Leveraging Large Language Models and Knowledge Graphs for Disease Theory Exploration and Causal Analysis

Isak Midtvedt
Oslo Metropolitan University
Norway
s350289@oslomet.no

Shanshan Jiang
SINTEF AS
Norway
shanshan.jiang@sintef.no

Dumitru Roman
SINTEF AS / Oslo Metropolitan University
Norway
dumitru.roman@sintef.no

Abstract

Clinical researchers need timely, evidence-supported overviews that clarify the underlying mechanisms and pathways of disease development. Manual curation of causal mechanisms is highly time-consuming and increasingly infeasible due to the exponential expansion of the biomedical literature. This paper introduces *KnowDisease* – an approach to automatically extracting disease theory mechanisms and causal relationships from biomedical literature using Large Language Models, exploiting Chain-of-Thought prompting with evidence traceability and constructing Knowledge Graph-based disease theories for causal analysis. An application is implemented based on this approach, enabling researchers to visualize complex disease pathways and causal relationships, navigate interconnected biological mechanisms and identify evidence for various disease theories based on constructed knowledge graphs. The work demonstrates how LLM-assisted tools can advance understanding of disease mechanisms and potentially accelerate biomedical research.

CCS Concepts

• **Computing methodologies** → **Knowledge representation and reasoning**; **Natural language generation**.

Keywords

Large Language Model, Knowledge Graph, Disease Theory

ACM Reference Format:
Isak Midtvedt, Shanshan Jiang, and Dumitru Roman. 2025. Leveraging Large Language Models and Knowledge Graphs for Disease Theory Exploration and Causal Analysis. In *Proceedings of The 2025 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'25).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Modern biomedical research faces the critical challenge of managing and interpreting an exponentially expanding body of research data. The extensive and complex nature of biomedical publications presents significant barriers to manually identifying causal relationships and underlying disease mechanisms.

Understanding disease progression relies on explanatory models, or *disease theories*, which are essential to improve diagnosis, treatment, and prevention. Representing these theories computationally is challenging due to their complexity and evolving nature. Effective models must capture causal mechanisms, supporting evidence, and biological context and at the same time remain flexible to incorporate new knowledge. While ontologies (e.g., SNOMED-CT and Gene Ontology) and description logics [11] provide standardized vocabularies and logical frameworks for medical knowledge bases, comprehensive modeling of explanatory and causal relationships remains an open research area. Capturing these causal links is vital for enabling targeted interventions, explainable decisions, and advancing precision medicine.

Knowledge Graphs (KGs) offer promising, effective representations of complex biomedical information, supporting reasoning and enhancing causal inference by uncovering implicit relationships in structured, semantic data. At the same time, Large Language Models (LLMs) have rapidly advanced, enabling domain-specific applications in scientific text analysis.

The objective of this paper is to leverage LLMs and KGs to assist researchers in exploring disease mechanisms, identifying causal pathways, and tracing supporting evidence with greater efficiency and transparency. In particular, we aim to explore *how to accumulate knowledge from multiple scientific papers while maintaining full traceability to supporting evidence.* To address these goals, we proposed an approach—*KnowDisease*—utilizing LLMs and KGs for disease theory extraction and analysis:

- To address KG challenges regarding maintaining evidence traceability and ensuring accessibility for non-technical domain experts, *KnowDisease* uses LLM for transparent term extraction and evidence linking, and presents results in an interactive Neo4j graph interface optimized for intuitive exploration.
- To address LLM limitations, such as hallucinations, limited handling of rare or specialized concepts, and insufficient capacity for robust causal reasoning, *KnowDisease* combines Retrieval Augmented Generation (RAG) with Chain-of-Thought (CoT) prompting, ensuring relevant context is provided and model reasoning remains transparent.

The main contribution of this work is the *KnowDisease* approach for disease theories establishment from scientific literature based on LLMs and KGs, demonstrated and evaluated for disease theory

KDD'25, Toronto, Canada

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

exploration, featuring automatic knowledge extraction and representation, interactive KG visualization, evidence-based theory exploration and causal pathway identification.

The remainder of this paper is organized as follows. Section 2 presents background and related work. Section 3 introduces the *KnowDisease* approach, and Section 4 provides details about *KnowDisease* implementation, while Section 5 outlines the evaluation framework and results. Finally, Section 6 concludes the paper.

