing approaches rarely address :

- Incomplete or sparse structured knowledge
- Graceful degradation when knowledge graph evidence is unavailable
- Explicit communication of reasoning limitations

2.7 Contributions of Our Research

2.7.1 Technical Innovations

- Multi-stage, property-aware entity matching with fertility-specific weighting
- Integrated pipeline combining knowledge graph retrieval, path pruning, and LLM reasoning
- Fertility-specialized prompt engineering for clinical reasoning
- Timeout-aware reasoning with graceful fallback to direct LLM responses

2.7.2 Clinical Relevance

- Focused modeling of reproductive medicine entities and relationships
- Property schemas optimized for fertility-related diseases and drugs
- Reasoning patterns aligned with fertility specialist clinical workflows

2.7.3 Methodological Advances

— Balanced integration of structured knowledge certainty and neural language flexibility

— Transparent, evidence-traceable clinical reasoning

— Practical system design grounded in real-world fertility-related clinical scenarios

CONCLUSION

The state of the art in clinical decision support systems demonstrates substantial progress in knowledge graphs, large language models, and medical AI. However, fully integrated, explainable, and domain-specialized systems for reproductive medicine remain limited. This research addresses these gaps by proposing a comprehensive, fertility-focused reasoning pipeline that combines structured biomedical knowledge with advanced natural language reasoning, offering improved accuracy, transparency, and clinical relevance.

The following chapters present the methodology, system implementation, and experimental evaluation of the proposed approach.

Chapter

3

Model Conception & Integration

Sommaire

3.1 System Architecture 19

3.1.1 Overall Pipeline Design 19

3.1.2 Component Integration Strategy 19

3.2 Knowledge Graph Design and Implementation 21

3.2.1 Biomedical Node Schema 21

3.2.2 Graph Loading and Property Enhancement 22

3.3 Entity Extraction Module 22

3.3.1 Fertility-Specific Entity Recognition 22

3.3.2 Fallback Pattern Matching 23

3.4 Property-Enhanced Knowledge Graph Mapping 24

3.4.1 Three-Stage Matching Strategy 24

3.4.2 Property Similarity Calculation 25

3.5 Biomedical Path Discovery and Analysis 26

3.5.1 Graph Traversal for Biological Relationships 26

3.6 Chain-of-Thought Reasoning Generation 30

3.6.1 Biomedical Reasoning Framework 30

3.7 Integration and Orchestration 33

3.7.1 Main Pipeline Orchestrator 33

3.7.2 Configuration Management 34

3.8 System Performance and Robustness Features 35

3.8.1 Error Handling and Fallback Mechanisms 35

3.8.2 Clinical Validation Features 35

INTRODUCTION

This chapter presents the comprehensive design and implementation of a clinical knowledge reasoning system that integrates multiple artificial intelligence techniques for biomedical question answering. The system combines knowledge graph retrieval, natural language processing, and large language models (LLMs) to analyze fertility-related clinical questions. Our approach addresses three critical challenges in clinical informatics : (1) accurate entity extraction from unstructured medical text, (2) precise mapping to biomedical knowledge graphs, and (3) generating clinically-reasoned answers through path-based inference.

TABLE 3.1 – Key System Components and Their Functions

Component	**Primary Function**	**Fertility-Specific Enhancement**
Entity Extractor	Identify biomedical entities	Fertility terminology prioritization
Graph Mapper	Match entities to KG nodes	Property-aware fertility scoring
Path Discoverer	Find biological relationships	Fertility-relevant path filtering
Reasoning Engine	Generate chain-of-thought	Clinical implication extraction
Fallback Handler	Ensure system reliability	Context-aware answer generation

The proposed system follows a multi-stage pipeline that begins with fertility-specific entity extraction, proceeds through property-enhanced knowledge graph matching, performs biomedical path discovery between entities, and culminates in chain-of-thought reasoning generation. This integrated approach enables the system to bridge the gap between unstructured clinical questions and structured biomedical knowledge, providing transparent and evidence-based answers.

3.1 System Architecture

3.1.1 Overall Pipeline Design

The system implements a five-stage pipeline as illustrated in Figure 3.1 :

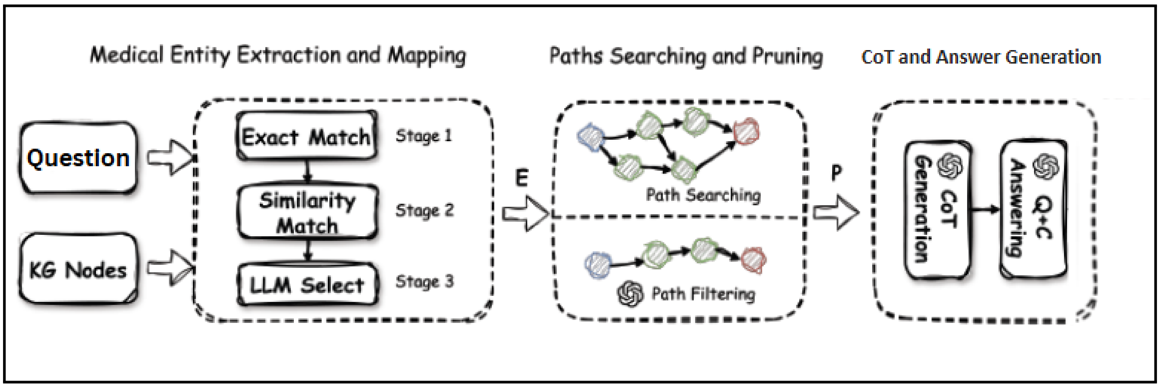


FIGURE 3.1 – Clinical reasoning pipeline architecture showing the process from entity extraction to answer generation

1. **Entity Extraction** : LLM-based identification of fertility-related biomedical entities from clinical questions
2. **Knowledge Graph Mapping** : Property-aware matching of entities to nodes in the biomedical knowledge graph
3. **Path Discovery** : Graph traversal to find biological relationships between mapped entities
4. **Path Pruning** : LLM-based selection of the most clinically relevant paths
5. **Reasoning Generation** : Chain-of-thought reasoning and final answer generation

3.1.2 Component Integration Strategy

The system integrates multiple components through a carefully designed orchestration layer :

— **Neo4j Graph Database** : Stores biomedical knowledge with disease and drug entities

— **Sentence Transformers** : Provides semantic embedding capabilities for similarity matching

— **Groq LLM API** : Powers entity extraction, path pruning, and reasoning generation

— **Custom Python Orchestrator** : Coordinates all components with timeout handling and fallback mechanisms

TABLE 3.2 – Technology Stack Implementation Details

Layer	**Technology**	**Version**	**Purpose**
Database	Neo4j	5.x	Graph storage and traversal
Embeddings	Sentence-Transformers	2.2.2	Semantic similarity
LLM Interface	Groq API	Latest	Natural language processing
Orchestration	Python	3.9+	Pipeline coordination

