Can Large Language Models Enhance Predictions of Disease Progression? Investigating Through Disease Network Link Prediction

Haohui Lu
Faculty of Engineering
The University of Sydney
Sydney, Australia
haohui.lu@sydney.edu.au

Usman Naseem
School of Computing
Macquarie University
Sydney, Australia
usman.naseem@mq.edu.au

Abstract

Large Language Models (LLMs) have made significant strides in various tasks, yet their effectiveness in predicting disease progression remains relatively unexplored. To fill this gap, we use LLMs and employ advanced graph prompting and Retrieval-Augmented Generation (RAG) to predict disease comorbidity within disease networks. Specifically, we introduce a disease **Com**orbidity prediction model using **LLM**, named ComLLM, which leverages domain knowledge to enhance the prediction performance. Based on the comprehensive experimental results, ComLLM consistently outperforms conventional models, such as Graph Neural Networks, achieving average area under the curve (AUC) improvements of 10.70% and 6.07% over the best baseline models in two distinct disease networks. ComLLM is evaluated across multiple settings for disease progression prediction, employing various prompting strategies, including zero-shot, few-shot, Chain-of-Thought, graph prompting and RAG. Our results show that graph prompting and RAG enhance LLM performance in disease progression prediction tasks. ComLLM exhibits superior predictive capabilities and serves as a proof-of-concept for LLM-based systems in disease progression prediction, highlighting its potential for broad applications in healthcare[1].

1 Introduction

The digital transformation of healthcare through Artificial Intelligence (AI) has reshaped health management (Jin et al., 2024). Electronic health records now provide a rich source of data for predictive analytics, improving patient care by forecasting outcomes such as mortality rates (Blom et al., 2019), length of stay (Levin et al., 2021), and disease progression (Lu et al., 2022; Uddin et al., 2023). Nevertheless, traditional methods relying on static data and uniform standards are insufficient for addressing individual patient needs (Shoham and Rappoport, 2023). This presents a significant challenge in utilising this extensive data for proactive health management.

This challenge extends into the crucial areas of disease progression and comorbidity prediction in healthcare research (Barnett et al., 2012). Comorbidity, or the co-occurrence of multiple health conditions, not only complicates clinical management but also leads to worse health outcomes and higher healthcare costs (Valderas et al., 2009). Conditions like arthritis and cardiovascular disease often coexist, emphasising the need for predictive models that focus on the relationships between comorbidities rather than solely on lab results (Hidalgo et al., 2009).

To enhance predictive accuracy for disease comorbidity, researchers have developed various methods. Folino and Pizzuti (2012) created a comorbidity network and applied link prediction techniques to forecast chronic diseases based on patients' current health statuses. However, heuristic link prediction techniques can lack precision, scalability, and robustness, often failing to consider the features of the nodes and edges in the network (Ma et al., 2024). Liu et al. (2016) analysed hypertension comorbidities using network analytics but found that their method was not robust or generalisable, relying heavily on specific datasets. Recently, Lu and Uddin (2022) introduced a framework utilising graph learning to explore chronic disease comorbidity and progression patterns. Nevertheless, this approach faces challenges, including dataset acquisition costs and generalisability issues.

Large Language Models (LLMs), defined as advanced AI systems capable of processing and generating human-like text based on vast amounts of training data (Thirunavukarasu et al., 2023), have demonstrated exceptional capabilities across various NLP tasks in medicine (Li et al., 2024; Xie

[1]The source code is available at https://github.com/haohuilu/llm_lp

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 17703–17715
November 12-16, 2024

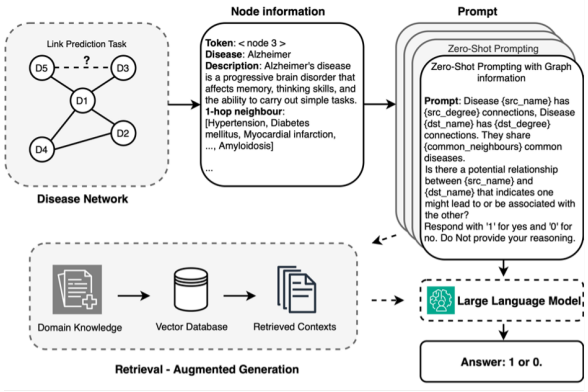

Figure 1: Overview of ComLLM. The process integrates domain knowledge and node-specific information to predict potential relationships between diseases.