2 Background and Related Work

2.1 Knowledge Graphs in Biomedical Research

Biomedical research produces vast, fragmented data across literature, databases, and clinical records. Early examples such as Hetionet [4] demonstrated the potential of KGs to generate novel hypotheses, such as drug repurposing, by revealing non-obvious connections. More recent efforts such as SPOKE [10] and KG-COVID-19 [12], scaled this approach by integrating diverse data sources to support personalized medicine and rapid crisis response. Despite their advances, key challenges remain, for example, regarding maintaining evidence traceability and ensuring accessibility for non-technical domain experts.

2.2 LLMs for Scientific Literature Analysis

The rapid development of LLMs has significantly enhanced capabilities in natural language processing (NLP), particularly in scientific text analysis. Early successes with general models such as BERT led to specialized variants, e.g., BioBERT [6] and PubMedBERT [3], which demonstrated superior performance on biomedical NLP tasks through training on domain-specific corpora. The emergence of generative LLMs, such as BioGPT [9], further extended capabilities to fluent generation, question answering, and zero-shot relation extraction, often outperforming larger general models on tasks like PubMedQA [5]. Retrieval-Augmented Generation (RAG) enhances these models by grounding outputs in retrieved source material, improving factual accuracy and traceability—critical features for scientific domains [7]. However, challenges need to be addressed: LLMs are prone to hallucinations, may struggle with rare or highly specialized concepts, and often lack robust causal reasoning.

2.3 RAG Pipelines for Biomedical Data

RAG has been applied in health domain to help LLM models generate more accurate answers using information from external sources, e.g., pipelines described in [14] and LLMDap[1]. LLMDap is a LLM-based pipeline for data enrichment. LLMDap improves metadata quality and discoverability by automatizing metadata assessment and enrichment, providing a sequential processing pipeline architecture that gives an essential foundation for structured data generation from biomedical literature. LLMDap focuses mainly on dataset metadata extraction, linking to the ArrayExpress database.

KnowDisease builds upon LLMDap and extends it by adapting the LLM pipeline and integrating with KGs for extraction and analysis of disease theories. Targeting improved reliability, transparency, and biomedical relevance, *KnowDisease* introduces several key enhancements over the original LLMDap pipeline:

- *Information Extraction:* Replaced regex constraints with CoT prompting and JSON schema validation using Outlines, improving reliability and interpretability.
- *Data Preprocessing:* Improved XML handling and chunking quality by preserving document structure and targeting papers with rich disease theory content.
- *Knowledge Representation:* Shifted from metadata profiling to constructing disease-specific KGs with causal links and full evidence provenance in Neo4j.
- *Semantic Retrieval:* Enhanced retrieval precision using domain-specific embeddings and relevance thresholds tailored to biomedical queries.
- *User Interface:* Introduced a Streamlit-based interface for interactive exploration of KGs, evidence, and causal diagrams.

3 KnowDisease Approach

KnowDisease is based on an end-to-end pipeline that automatically extracts disease theory mechanisms and causal relationships from biomedical literature using LLMs and constructs KGs-based disease theories, as illustrated in Figure 1.

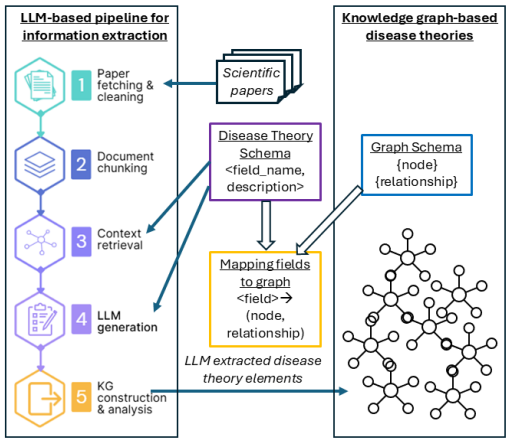

Figure 1: KnowDisease Approach: Graph-based disease theories construction with LLM-based pipeline.