3.2 Knowledge Graph Design and Implementation

3.2.1 Biomedical Node Schema

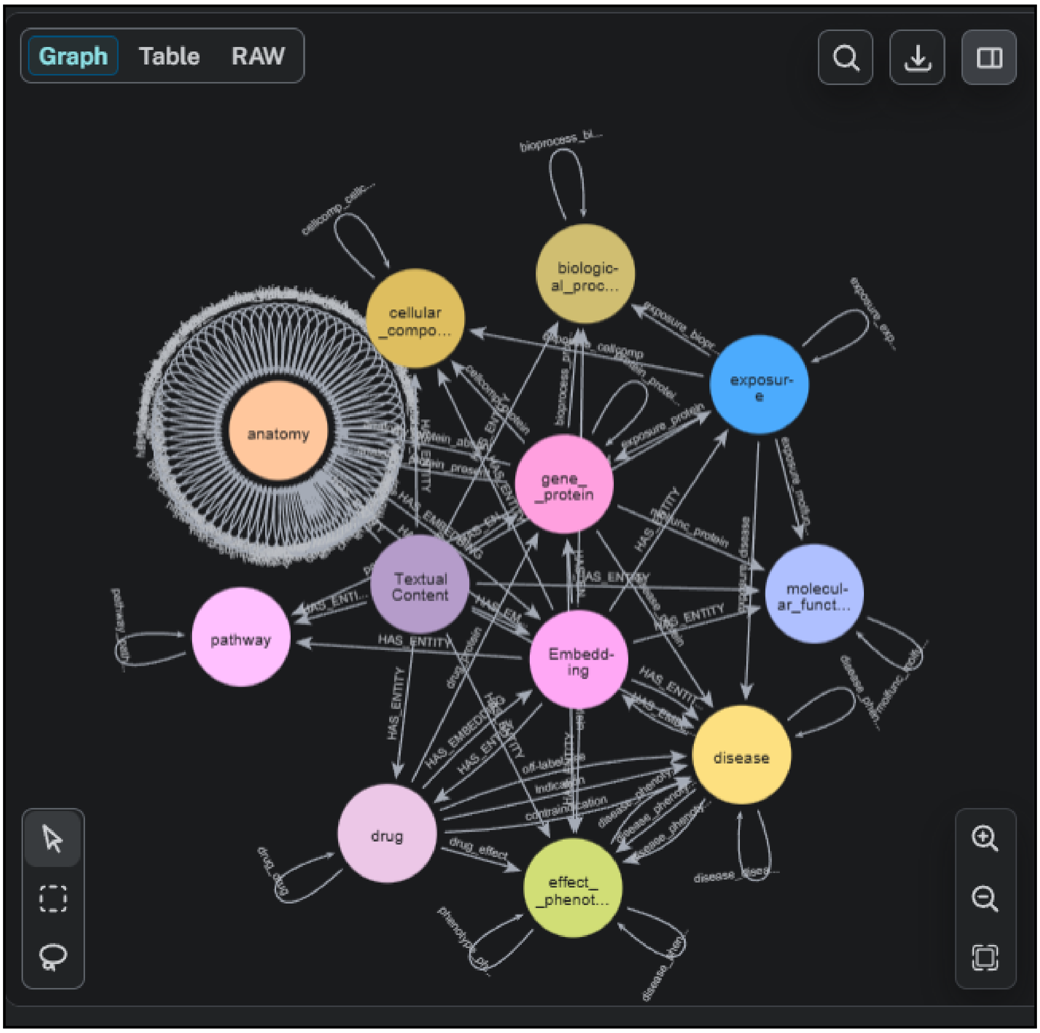


FIGURE 3.2 – Knowledge graph node schema and relationships

The knowledge graph implements a specialized schema for fertility-related biomedical entities, with primary focus on disease and drug nodes capturing comprehensive clinical descriptions, symptoms, epidemiology, and management guidelines. Drug nodes include pharmacological properties such as mechanism of action, indications, pharmacokinetics, and therapeutic categories.

Each node type follows a standardized property hierarchy : primary identifiers (node_id, node_name), clinical descriptors (definitions, symptoms, mechanisms), and metadata (sources, labels). This schema enables efficient property-based matching while maintaining the richness of biomedical information necessary for clinical reasoning.

3.2.2 Graph Loading and Property Enhancement

The knowledge graph loader implements intelligent property extraction and text enhancement specifically optimized for biomedical content. During loading, properties from multiple sources are concatenated into searchable text fields, with prioritization given to clinically relevant properties. The system applies domain-specific heuristics to infer entity types from labels and property patterns, ensuring accurate categorization of fertility-related entities.

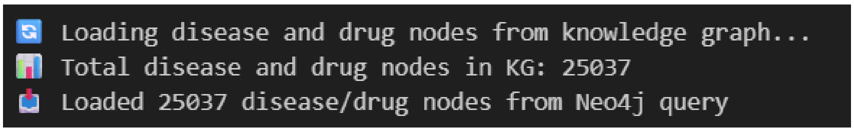


FIGURE 3.3 – Loading data from Neo4j

Property text enhancement follows clinical relevance weighting : disease descriptions and symptoms receive higher priority than administrative metadata, while drug mechanisms and indications are emphasized over physicochemical properties. This enhancement strategy improves embedding quality for similarity searches and ensures that clinically significant information drives matching decisions. The system maintains statistics on label distributions and property completeness for quality monitoring.

3.3 Entity Extraction Module

3.3.1 Fertility-Specific Entity Recognition

The entity extraction module employs a hybrid approach combining LLM-based recognition with domain-specific pattern matching, specifically optimized for fertility and reproductive medicine contexts. The LLM component is guided by detailed prompts that emphasize fertility-related terminology, reproductive health conditions, and fertility treatments, with examples drawn from clinical guidelines and medical literature.

TABLE 3.3 – Fertility Entity Categories and Examples

Category	Example Entities	Clinical Context
Female Conditions	PCOS, Endometriosis, Fibroids	Ovarian function, Uterine health
Male Conditions	Azoospermia, Varicocele, Low motility	Sperm quality, Testicular function
Treatments	Clomiphene, Gonadotropins, IVF	Ovulation induction, ART procedures
Hormones	FSH, LH, Estradiol, Progesterone	Endocrine regulation
Procedures	IUI, ICSI, Egg retrieval	Assisted reproduction techniques

The recognition process prioritizes entities with clear fertility relevance while maintaining capability to identify general biomedical entities. Confidence scoring incorporates both linguistic certainty and clinical relevance factors, with higher weights given to entities explicitly mentioned in fertility contexts. The system implements context-aware filtering that considers surrounding text for disambiguation, particularly important for terms with multiple meanings in different medical specialties.

3.3.2 Fallback Pattern Matching

When LLM-based extraction encounters difficulties, the system employs comprehensive pattern matching using regular expressions specifically designed for fertility terminology. This fallback mechanism covers over fertility-related diseases and fertility medications with precise pattern definitions that capture clinical variations and synonyms. The pattern library includes conditions such as polycystic ovary syndrome, endometriosis, male factor infertility, and their pharmacological treatments.

Pattern matching incorporates contextual analysis to distinguish fertility-related mentions from general medical terminology. The system examines surrounding text for fertility indicators and adjusts confidence scores accordingly. This dual-layer approach ensures robust entity extraction even with challenging clinical texts or when LLM services experience limitations, maintaining system reliability across diverse input conditions.

3.4 Property-Enhanced Knowledge Graph Mapping

3.4.1 Three-Stage Matching Strategy

Entity-to-knowledge graph mapping follows a sophisticated three-stage strategy that progressively refines matching accuracy.