et al., 2023). Despite these successes, applying LLMs in predicting disease progression and transforming medical data into actionable narratives is largely underexplored. To address the gap in disease progression research, we introduce a framework for Disease **Com**orbidity Prediction using **LLM**, named ComLLM, that utilises LLMs to extract and analyse disease networks. This method capitalises on the extensive clinical knowledge, employing LLMs to predict disease progression within disease networks.

Figure 1 illustrates a framework designed to predict potential disease progression within a network using an LLM. The framework starts by mapping diseases as nodes within a network, linked by known relationships. For each disease, such as Alzheimer's, comprehensive details including descriptions and 1-hop neighbours are utilised in a link prediction task to identify possible new relationships between diseases. This task incorporates a Retrieval-Augmented Generation (RAG) strategy, which enriches the model's predictions by integrating relevant domain knowledge. The LLM processes prompts that incorporate graph information to describe potential relationships between diseases and assesses the likelihood of a link, ultimately providing a binary response to indicate the presence or absence of a connection. This framework aims to significantly improve the accuracy of disease relationship predictions by combining disease network information with advanced NLP techniques.

Specifically, the system uses two disease networks, the human disease network (Goh et al., 2007) and the human symptoms-disease network (Zhou et al., 2014), with different graph sizes to compare the LLM's performance using various prompts and strategies. Our system incorporates automated disease feature generation from GPT-4. Additionally, the framework utilises the Langchain (LangChain, 2024) framework to retrieve domain knowledge via RAG. Finally, the system is tested with different prompts to improve prediction accuracy. Through comprehensive experiments across various settings (zero-shot, few-shot, Chain-of-Thought (COT), Graph Prompt and RAG), we observed that ComLLM surpasses numerous baseline models, and when integrated with RAG, it exceeds the performance of Graph Neural Network (GNN)-based models (Stamile et al., 2021; Lu and Uddin, 2023) in link prediction tasks. The experimental outcomes indicate that ComLLM, with the RAG, records average AUC enhancements of 10.70% and 6.07% over the best-performing model in baselines in the human disease network and the disease-symptoms network, respectively. Additionally, our method achieves these results with considerably fewer disease records and information, offering strong support for integrating open-world knowledge into healthcare prediction. Our main contributions include:

- We explore the effectiveness of LLMs in performing link prediction tasks within the domain of disease prediction. We enhance the LLMs' ability to predict the relationships in disease networks by incorporating graph information into the prompts. This study provides a comprehensive summary of how LLMs perform under various conditions and offers practical recommendations for leveraging LLMs in disease progression prediction.

- We investigate the application of LLMs in predicting disease progression by converting graph data into natural language narratives and evaluating their capabilities in zero-shot and few-shot learning environments. Additionally, we assess their efficacy when integrated with an RAG approach and various prompting strategies. This research introduces the innovative ComLLM framework, which combines feature generation, domain knowledge extraction, and LLM methodologies to enhance the analysis of disease comorbidities using healthcare data.

2 Related Works

Traditional Methods for Disease Prediction: The evolution of AI in disease prediction has transi-

tioned through various stages, from simple regression models to advanced deep learning, reflecting significant methodological advancements (Jin et al., 2024; Lu and Uddin, 2023; Uddin et al., 2019). Early methods, like DeepPatient (Miotto et al., 2016) and HRFLM (Mohan et al., 2019), often focused on existing data without considering extensive historical patient information. Subsequent models, including those for chronic kidney disease and COVID-19, began analysing sequential patient data, integrating deep learning techniques such as recurrent neural network-based Doctor AI (Choi et al., 2016) and attention-based GRAM for heart failure (Choi et al., 2016). Knowledge graph-based models, such as GNDP (Li et al., 2020) and GAT-ETM (Zou et al., 2022), have also shown promise. However, a common drawback of these models is their failure to connect to external medical domain knowledge, which is rich in valuable relational information and often overlooks interrelations between diseases.