3.1 KG-based Disease Theory Representation

In *KnowDisease*, disease theories are represented as KGs where the nodes represent the key aspects in the disease theories and the edges represent their relationships, including causal relationship. This approach preserves the inherent structure of the theories, enabling the systematic exploitation of KGs reasoning capabilities.

An example KG structure for our experiments is designed with six primary nodes (*disease_name, etiology_factor, diagnostic_method, biomarker, treatment_intervention and prognostic_indicator*), representing key disease theory elements and their relationships. A *Graph Schema* is defined to store disease theory KGs in graph database following this KG structure.

A *Disease Theory Schema* is defined based on the KG structure to facilitate extracting structured, meaningful disease theory elements

[1] https://github.com/SINTEF-SE/LLMDap

from biomedical literature using LLMs. The schema consists of six primary fields corresponding to the six nodes defined in the KG structure. Each field has a detailed description used in both *context retrieval* and *LLM generation* steps of the LLM-based pipeline (see details in Section 3.2).

A *mapping of Disease Theory schema fields to graph schema* is designed to reflect the structure and preserve the meaning of relationships between disease concepts, facilitating storing the extracted disease theory elements in graph databases for disease theory construction and causal analysis. Each relationship type is clearly defined as to what it represents, *HAS_ETIOLOGY_FACTOR* for instance indicates a causal relationship, while *HAS_BIOMARKER* represents a measurable indicator relationship. This semantic structure enables sophisticated queries about disease mechanisms and their supporting evidence.

3.2 Pipeline for Disease Theory Population

An end-to-end pipeline is exploited to extract relevant information from biomedical literature and construct KGs for disease theories. This pipeline consists of five components representing key steps in the pipeline as depicted in Figure 1:

- *Paper fetching and cleaning*: Retrieve publications with metadata from scientific literature databases and transforme them into clean, structured text. To retain paper section hierarchy, title passages are converted into markdown headers (e.g., ## for level 2) to define natural content boundaries.
- *Document chunking*: Split the structured text into manageable, semantically coherent chunks (i.e., text segments) by leveraging the markdown headers inserted in the first step.
- *Context retrieval*: Rank chunks based on semantic similarity with Disease Theory Schema fields using domain-specific embeddings, and output high-relevance segments per field, providing the LLM with context-rich input tailored for disease theory extraction.
- *LLM generation*: Generate answers and extract evidence for fields representing disease theory elements with LLM using retrieved context. A CoT approach is exploited for verification and interpretation of outputs and extraction of evidence quotes, ensuring trustworthiness.
- *KG-based disease theory construction and causal analysis*: Build disease theory graphs from the LLM outputs and produce causal diagrams with accumulated knowledge from multiple papers.

The first four components build upon the original LLMDap with notable adaptions to optimize the performance, while the last component is a new one dedicated for disease theory analysis, leveraging KG's inference power.

3.2.1 Paper fetching and cleaning. The component retrieves research articles about a specified disease in XML format, using specialized Medical Subject Headings (MeSH)[2] terms and well-constructed queries to target papers with causal and mechanistic disease information.

[2]Medical Subject Headings (MeSH) from National Library of Medicine: https://meshb.nlm.nih.gov.

The query construction uses the disease name as a major topic, incorporates MeSH terms related to disease mechanisms, targets keywords like "pathogenesis", "mechanism" and "theory", and filters for open-access papers. These strategies ensure retrieval of highly relevant literature for populating the KG. The query outputs XML files with paper text and structural metadata, which are converted into clean, structured text suitable for downstream processing.

The processing identifies and filters <passage> elements based on type and section_type metadata, excluding non-informative sections (e.g., references, acknowledgments, licenses). Unlike the original LLMDap, which minimally structured the output, our approach retains section hierarchy by converting title passages into markdown headers, aiding coherent chunking.

Remaining content passages are included as-is, and all chunks are joined with double newlines to preserve paragraph separation. The output is a markdown-enhanced document that preserves logical structure and provides cues for semantically aware chunking, enabling more effective disease theory extraction.

3.2.2 Document chunking. The component splits the structured, markdown-enhanced text into manageable, semantically coherent units (i.e., chunks) for downstream analysis and retrieval. Unlike the original pipeline, which relied on fixed section titles (e.g., "METHODS," "RESULTS"), our approach uses markdown headers inserted during *Paper fetching and cleaning* to define natural content boundaries. This structure-aware method improves adaptability across diverse biomedical documents, even those lacking standard section labels. Chunks are initially split at markdown headers, with further subdivision guided by a hierarchy of breakpoints—prioritizing paragraphs, then line breaks, sentence ends, and only splitting within sentences as a last resort.