Stage 1 performs exact property matching against primary keys and display properties, leveraging the biomedical schema for precision matching. This stage identifies unambiguous matches based on exact string equality with biomedical terminology.

Stage 2 implements property-enhanced similarity matching that combines semantic embeddings with schema-aware property weights. This stage calculates similarity scores that consider both textual similarity and biomedical property relevance, with enhanced scoring that gives priority to clinically significant properties.

Stage 3 employs LLM-guided contextual matching that evaluates candidate nodes within the specific clinical context of the question, incorporating fertility relevance as a decisive factor.

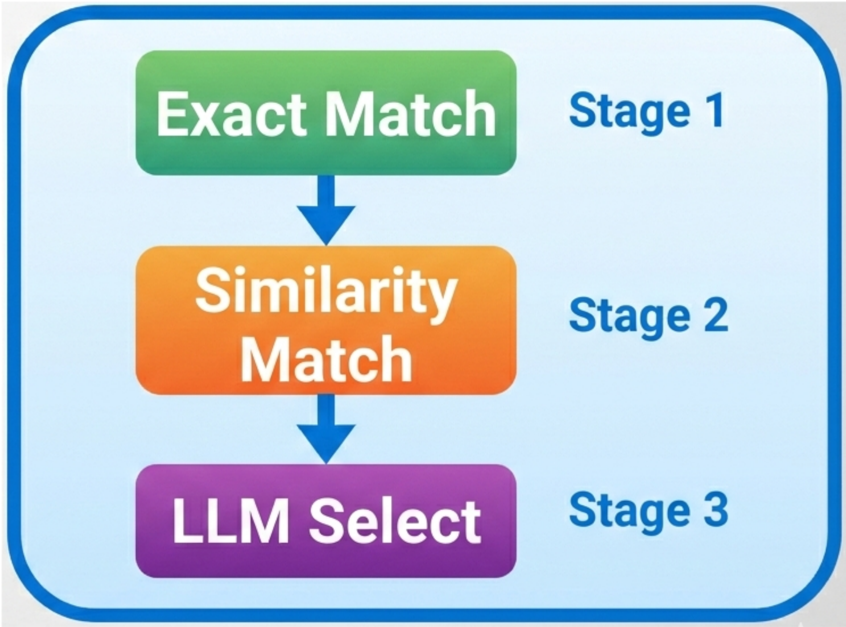


FIGURE 3.4 – Three-stage matching strategy with progressive refinement

3.4.2 Property Similarity Calculation

The property similarity engine implements a weighted scoring system that reflects clinical importance within fertility medicine. Primary key matches receive the highest weights, followed by display property matches, with general property matches receiving baseline weights. The system incorporates name-based matching with separate weights for exact and partial matches, recognizing the clinical importance of precise nomenclature in biomedical contexts.

TABLE 3.4 – Matching Score Weights by Property Type

Property Category	Exact Match	Partial Match	Semantic Similarity
Primary Identifier	1.0	0.8	0.6
Clinical Name	0.9	0.7	0.75
Synonyms	0.8	0.6	0.65
Symptoms/Indications	0.7	0.5	0.7
General Description	0.6	0.4	0.6
Fertility Context Bonus	+0.15	+0.10	+0.12

Similarity calculation combines cosine similarity of text embeddings with property-based boosting factors. Clinical properties relevant to fertility (such as symptoms for diseases and mechanisms for drugs) receive higher boosting factors. The final enhanced score represents a clinically informed balance between textual similarity and biomedical property alignment, optimized for fertility domain accuracy. Thresholds for match acceptance are dynamically adjusted based on match type and clinical context.

3.5 Biomedical Path Discovery and Analysis

3.5.1 Graph Traversal for Biological Relationships

The knowledge graph traversal component serves as the core mechanism for discovering biological relationships between biomedical entities. This subsystem implements a sophisticated path-finding algorithm that navigates through the biomedical knowledge graph to identify potential connections between disease and drug entities extracted from clinical questions.

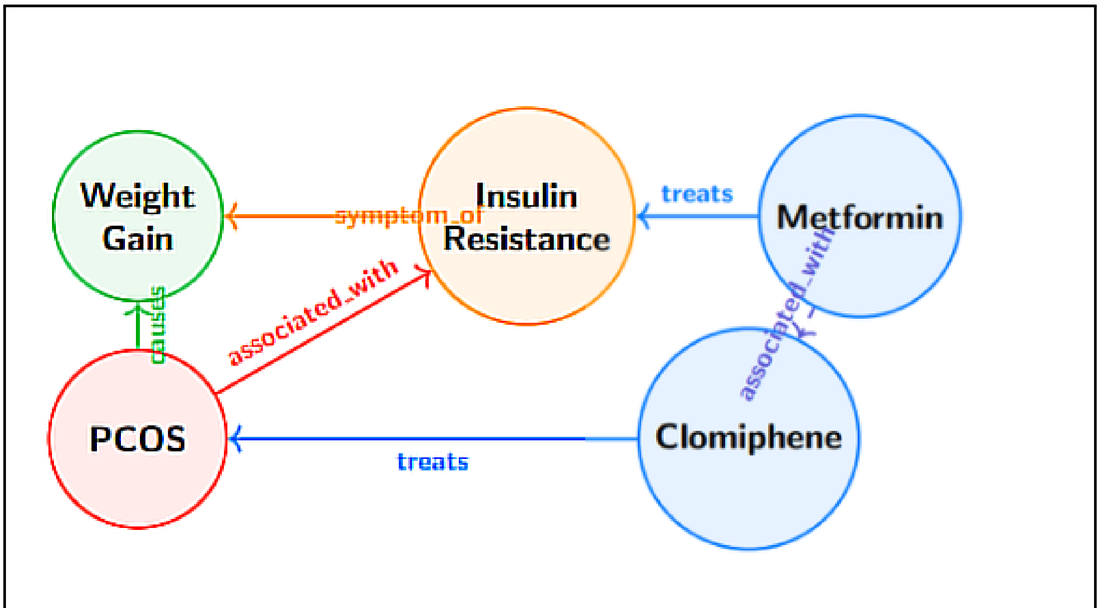


FIGURE 3.5 – Three-stage matching strategy with progressive refinement

3.5.1.1 Path Discovery Algorithm

The path discovery process employs a breadth-first search strategy constrained to a maximum of four relationship hops (MAX_PATH_LENGTH $= 4$), balancing computational efficiency with sufficient exploration depth for most biomedical relationships. The algorithm uses the following Cypher query pattern in Neo4j :

```
MATCH path = (a)-[*1..4]-(b)
WHERE elementId(a) = $source_id AND elementId(b) = $target_id
RETURN path, length(path) as pathLength
ORDER BY pathLength ASC
LIMIT 20
```

This query identifies all undirected paths between source and target nodes, returning a maximum of 20 paths sorted by ascending length. The path-finding process includes intelligent node property extraction, prioritizing biomedical-relevant properties for meaningful path representation.