Large Language Models in Disease Prediction: LLMs have demonstrated impressive capabilities across several natural language processing tasks in the healthcare domain. However, their potential to predict disease progression and convert medical data into actionable narratives remains untapped. Wang et al. (2023) proposed a framework called CoAD, which highlights the ability of LLMs to analyse health reports and assess medical conditions. Another significant advancement is the Clinical Prediction with Large Language Models (CPLLM) method, introduced by Shoham and Rappoport (2023), which leverages historical diagnosis data and outperforms traditional logistic regression and advanced models like Med-BERT (Rasmy et al., 2021) in predicting future disease diagnoses. Furthermore, Jin et al. (2024) developed the Health-LLM framework, combining feature extraction with medical knowledge scoring to enhance disease prediction and personalise healthcare. Jiang et al. (2023) introduced the Graph-CARE framework, which merges external knowledge graphs with electronic health records via a novel Bi-attention Augmented GNN, significantly improving key healthcare outcomes such as mortality, readmission, length of stay, and drug recommendation. Although these methods represent significant progress in disease diagnosis prediction, they have yet to completely unravel the complex interrelations among diseases, which is crucial for further enhancing their efficacy in predicting disease progression. LLMs in disease prediction often face limitations such as generating outputs that may be factually incorrect or irrelevant, particularly when dealing with complex data sets in healthcare (Ke et al., 2024). This is where RAG becomes crucial. RAG addresses these limitations by integrating external, verified information, thus enhancing the model's ability to produce accurate and relevant responses. For instance, in the Health-LLM framework, RAG uses external medical databases to refine its predictions and reduce the likelihood of erroneous outputs, ensuring that the advice and diagnostics provided are based on the most current and comprehensive data available (Kim et al., 2024). This integration is vital in healthcare, where the accuracy of information can significantly impact patient outcomes.

Link Prediction on Disease Networks: Initial studies on link prediction within disease networks relied on heuristic techniques that calculated node similarity based on the graph's structure to predict potential links from these similarity scores. However, these methods primarily focused on the structural properties of the graph, neglecting individual node features and the broader domain knowledge associated with each node (Aziz et al., 2021; Folino and Pizzuti, 2012). Advanced ML algorithms, including GNNs and matrix factorisation, have since been utilised to glean deeper insights from complex disease networks (Lu and Uddin, 2023). While GNNs effectively capture hierarchical and non-linear structures within the network, they often face challenges with complex medical data due to insufficient integration of external domain knowledge.

Building on previous research, our work uniquely applies LLMs to link prediction within disease networks. This approach surpasses traditional methodologies by dynamically integrating extensive medical knowledge, enabling a more precise understanding of disease progression.

3 Method

3.1 Data Preprocessing and Feature Generation for Diseases

In this study, we analyse a disease network where diseases are represented as nodes and links between nodes signify relationships between diseases, as depicted in the top left part of Figure 1. We utilise data on these relationships to predict unknown disease progressions. Initially, our disease network in-

cluded only labels for the nodes without additional features. To enrich this data, we employ GPT-4 to systematically extract disease features, as shown in Figure 2. This feature generation employs zero-shot prompting. To enhance the accuracy of our predictions, we integrate a medical knowledge base, such as PubMed (National Library of Medicine, 2024), and utilise a RAG mechanism for improved knowledge retrieval. Given the general nature of LLMs and their limited specialised medical knowledge, we embed contextual information within the prompts. This RAG technology ensures that our queries are aligned with the knowledge base, identifying and retrieving the most relevant information about the diseases referenced in our queries.

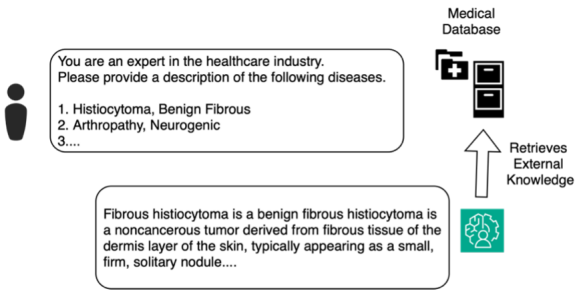

Figure 2: Disease features generation.