To enhance chunk quality, an overshoot factor is applied for flexible sizing and filter out fragments below a minimum length. The output is a set of TextNode objects [8] containing well-structured text segments and metadata, forming the basis for targeted retrieval in disease theory modeling.

3.2.3 Context retrieval. This step identifies the most relevant chunks for each field in the disease theory schema. It filters and selects content to provide the LLM with targeted input for accurate extraction.

Building on the original LLMDap architecture, we optimized retrieval by incorporating domain-specific embeddings. Document chunks and schema field descriptions are embedded into a shared vector space using the fine-tuned model *pritamdeka/S-PubMedBert-MS-MARCO*[2], which improves semantic matching of biomedical concepts.

Semantic similarity is computed via cosine similarity, enabling robust identification of relevant content despite lexical variation. A configurable relevance threshold excludes low-scoring chunks, reducing noise.

The output is a concatenated string of high-relevance segments per field, feeding the LLM with tailored, context-rich input to enable disease theory extraction.

3.2.4 LLM generation. This component represents a major advancement over the original LLMDap pipeline. Replacing regex-constrained outputs, a CoT approach is adopted using a CoT model schema consisting of:

- *reasoning:* step-by-step reasoning process for deriving the answer.
- *evidence:* mapping from each extracted term to the verbatim quote(s) from the supporting context.
- *final_answer:* up to three distinct terms that directly and specifically answer the field's question and at the same time serve as key indicators of supporting evidence.

This design requires the LLM to return structured JSON containing the extracted term, its reasoning, and supporting evidence, enhancing not only performance, but also transparency and traceability. By prompting step-by-step reasoning, the model reveals its decision process, enabling verification and interpretation of outputs.

The Outlines library [13] is used to enforce schema-conformant JSON generation, eliminating parsing errors common with regex and supporting more complex, interpretable responses. Verbatim evidence quotes strengthen the reliability of extractions and support expert validation, ensuring high trustworthiness in a domain where precision is critical.

3.2.5 KG construction and causal analysis. The structured information extracted with the previous steps (Subsections 3.2.1-3.2.4) is processed to construct KGs using graph databases. This includes entities normalization, creation or update of nodes representing disease elements and establishment of relationships between them. In particular, connections between source publication and extracted supporting textual evidence are created to facilitate transparency and traceability.

Evidence Extraction and Merging. This component addresses the challenge stated in the introduction: *how to accumulate knowledge from multiple research papers while maintaining full traceability to supporting evidence*. The most critical feature is the *evidence accumulation strategy* through building relationships that grow richer with each processed paper. This approach ensures that when a relationship already exists between two nodes, new evidence and source information is appended rather than overwritten and allows for merge related information without losing the evidence trail from individual sources when processing multiple papers about the same disease.

Furthermore, the evidence storage mechanism preserves the full CoT reasoning from the extraction process. A flexible format that accommodates both the enhanced CoT evidence structure and simpler string-based evidence for backward compatibility is designed.

Normalization and De-duplication. To handle variations in terminology across different papers, data normalization is applied to ensure that semantically equivalent terms are merged appropriately while filtering out non-informative entries.

The resulting KG provides the structured foundation for interactive exploration through a Streamlit application (Section 4.2). Disease theories are stored as interconnected nodes with rich relationship properties, which enables sophisticated queries about disease mechanisms, evidence comparison across papers, and identification of knowledge gaps in the literature. This graph structure transforms isolated extracted data into a coherent, queryable knowledge representation that supports the deeper understanding of disease theories.

4 KnowDisease Implementation

To demonstrate the feasibility of the *KnowDisease* approach, an implementation of the approach is provided (available on Github[3]). The implementation consists of three main elements: (1) a multi-stage backend processing pipeline; (2) the Neo4j graph database for persistent storage; and (3) a Streamlit Web application that provides users intuitive access to the application features. The Streamlit application provides a user-friendly interface for researchers to interact with a knowledge base built from biomedical literature. Users can submit papers for analysis and inclusion in the knowledge base, explore extracted theories, and view causal diagrams illustrating disease relationships. All information is directly linked to supporting evidence and source publications.