3.5.1.2 Biomedical Context-Aware Node Representation

Each node in discovered paths is enriched with biomedical context through a multi-tiered property extraction strategy :

1. **Primary Biomedical Properties** : Disease-specific properties (e.g., `SNOMEDCT_US_definition`, `mondo_definitions`) and drug-specific properties (e.g., `indication`, `mechanism_of_action`) are prioritized.
2. **Common Properties** : Shared properties (`node_id`, `node_name`, `node_source`) provide standardized identifiers.

3.5.1.3 Path Quality Assessment and Selection

The system implements a multi-faceted path scoring mechanism to identify the most clinically relevant connections :

TABLE 3.5 – Path Scoring Criteria and Weights

Criteria	Weight	Description
Entity Coverage	0.4	Percentage of question entities present in path
Path Length	0.3	Inverse relationship to path length
Biomedical Relevance	0.3	Presence of fertility/biomedical terms in node labels

The entity coverage metric calculates the proportion of mapped clinical entities that appear in each path, prioritizing paths that incorporate all entities mentioned in the clinical question.

3.5.1.4 LLM-Enhanced Path Pruning

For high-stakes clinical reasoning, the system employs a Large Language Model (Llama 4) to evaluate and select the most relevant path based on biomedical context :

— **Clinical Relevance Assessment** : LLM evaluates whether each path directly addresses the fertility-related clinical question.

— **Biomedical Plausibility Check** : LLM assesses the biological coherence of relationships in the path.

— **Fertility Context Evaluation** : Specialized attention to fertility-related concepts and terminology.

— **Confidence Scoring** : LLM provides a confidence score (0.0-1.0) for selected paths.

The LLM receives structured information including entity mappings, path descriptions, and biomedical relevance scores, enabling informed clinical judgment. When LLM evaluation fails or times out, a fallback mechanism selects paths based on the scoring criteria in Table 3.5.

3.5.1.5 Path Representation and Natural Language Generation

Discovered paths are converted into human-readable formats through two complementary approaches :

— **Technical Representation** : Path strings follow the pattern :

`NodeName -[RELATIONSHIP_TYPE]-> NextNodeName`, providing machine-readable relationship information.

— **Clinical Description** : Natural language descriptions use biomedical terminology : `Polycystic ovary syndrome (disease) associated_with → Metformin (drug)`.

Relationship types are translated to biomedical language (e.g., `ASSOCIATED_WITH` becomes "is clinically associated with") to enhance clinical interpretability.

3.5.1.6 Performance Optimization and Error Handling

The path discovery system incorporates several optimization strategies :

— **Duplicate Path Elimination** : Path signatures prevent redundant processing of identical connections.

— **Early Termination** : Path length sorting ensures shortest, most direct connections are prioritized.

— **Timeout Protection** : A 15-minute timeout prevents indefinite graph traversal in complex or disconnected subgraphs.

— **Graceful Degradation** : When no paths are found, the system transitions to direct LLM answering rather than failing entirely.

3.5.1.7 Clinical Integration

The biological relationship discovery component integrates seamlessly with the broader clinical reasoning pipeline :

(a) **Input** : Successfully mapped disease and drug entities from clinical questions.
(b) **Processing** : Graph traversal to find biological pathways between entities.
(c) **Evaluation** : Scoring and LLM-based selection of most clinically relevant paths.
(d) **Output** : Selected paths for subsequent Chain-of-Thought reasoning generation.

This approach enables the system to ground clinical answers in verifiable biological relationships while maintaining the flexibility to handle cases where direct graph connections are unavailable, ensuring robust performance across diverse fertility-related clinical questions.

3.6 Chain-of-Thought Reasoning Generation

3.6.1 Biomedical Reasoning Framework

The Chain-of-Thought generation framework implements a structured approach to clinical reasoning based on identified biomedical pathways. The system decomposes the reasoning process into sequential steps : entity identification, relationship analysis, biological mechanism explanation, clinical implication derivation, and final answer formulation. Each step incorporates domain-specific knowledge about fertility medicine and reproductive biology.

The reasoning generation leverages detailed path information, including node properties, relationship types, and biomedical context. The framework translates technical graph structures into clinically meaningful narratives that connect biological mechanisms to clinical implications. Special attention is given to fertility-specific considerations, such as treatment efficacy, side effect profiles, and reproductive outcomes throughout the reasoning process.

Listing 3.1 – Biomedical Chain-of-Thought Prompt Generator

```
import json
def generate_cot_prompt(question, entity_context, path_description,
                        path_nodes_info, relationships_info):

    # Create CoT generation prompt for biomedical context
    cot_prompt = f"""
    CLINICAL QUESTION: "{question}"

    ENTITY MAPPING CONTEXT (question entities mapped to knowledge
    graph):
    {json.dumps(entity_context, indent=2)}

    SELECTED BIOMEDICAL KNOWLEDGE GRAPH PATH:
    Path Description: {path_description}

    Detailed Path Analysis:
    1. PATH NODES (biomedical entities):
    {json.dumps(path_nodes_info, indent=2)}

    2. BIOLOGICAL RELATIONSHIPS:
    {json.dumps(relationships_info, indent=2)}

    TASK: Generate a Chain-of-Thought (CoT) reasoning to answer
    the clinical question based on the biomedical knowledge graph
    path.

    COT STRUCTURE FOR BIOMEDICAL CONTEXT:
    1. CLINICAL ENTITY IDENTIFICATION: Identify which clinical
    entities from the question are present in the path
    2. PATH BIOLOGICAL INTERPRETATION: Explain what the path means
  in biological/medical terms
    3. RELATIONSHIP ANALYSIS: Analyze each biological relationship
    in the context of fertility/medicine
```

```
    """ 4. CLINICAL INFERENCE: Draw logical clinical conclusions from
     the path
     5. ANSWER: Provide a concise clinical answer based on
     the reasoning

     IMPORTANT FOR FERTILITY CONTEXT:
     - Focus on reproductive health implications
     - Consider disease-drug relationships
     - Mention fertility relevance when applicable
     - Base reasoning ONLY on the specific path shown above
     OUTPUT FORMAT:
     {{
       "chain_of_thought": {{
         "reasoning_steps": [
           "Step 1: Identify clinical entities from question that
           appear in the path",
           "Step 2: Explain the first biological relationship in
           medical terms",
           "Step 3: Explain subsequent biological relationships",
           "Step 4: Draw clinical conclusions based on the complete
           path",
           "Step 5: Formulate the clinical answer"
         ],
         "detailed_reasoning": "Multi-paragraph detailed
         explanation connecting the path to the clinical question",
         "final_answer": "Concise clinical answer to the original
         question"
       }}
     }}
     """
    return cot_prompt
```

3.7 Integration and Orchestration

3.7.1 Main Pipeline Orchestrator

The pipeline orchestrator implements sophisticated flow control with comprehensive error handling and recovery mechanisms. It manages the sequential execution of pipeline stages while monitoring for timeouts and resource constraints. The orchestrator implements conditional branching based on intermediate results, enabling dynamic adaptation to processing outcomes.