3.2 Integrating Graph Information for Enhanced Predictions

We aim to determine if LLMs can enhance disease comorbidity predictions by utilising features from disease networks.

Figure 3 illustrates the differences between zero-shot prompting without disease network information and zero-shot prompting with graph information for predicting disease progression. The standard zero-shot prompting method directly queries LLMs about potential relationships between two diseases. In contrast, zero-shot prompting with graph information enriches the query by including structural data from the disease networks, such as the degree of the source and target nodes and their common neighbors. In the advanced prompt strategy, features such as node centrality and anchor nodes are incorporated into the prompt input. We hypothesise that integrating graph reasoning with LLMs will enhance disease progression predictions by utilising the structural and relational information inherent in disease networks. This method aims to provide a nuanced understanding that could lead to more accurate and clinically relevant predictions.

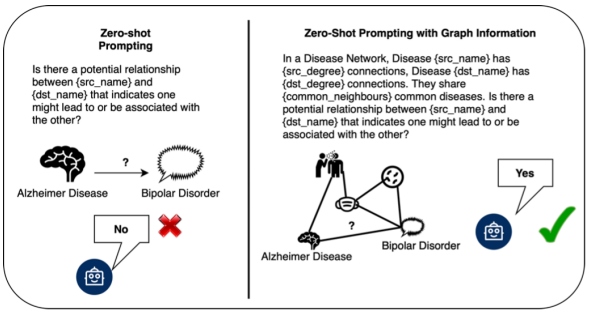

Figure 3: Using features from disease networks for prompting.

3.3 Prompting

In our study, we employed various prompt engineering techniques to enhance the performance of LLMs in predicting disease comorbidity. Zero-shot learning was used to establish a baseline by assessing the LLM's ability to infer disease relationships based on its general knowledge without specific training. Few-shot learning followed, providing the model with a small but crucial set of examples from our disease network to refine its accuracy in identifying disease links. The COT approach structured prompts to guide the LLM through complex reasoning tasks, improving its capacity to handle intricate medical queries. Lastly, Graph Prompting incorporated specific structural features from disease networks, such as node connectivity and centrality, enabling the LLM to utilise relational data more effectively. These methods collectively tailored the LLM's capabilities to the unique challenges of predicting disease comorbidity, leveraging its inherent strengths and the detailed information from disease networks. The prompt example is shown in Appendix A.

4 Experimental Settings

Datasets: For the disease network dataset, we utilise publicly available datasets: the human disease network (Goh et al., 2007) and the disease-symptom network (Zhou et al., 2014). Table 1 displays the graph statistics for these disease networks. The "Human disease network" and the "Disease-symptom network" exhibit notable differences in their structures. The former has 516 nodes and 1,188 edges, with an average degree of 4.6047 and a clustering coefficient of 0.6358, suggesting tight clustering. It also shows a slight preference for connecting similar nodes with an assortativity of 0.0666. In contrast, the larger size "Disease-symptom network" contains 1,596 nodes

Dataset	No. of nodes	No. of edges	Avg degree	Avg Clustering	Assortativity	Density
Human disease network	516	1,188	4.6047	0.6358	0.0666	0.0089
Disease-symptom network	1596	1,133,106	166.7995	0.5941	-0.1878	0.1046

Table 1: Network characteristics for different datasets

and 1,133,106 edges, with a significantly higher average degree of 166.7995 and a clustering coefficient of 0.5941. However, it has a negative assortativity of -0.1878, indicating diverse node connections. These metrics reflect the distinct interaction dynamics within each network. Further, the disease-symptom network, with a density of 0.1046, is much more interconnected, showing that diseases share many common symptoms. In contrast, the human disease network is sparser, with a density of 0.0089, indicating fewer connections between diseases. To construct features for the disease entities, we employ GPT-4 (Model name: *GPT-4-turbo*) (Achiam et al., 2023) as the language model for generating features. We use the *text-embedding-3-large* embedding model from OpenAI (OpenAI, 2024a) to obtain word embeddings for the baseline models, such as GNN-based models. For the RAG experiment, we utilised 892 papers published in April 2024 from PubMed (National Library of Medicine, 2024), selected for its coverage of peer-reviewed biomedical literature. This choice demonstrated our current study; however, we are open to exploring other sources based on different business cases or research needs. We also employed the *text-embedding-3-large* model to embed the articles and convert them into a vector database using LangChain (LangChain, 2024).