4.1 Disease Theory Population through the UI

Users can use *KnowDisease* UI to specify the publications for population of the disease theory graph database, either by uploading biomedical scientific papers, or by specifying their PMIDs or disease name for remote database query. This input then activates the backend pipeline, which processes documents through the five sequential steps described in Section 3.2 and parses LLM output into Neo4j database as a disease theory.

4.2 Disease Theory Exploration and Analysis

The UI provides pages for interactive exploration of the KG, visualization of disease specific causal diagrams, and detailed examination of evidence supporting each extracted element, facilitating deeper understanding of disease mechanisms and progression.

The *Disease theory KG page* (Figure 2) visualizes the structured disease theory knowledge base as an interactive network reflecting the interconnected nature of biomedical knowledge development, enabling researchers to explore complex relationships among diseases, etiologies, treatments, biomarkers, diagnostics, and other biomedical concepts. This graph-based interface supports cross-paper integration and evidence accumulation, offering a comprehensive view of disease theories. This page also presents key metrics (such as total nodes and relationships in the Neo4j database) to contextualize the scope of the knowledge base. The graph is rendered using PyVis[4] and NetworkX[5], preserving the structure of the Neo4j database while enabling intuitive interaction. Disease nodes are visually emphasized as central hubs, surrounded by color-coded nodes representing related biomedical concepts, allowing users to quickly recognize conceptual structures and identify patterns, such as the density of treatment information linked to specific diseases. The search feature on the page enables users to filter the visualization by specific terms or concepts, supporting targeted exploration. An organic discovery-oriented layout clusters related concepts naturally, highlighting overlaps and complementarities across studies.

[3] https://github.com/SINTEF-SE/KnowDisease
[4] https://pyvis.readthedocs.io/en/latest/
[5] https://networkx.org/

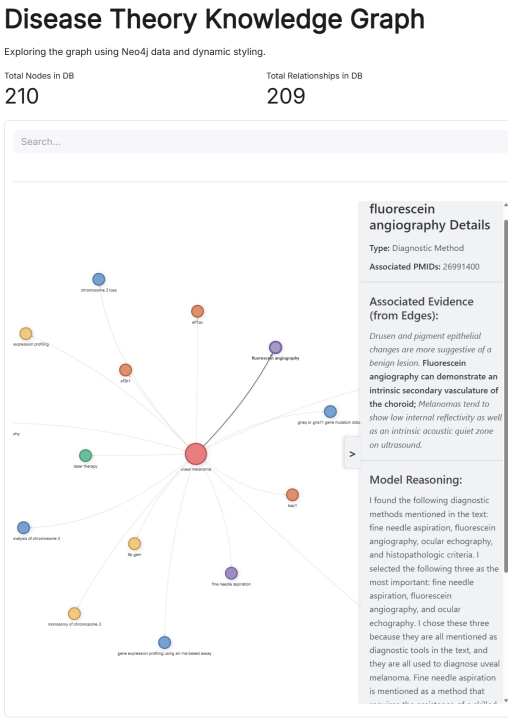

Figure 2: Disease theory KG page.

A notable aspect of the visualization page is the integrated *evidence panel* in the lower right part, which maintains the full traceability chain from the extraction process. When a node is selected, the panel displays its type, linked PMIDs, and the original textual evidence supporting its associated relationships. This ensures transparency by allowing users to examine the exact sentences and contextual information from source papers that underpin each connection.

The *Causal Diagram* page (Figure 3) provides paper-specific visualizations of disease theories where extracted concepts are arranged by category and color. It highlights how individual papers relate concepts to a disease, revealing their research focus. The causal diagram is fully integrated with the KG and evidence traceability and offers paper-specific insights, helping researchers assess study scope, compare perspectives, and verify source text for each concept.