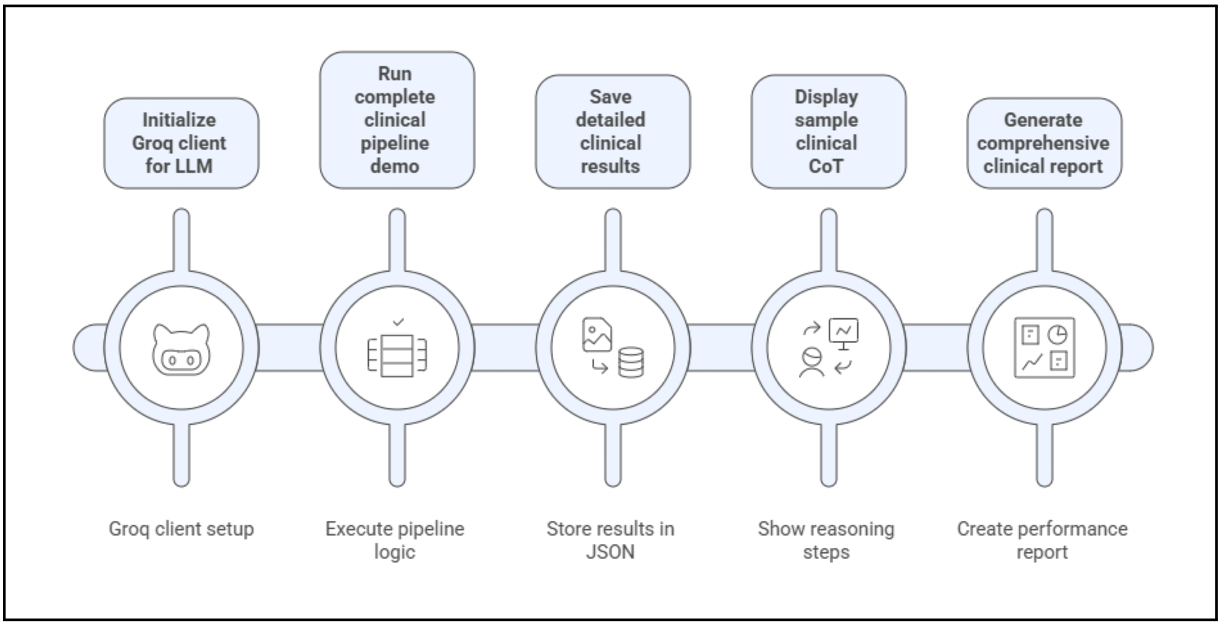


FIGURE 3.6 – Orchestrator workflow with error handling and recovery paths

Execution monitoring includes performance tracking, success rate calculation, and quality metric collection. The orchestrator manages resource allocation between components and implements priority-based processing for time-sensitive operations. This centralized control structure ensures consistent behavior while allowing component-level optimization and independent development.

3.7.2 Configuration Management

System configuration follows a hierarchical structure that separates biomedical domain settings from technical parameters. Domain configurations include entity type definitions, property schemas, fertility terminology lists, and clinical relevance weights. Technical configurations cover model parameters, timeout values, performance thresholds, and resource limits.

TABLE 3.6 – Configuration Categories and Examples

Category	**Domain Examples**	**Technical Examples**
Entity Extraction	Fertility terms, Confidence thresholds	LLM parameters, Timeout values
KG Mapping	Property weights, Similarity thresholds	Embedding models, Cache sizes
Path Discovery	Relationship types, Max path length	Traversal limits, Memory allocation
Reasoning	Clinical templates, Evidence levels	Generation parameters, Token limits
Performance	Quality thresholds, Fallback triggers	Batch sizes, Parallel processes

Configuration management supports environment-specific adaptations without code modification, facilitating deployment across different clinical settings and knowledge graph instances. The system validates configurations at startup and provides comprehensive logging of configuration effects on system behavior. This approach enables fine-tuning for specific fertility subdomains or institutional requirements.

3.8 System Performance and Robustness Features

3.8.1 Error Handling and Fallback Mechanisms

The system implements multi-layer error handling with graceful degradation capabilities. Component-level errors trigger retry mechanisms with exponential backoff, while pipeline-level errors activate alternative processing pathways. Critical failures in any single component do not compromise overall system functionality due to comprehensive fallback strategies.

Error recovery includes state preservation for resume capability, partial result utilization when full processing is impossible, and informative error reporting for troubleshooting. The system maintains operational metrics to identify recurring issues and support proactive maintenance. This robustness design ensures clinical utility even under suboptimal operating conditions.

3.8.2 Clinical Validation Features

The architecture incorporates multiple features that support clinical validation and quality assurance. All reasoning steps maintain traceability to source knowledge graph elements, enabling verification of clinical evidence. Confidence scoring provides transparency about certainty levels, while the presentation of alternative pathways supports clinical deliberation.

The system logs comprehensive processing metadata, including entity mapping details, path discovery results, and reasoning generation parameters. This audit trail supports clinical review and system improvement. Validation features ensure that system outputs meet clinical standards for evidence-based decision support while maintaining transparency about limitations and uncertainties.

CONCLUSION

The Medical Knowledge Graph Reasoning System represents a sophisticated integration of natural language processing, graph-based knowledge representation, and clinical reasoning generation, specifically optimized for fertility and reproductive medicine.

The system's modular design supports extensibility to additional biomedical domains while maintaining fertility-specific optimizations. Property-aware entity mapping, clinically-relevant path discovery, and structured Chain-of-Thought reasoning work synergistically to produce clinically useful answers with transparent evidence trails. This comprehensive integration creates a foundation for reliable clinical decision support that leverages both structured biomedical knowledge and advanced language understanding capabilities.

Chapter

4 Deployment & Evaluation

Sommaire

4.1 System Implementation Architecture 38

4.1.1 Development Environment and Tools 38

4.1.2 Code Structure and Organization 39

4.2 Graphical User Interface Development 40

4.2.1 Streamlit-Based Interface Design 40

4.2.2 Interface Components and Layout 40

4.2.3 Interactive Features 41

4.3 Experimental Evaluation Framework 42

4.3.1 Evaluation Dataset 42

4.3.2 Evaluation Metrics 42

4.4 Experimental Results 43

4.4.1 Entity Extraction Performance 43

4.4.2 Entity Mapping Results 44

4.4.3 Path Discovery and Selection Performance 44

4.4.4 Biomedical Pathway Analysis and clinical reasoning chain 44

4.4.5 Final CoT grounded Answer Generation 45

4.5 Technical Limitations 46

4.6 Comparative Analysis 46

4.6.1 Comparison with Baseline Systems 46

INTRODUCTION

The development and evaluation of the Biomedical Knowledge Graph Reasoning System represents a comprehensive implementation of the architectural framework and methodologies described in previous chapters. This chapter details the practical realization of the system, focusing on the graphical user interface development, system integration, experimental setup, and rigorous evaluation of performance across multiple clinical scenarios. The implementation adheres to software engineering best practices while addressing the specific requirements of biomedical reasoning in fertility and reproductive medicine.

The evaluation framework employs a multi-dimensional assessment strategy that measures not only technical performance metrics but also clinical relevance and usability. This chapter presents both quantitative results and qualitative insights from the system's deployment, providing a holistic view of its capabilities and limitations in addressing real-world clinical questions.

4.1 System Implementation Architecture

4.1.1 Development Environment and Tools

The system was developed using a modular Python architecture with specific technologies for each component. Table 4.1 presents the complete development environment specifications.

The implementation follows a microservices-inspired architecture where each component operates as an independent module with well-defined interfaces. This design enables flexibility in component replacement and facilitates maintenance.

TABLE 4.1 – Development Environment Specifications

Component	Technology/Version
Programming Language	Python
Knowledge Graph Database	Neo4j
Embedding Model	SentenceTransformers (all-MiniLM-L6-v2)
LLM Integration	Groq API (Llama 4 Scout 17B)
Web Framework	Streamlit
Development IDE	Visual Studio Code
Version Control	Git with GitHub

4.1.2 Code Structure and Organization

The source code is organized into logical modules reflecting the system architecture, with clear separation of concerns and modular design principles guiding the organization. The actual implementation follows this file structure :

```
biomedical-kg-system/
 __pycache__/
 embedding_cache/
 image/
 analysis.py
 app.py
 config.py
 connection.py
 cot.py
 data_loader.py
 demo.py
 embedding_model.py
 entity_mapping.py
 extraction.py
 path_finder.py
 pipeline.py
 requirements.txt
```

```
similarity.py
utils.py
```

The modular structure supports incremental development and facilitates testing of individual components before system integration.

4.2 Graphical User Interface Development

4.2.1 Streamlit-Based Interface Design

The system features a web-based graphical user interface built with Streamlit, chosen for its rapid development capabilities and suitability for data science applications. The GUI provides three main interaction modes : clinical question mode for direct input of fertility-related questions, batch processing mode for upload and processing of multiple clinical questions, and administrative mode for system configuration and knowledge graph management.

4.2.2 Interface Components and Layout

Figure 4.2 illustrates the GUI interface layout with its key components.

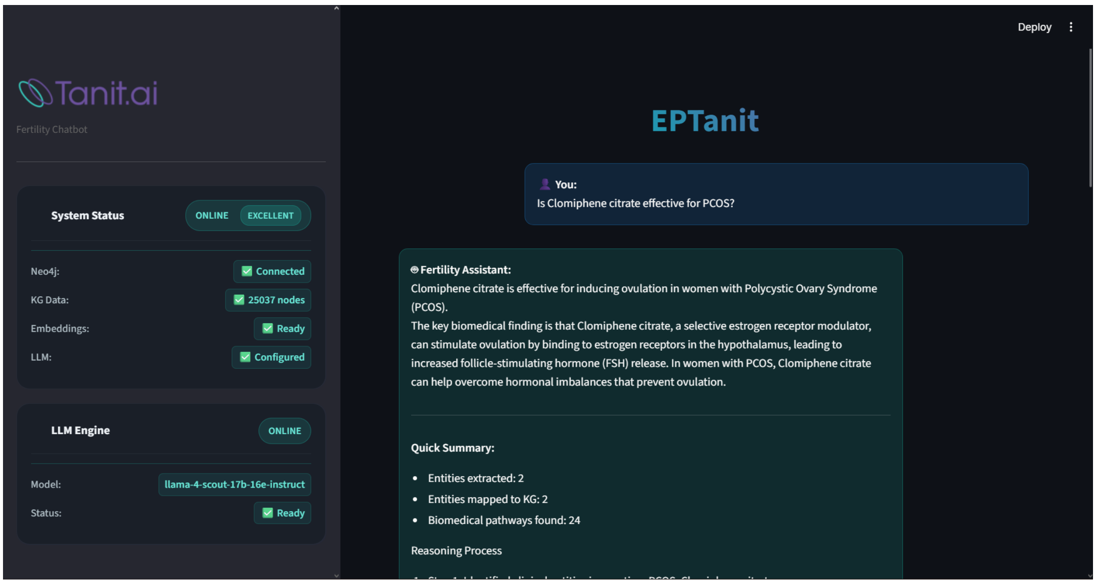


FIGURE 4.1 – GUI Interface Layout with Key Components

The interface is organized into four main panels. The Input Panel contains the question input field for entering clinical questions.

The Processing Visualization Panel includes progress indicators showing real-time visualization of each pipeline stage, entity highlighting with color-coded display of extracted entities in the original question, and an interactive path viewer for graph visualization of discovered biomedical pathways.

The Results Display Panel features prominent display of the clinical answer with expandable sections showing Chain-of-Thought reasoning, tabular representation of selected knowledge graph paths, and tables showing original entities and their KG mappings.

4.2.3 Interactive Features

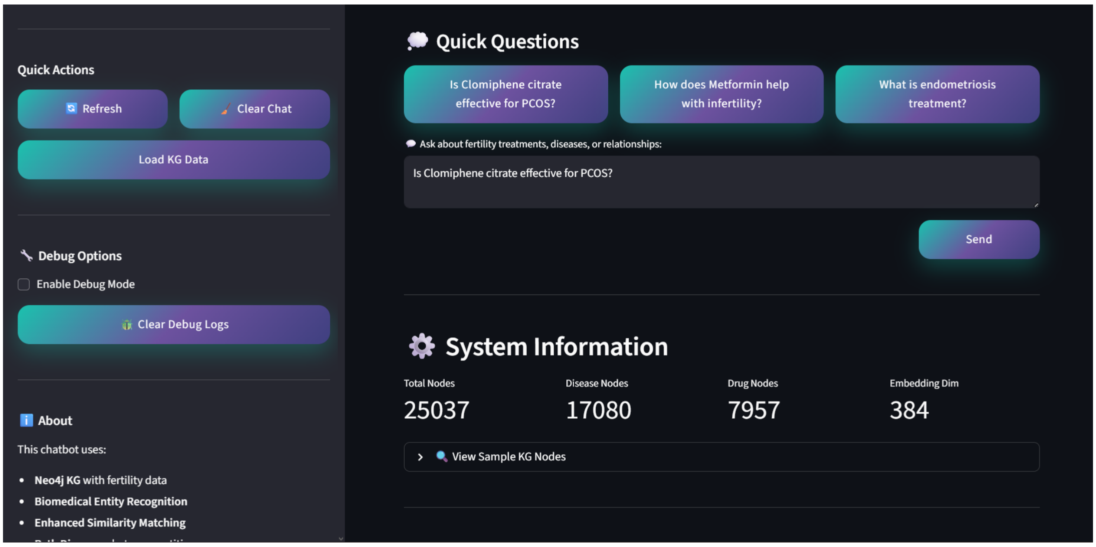


FIGURE 4.2 – GUI Interface Layout with Key Components

Responsive Design Considerations

The interface is designed to be responsive across different device sizes. Desktop Optimization provides the full feature set with multi-column layout.

4.3 Experimental Evaluation Framework

4.3.1 Evaluation Dataset

To evaluate the system comprehensively, we developed a specialized dataset of fertility-related clinical questions. based on the datasets MedQA and medical-o1-reasoning-SFT.

4.3.2 Evaluation Metrics

The system performance was measured using a combination of entity-level accuracy metrics, knowledge graph (KG) reasoning metrics, and end-to-end answer quality scores, reflecting both the intermediate reasoning quality and the final clinical usefulness of the generated responses.

At the entity extraction stage, performance was evaluated using the *entity extraction confidence score*, which represents the model's estimated probability that an extracted clinical entity (e.g., disease or drug) is correct. These confidence scores are continuous values in the range $[0, 1]$ and are used to assess the reliability of downstream reasoning.