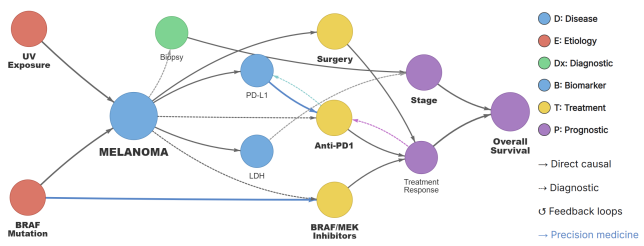

Figure 3: Causal diagram for *Melanoma*.

5 Evaluation

5.1 Evaluation Framework

To evaluate *KnowDisease*, we created a baseline and a systematic evaluation method.

5.1.1 Baseline. In the absence of a benchmark, we applied the "LLM-as-a-Judge" approach [15], leveraging OpenAIs GPT-o3, Gemini 2.5 Pro, and Claude Sonnet 3.7 to simulate expert biomedical researchers. Each model extracted key concepts from 10 shared baseline papers into a predefined Pydantic schema. The outputs were cross-compared and consolidated into a single "silver truth" per paper, providing an approximate but effective reference for system performance evaluation.

5.1.2 Multi-metric approach and custom scoring metric. Multiple metrics were selected for evaluation of system performance, including precision, recall, processing time, F1 score and field specific analysis. A custom F1 scoring function was developed to address abbreviation handling, term specificity, and semantic overlap using a layered strategy with exponential penalties: exact matches score fully; otherwise, semantic similarity (*SimScore*) guides scoring. Terms with *SimScore* ≥ 0.9 (or ≥ 0.88 for abbreviations ≤ 3 characters) are accepted. Scores between 0.85–0.9 are penalized based on substring overlap; 0.75–0.85 receive reduced weight (max 0.4); below 0.75 score zero. Parameters were empirically tuned for optimal performance.

5.1.3 Optimization framework. Systematic optimization was applied using OptWuna [1], a hyperparameter optimization framework that automatically searches for best system configurations defined by a given metric. To explore a large parameter space, we evaluated chunk sizes from 512 to 1408 tokens (step 128), overlaps from 0 to 512, and top_k values from 2 to 7. Beyond retrieval settings, we compared three schema designs (QA-based, keyword-based, and natural language) and CoT vs. non-CoT extraction. Approximately 250 full-pipeline trials were run, providing a comprehensive assessment of configuration impacts on system performance.

5.2 Evaluation Results

Figure 4 shows that CoT outperforms non-CoT, while CoT works best with focused retrieval (k=3), and non-CoT improves with more context until k=6.

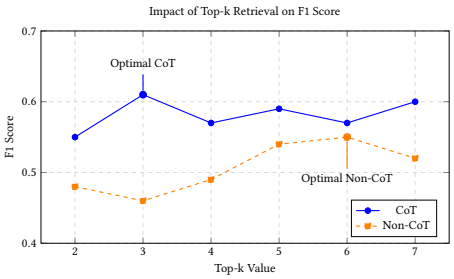

Figure 4: Impact of Top-k Retrieval on F1 score.

The field-specific performance analysis (Figure 5) highlights variability in extraction difficulty across disease theory components,

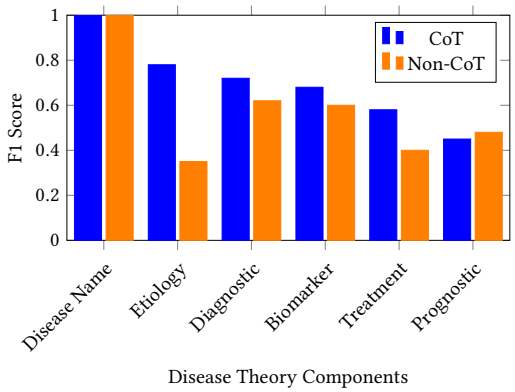

Figure 5: Field-specific F1 performance showing extraction accuracy across different disease theory components.

reflecting the inherent complexity of biomedical knowledge structures. For example, CoT reasoning yields substantial performance improvements in extracting etiological factors, achieving an F1 score of 0.78 compared to 0.35 for non-CoT approaches. This highlights the importance of step-by-step reasoning in distinguishing causal relationships from loosely associated terms.