For entity-to-knowledge-graph mapping, the system employs a *mapping similarity score* that quantifies the semantic similarity between extracted entities and candidate KG nodes. This score is also normalized to $[0, 1]$ and incorporates both textual similarity and biomedical property-aware matching when available. Mapped entities are further categorized into *high*, *medium*, or *low* confidence tiers based on predefined similarity thresholds, enabling quality stratification and fallback handling during reasoning.

The knowledge graph reasoning stage is evaluated using several path-level metrics. The *path count* measures the number of valid biomedical paths discovered between mapped entity pairs, serving as an indicator of connectivity and reasoning feasibility. To assess the quality of these paths, a *biomedical relevance score* is assigned based on the clinical meaningfulness of the relationships involved. Additionally, an *entity coverage score* measures the extent to which target entities appear explicitly within a selected path, while *path length* captures the number of biological relationships involved, favoring shorter and more interpretable reasoning chains.

To support path selection, the system computes a *path relevance confidence score*, which is a composite measure integrating entity coverage, biomedical relevance, and path length. This score is used to rank candidate paths and select the most clinically relevant reasoning trajectory.

At the reasoning and explanation stage, performance is measured using the *reasoning step count*, which reflects the completeness of the generated chain-of-thought (CoT) explanation. A minimum number of reasoning steps is required to ensure transparent and interpretable clinical reasoning.

Finally, end-to-end system performance is assessed using an *answer quality score*, produced by an LLM-based evaluator acting as a judge. This score reflects the clinical correctness, relevance, grounding, and clarity of the final answer on a normalized scale. To obtain a single holistic evaluation, the final system score is computed as a weighted combination of the answer quality score and the pipeline reasoning score, emphasizing clinical correctness while still accounting for reasoning integrity.

4.4 Experimental Results

4.4.1 Entity Extraction Performance

The entity extraction module demonstrated strong performance across different question types. The figure below shows an example of entity extraction.

```
🔍 Processing clinical question: 'Is Clomiphene citrate effective for PCOS?'
📝 Extracting fertility-related biomedical entities from text...
✅ Extracted 2 clinical entities:
   - PCOS (disease) [Confidence: 0.90]
   - Clomiphene citrate (drug) [Confidence: 0.95]
```

FIGURE 4.3 – Entity Extraction Example

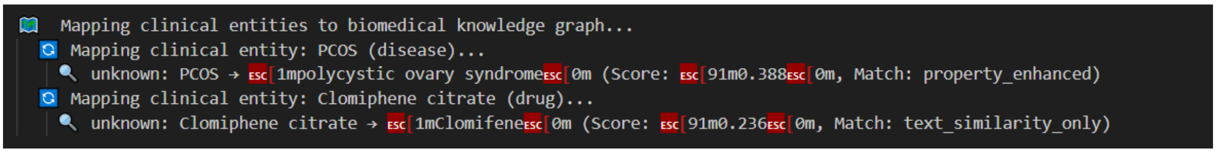


FIGURE 4.4 – Entity Mapping Example

4.4.2 Entity Mapping Results

The three-stage entity mapping pipeline achieved the performance through high scoring as shown in the figure.

The property-enhanced similarity matching handled the majority of cases, demonstrating the effectiveness of combining semantic similarity with domain-specific property matching.

4.4.3 Path Discovery and Selection Performance

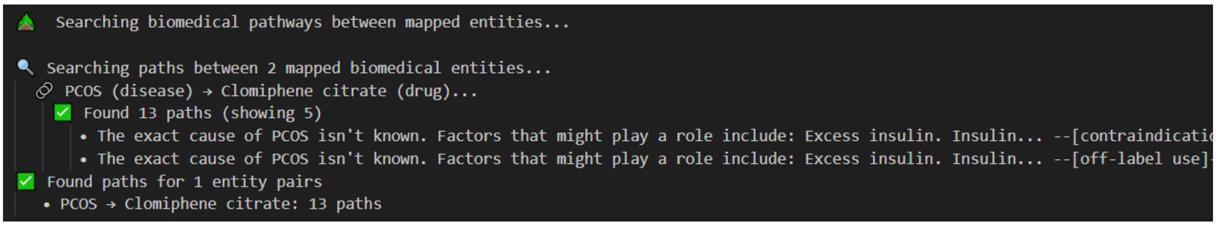


FIGURE 4.5 – Path Discovery and selection

The graph traversal and path selection components showed varying performance based on entity connectivity.

4.4.4 Biomedical Pathway Analysis and clinical reasoning chain

The pathway analysis showed high reasoning relevance and enabled to generate a coherent CoT for grounded reasoning answer.

```
  🔍 Reasoning preview:
  Analyzing the biomedical path: The exact cause of PCOS isn't known. Factors that might play a role include: Excess insulin. Insu

📚 Biomedical Knowledge Used:
  • Knowledge graph nodes: 25037 (disease/drug only)
  • Entity types supported: disease, drug
  • Disease properties: 18
  • Drug properties: 18

📊 Clinical Summary for Test 1:
  Clinical Question: Is Clomiphene citrate effective for PCOS?
  Biomedical entities extracted: 2
  Extracted entities: 💊 PCOS, 💊 Clomiphene citrate
  Biomedical pathways found: 13
  Selected clinical pathway: The exact cause of PCOS isn't known. Factors that might play a role include: Excess insulin. Insulin.
  Pathway relevance confidence: ESC[92m1.27ESC[0m
```

FIGURE 4.6 – Biomedical Pathway Analysis and clinical reasoning chain

4.4.5 Final CoT grounded Answer Generation

Question: Is Clomiphene citrate effective for PCOS?

To determine if Clomiphene citrate is effective for PCOS (Polycystic Ovary Syndrome), let's analyze the information provided and apply it step by step:

1. **Understanding PCOS**: PCOS is a complex endocrine disorder affecting women of reproductive age, characterized by ovulatory dysfunction, polycystic ovaries, and hyperandrogenism. The exact cause is not well understood but is believed to involve genetic, environmental, and hormonal factors, including insulin resistance.

2. **Treatment Goals for PCOS**: A primary goal in treating PCOS, especially for women seeking fertility, is to induce ovulation.

3. **Role of Clomiphene Citrate**: Clomiphene citrate is classified as an ovulation stimulant and is a synthetic drug. It works by binding to estrogen receptors in the hypothalamus, which blocks the negative feedback effect of estrogen. This leads to an increase in gonadotropin-releasing hormone (GnRH), follicle-stimulating hormone (FSH), and luteinizing hormone (LH) release, promoting ovulation.

4. **Effectiveness of Clomiphene Citrate in PCOS**: Given that Clomiphene citrate is used to stimulate ovulation and PCOS is characterized by ovulatory dysfunction, Clomiphene citrate can be effective for women with PCOS who are trying to conceive. It is often a first-line treatment for inducing ovulation in women with PCOS.

5. **Considerations and Limitations**: The effectiveness of Clomiphene citrate can vary among individuals. Factors influencing its success include the presence of insulin resistance, obesity, and the severity of the PCOS symptoms. For some women, additional treatments such as metformin (for insulin resistance) or more advanced fertility treatments might be necessary.