Our optimizations significantly enhanced the effectiveness of semantic retrieval for disease theory extraction. Our evaluation results showed that replacing the general-purpose model *all-MiniLM-L6-v2* with the domain-specific *pritamdeka/S-PubMedBert-MS-MARCO* significantly improved retrieval precision over original pipeline. In addition, approximately 250 Optuna trials revealed that disease theory extraction required task-specific retrieval settings, with a CoT-enabled configuration using 1024-token chunks and 192-token overlap yielding superior performance compared to the original pipeline defaults.

6 Conclusion

This paper introduced the *KnowDisease* approach to leverage the power of LLMs and KGs to extract and analyze disease theories from biomedical literature. By enhancing an existing LLM-based pipeline with CoT prompting, the approach enables automatic extraction of disease mechanisms and supporting evidence. By representing extracted structured information into KGs, the work supports deeper analysis of disease mechanisms and therapeutic targets.

The work contributes to AI-assisted scientific discovery by demonstrating how to integrate LLM capabilities with evidence linking and transparent reasoning, enabling verification and critical evaluation over blind trust on LLM outputs. The work also contributes to biomedical KG applications by aggregating evidence across multiple sources with full provenance, yielding more robust and trustworthy representations than simple, isolated fact extraction.

The approach has been evaluated using the *KnowDisease* implementation. One future work is to incorporate domain expert validation and feedback loop to enhance accuracy and enable ground truth dataset creation. Another improvement is to design more sophisticated causal diagrams to capture complex biological hierarchies. Deeper integration with biomedical ontologies could improve terminology normalization and extraction specificity. Finally, task-specific model fine-tuning, informed by expert feedback, presents a path to further performance gains and iterative system refinement.

Acknowledgments

The work is funded through the projects UPCAST (HE 101093216), enRichMyData (HE 101070284), Graph-Massivizer (HE 101093202), and DataPACT (HE 101189771).

References

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. arXiv:1907.10902 [cs.LG]

[2] Pritam Deka, Anna Jurek-Loughrey, and Deepak Padmanabhan. 2022. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence* 3, 4 (Nov. 2022), 474–505.

[3] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* 3, 1, Article 2 (Oct. 2021), 23 pages.

[4] Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. 2017. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife* 6 (sep 2017), e26726.

[5] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 2567–2577.

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (09 2019), 1234–1240.

[7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793.

[8] Jerry Liu. 2022. *LlamaIndex*. doi:10.5281/zenodo.1234

[9] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics* 23, 6 (09 2022), bbac409.

[10] John H Morris, Karthik Soman, Rabia E Akbas, Xiaoyuan Zhou, Brett Smith, Elaine C Meng, Conrad C Huang, Gabriel Cerono, Gundolf Schenk, Angela Rizk-Jackson, Adil Harroud, Lauren Sanders, Sylvain V Costes, Krish Bharat, Arjun Chakraborty, Alexander R Pico, Taline Mardirossian, Michael Keiser, Alice Tang, Josef Hardi, Yongmei Shi, Mark Musen, Sharat Israni, Sui Huang, Peter W Rose, Charlotte A Nelson, and Sergio E Baranzini. 2023. The scalable precision medicine open knowledge engine (SPOKE): a massive knowledge graph of biomedical information. *Bioinformatics* 39, 2 (02 2023), btad080.

[11] Alan Rector and Jeremy Rogers. 2006. *Ontological and Practical Issues in Using a Description Logic to Represent Medical Concept Systems: Experience from GALEN*. Springer Berlin Heidelberg, Berlin, Heidelberg, 197–231.

[12] Justin T. Reese, Deepak Unni, Tiffany J. Callahan, Luca Cappelletti, Vida Ravanmehr, Seth Carbon, Kent A. Shefchek, Benjamin M. Good, James P. Balhoff, Tommaso Fontana, Hannah Blau, Nicolas Matentzoglu, Nomi L. Harris, Monica C. Munoz-Torres, Melissa A. Haendel, Peter N. Robinson, Marcin P. Joachimiak, and Christopher J. Mungall. 2021. KG-COVID-19: A Framework to Produce Customized Knowledge Graphs for COVID-19 Response. *Patterns* 2, 1 (Jan. 2021). Publisher: Elsevier.

[13] Brandon T. Willard and Rémi Louf. 2023. Efficient Guided Generation for Large Language Models.

[14] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems* 2, 1 (2025), 1–5.

[15] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]