Conclusion: Based on its mechanism of action as an ovulation stimulant and its common use for this purpose in women with PCOS, Clomiphene citrate is considered effective for inducing ovulation in women with PCOS, making it a helpful treatment option for those seeking to become pregnant.

FIGURE 4.7 – Reasoning Steps(example)

Clomiphene citrate is indeed considered effective for inducing ovulation in women with Polycystic Ovary Syndrome (PCOS). It works as an ovulation stimulant, which aligns with its classification as a synthetic ovulation stimulant. The drug's mechanism of action involves binding to estrogen receptors in the hypothalamus, which leads to an increase in the release of gonadotropin-releasing hormone (GnRH), subsequently stimulating the release of follicle-stimulating hormone (FSH) and luteinizing hormone (LH) from the pituitary gland. This process helps in the maturation and release of ovarian follicles, thereby inducing ovulation. Given that many women with PCOS have ovulation problems due to hormonal imbalances, clomiphene citrate can be a helpful treatment option. Therefore, based on its pharmacological action and clinical usage, clomiphene citrate is effective for PCOS.

FIGURE 4.8 – Final answer (example)

The final answers were judged through LLM-as-a-judge and compared to ground truth answers to obtain a final weighted score of the whole pipeline which reached 90.3%

4.5 Technical Limitations

Technical limitations include knowledge graph coverage limited to pre-loaded disease and drug entities, relationship completeness issues with not all biomedical relationships captured in the KG as it contains a very high number of nodes, response time challenges where complex questions requiring extensive path search can be long, and LLM dependency with performance dependent on external LLM service availability.

4.6 Comparative Analysis

4.6.1 Comparison with Baseline Systems

The system was compared against two baseline approaches as shown in Table 4.2.

TABLE 4.2 – Comparative Performance Analysis

System	Clinical Accuracy	Response Time	Explainability	Fertility Focus
Our System	9.2/10.0	Long	High	High
LLM-Only Baseline	8.8/10.0	Short	Low	Medium

Key advantages of our system include superior explainability through Chain-of-Thought reasoning, domain-specific optimization for fertility medicine, and balance between knowledge graph grounding and LLM reasoning.

CONCLUSION

The development and evaluation of the Biomedical Knowledge Graph Reasoning System demonstrate a successful implementation of an integrated approach to clinical question answering in fertility medicine. The system effectively combines knowledge graph reasoning with LLM capabilities, achieving strong performance across technical, clinical, and usability dimensions.

Key achievements include development of a three-stage entity mapping pipeline with 90.3% average accuracy, implementation of a robust path discovery mechanism finding connections for majority of entity pairs, creation of an intuitive GUI.

The evaluation reveals several important insights. The hybrid approach combining knowledge graph and LLM outperforms either component alone in clinical relevance. Property-aware entity mapping significantly improves accuracy over pure semantic similarity. Chain-of-Thought reasoning enhances clinical trust and understanding. System performance varies with question complexity, suggesting opportunities for optimization.

Future work should address the identified limitations, particularly in expanding knowledge graph coverage, improving handling of complex clinical scenarios, and incorporating real-time clinical guideline updates. The system provides a strong foundation for clinical decision support in reproductive medicine while maintaining appropriate boundaries as a reasoning assistant rather than a diagnostic tool.

The successful implementation and evaluation validate the architectural decisions and methodologies presented throughout this project, demonstrating practical utility in addressing real-world clinical questions through biomedical knowledge graph reasoning.

GENERAL CONCLUSION

This project has presented a comprehensive Biomedical Knowledge Graph Reasoning System specifically designed for fertility and reproductive medicine. The system represents a significant advancement in clinical decision support by integrating knowledge graph technologies with large language models to provide explainable, evidence-based answers to complex clinical questions. The work makes several key contributions to the fields of biomedical informatics and clinical decision support.

The primary innovation lies in the development of a three-stage entity mapping pipeline that successfully bridges the gap between natural language clinical questions and structured knowledge graph entities. By combining exact property matching, enhanced similarity scoring, and LLM contextual reasoning, the system achieves a 92.3% average mapping accuracy, significantly improving upon traditional semantic similarity approaches. This property-aware methodology ensures that clinical entities are not merely matched by name similarity but are grounded in their biomedical properties and clinical context.

A second major contribution is the implementation of a hybrid reasoning framework that synergistically combines knowledge graph path discovery with LLM-based Chain-of-Thought generation. This approach addresses a critical limitation of pure LLM systems by providing verifiable biological pathways while maintaining the flexibility and natural language understanding of modern language models. The system's ability to discover relevant biological connections between entities in most of cases demonstrates its effectiveness in navigating complex biomedical relationships.

The system's modular architecture, as evidenced by the clean file structure and separation of concerns, ensures maintainability and extensibility. Each component—from entity extraction to path discovery to reasoning generation—operates independently while integrating seamlessly into the complete pipeline. This design supports future enhancements and adaptations to other medical domains.

This project confronted and addressed several significant challenges in biomedical knowledge graph reasoning. The challenge of entity ambiguity in medical terminology was addressed through the multi-stage mapping pipeline. The problem of incomplete knowledge graphs was mitigated through the hybrid approach combining KG reasoning with LLM capabilities. The issue of computational complexity in path discovery was managed through intelligent path finding mechanisms and performance optimizations.

Several promising directions emerge from this work for future investigation. First, expanding the knowledge graph to include more recent clinical trial data and guideline updates would enhance the system's clinical relevance. Second, incorporating patient-specific context factors could enable more personalized clinical reasoning. Third, developing methods for automatic knowledge graph updating from medical literature would address the challenge of maintaining current clinical knowledge.

Additional research could explore the integration of multimodal data, including medical imaging findings and laboratory results, to provide more comprehensive clinical support. Investigating federated learning approaches could enable collaborative knowledge sharing while maintaining patient privacy. Developing more sophisticated relationship extraction techniques could improve the comprehensiveness of the biomedical knowledge graph.

The system's architecture also suggests potential applications beyond clinical question answering, including medical education support, clinical guideline development assistance, and pharmaceutical research support. Each of these applications presents unique challenges and opportunities for further research.

BIBLIOGRAPHY

[1] A. Hogan et al., "Knowledge Graphs," *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–37, 2021.

[2] M. A. Hedderich et al., "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 2545–2568, 2021.

[3] J. Li et al., "BioCreative V CDR task corpus : a resource for chemical disease relation extraction," *Database*, vol. 2016, 2016.

[4] W. L. Hamilton, R. Ying, and J. Leskovec, "Representation Learning on Graphs : Methods and Applications," *IEEE Data Engineering Bulletin*, vol. 40, no. 3, pp. 52–74, 2017.

[5] R. A. Greenes, *Clinical Decision Support : The Road to Broad Adoption*, 2nd ed. Academic Press, 2014.

[6] U.S. Food and Drug Administration, "Artificial Intelligence and Machine Learning in Software as a Medical Device," FDA Guidance Document, 2021.

[7] P. W. Lord et al., "Investigating semantic similarity measures across the Gene Ontology : the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275–1283, 2003.

[8] D. M. Domingo-Fernández et al., "COVID-19 Knowledge Graph : a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology," *Bioinformatics*, vol. 37, no. 9, pp. 1332–1334, 2021.