Baselines: We assess the performance of our proposed model by comparing it with leading link prediction methods. Initially, we employ heuristic approaches, such as Common Neighbour (CN) and Adamic-Adar index (AA) (Liben-Nowell and Kleinberg, 2003), which have been widely used in prior research (Aziz et al., 2021; Folino and Pizzuti, 2012). Subsequently, we explore more advanced methods, such as Matrix Factorisation (MF), which integrates latent and explicit features within the graph. In this method, each node is assigned various embeddings, which are trained end-to-end using a multilayer perceptron (MLP) predictor (Menon and Elkan, 2011). We use an embedding size of 64 and train for 100 epochs for this method. Following this, we implement Node2Vec, a method that maps nodes into an embedding space (Grover and Leskovec, 2016). In this study, these embeddings are utilised in a downstream link prediction task using a three-layer MLP. The dimensions are set to 64, the walk length to 30, and the number of walks to 200. Additionally, we benchmark our model against the latest GNN-based models. Graph Convolutional Network (GCN), a type of GNN, utilises convolutional neural networks on graphs (Kipf and Welling, 2016), while GraphSAGE is another GNN variant designed for inductive representation learning on large graphs (Hamilton et al., 2017). Moreover, Contrastive Multi-View Representation Learning on Graphs (MVGRL) employs a self-supervised approach to learn representations by contrasting structural graph views (Hassani and Khasahmadi, 2020), and learning from Subgraphs, Embeddings, and Attributes for Link prediction (SEAL) uses GNNs in a deterministic manner, leveraging node features and hand-crafted labels (Zhang and Chen, 2018). Furthermore, we include recent GNN-based disease progression models, CPLLM (Shoham and Rappoport, 2023) and Health-LLM (Jin et al., 2024), as well as the LLM-based disease prediction model, GraphPrompter (Liu et al., 2024), for comparison. For these baselines, we used default settings for the hyperparameters in DGL (Wang et al., 2019).

Evaluation setting: The focus of this study is on link prediction, defined as a binary classification challenge. In this context, the inputs to the model are pairs of nodes. To construct these inputs, we split the positive edges of our network into training and testing sets. During the training phase, we generate an equivalent number of negative edges through random sampling. Specifically, for the dis-

ease network, we allocate 10% of the existing edges to the train set as positive samples and another 10% to the test set while also creating an equal number of non-existent edges as negative samples. In the case of the broader disease symptoms network, we use 10% of the existing edges for training and 1% for testing, corresponding to 11,311 positive edges, and similarly generate an equal number of non-existent edges for negative samples. For the LLMs, we utilise GPT-3.5 and GPT-4-Turbo from OpenAI (OpenAI, 2024b), LLaMA 2, LLaMA 3 and LLaMA 3.1 from Meta AI (Meta, 2024). GPT-3.5 acts as a transitional model between GPT-3 and GPT-4, providing enhanced reasoning and reduced computational needs. GPT-4 expands on this with more advanced algorithms and a larger training dataset, significantly improving accuracy and knowledge breadth. LLaMA 2, LLaMA 3 and LLaMA 3.1, developed by Meta AI, offer a scalable and efficient architecture, allowing customisation to match specific computational limits and application requirements. These models are instrumental in advancing AI applications across various fields, reflecting rapid innovations and the increasing sophistication of language models. The specific models we use for GPT-3, GPT-4, Llama 2, Llama 3, and Llama 3.1 are *GPT-3.5-Turbo*, *GPT-4-Turbo*, *Llama-2-7b-hf*, *Meta-Llama-3-8B* and *Meta-Llama-3.1-405B-Instruct*, respectively. We set the temperature to zero and the maximum token to 64. In assessing the performance of models for link prediction tasks, we use the Area Under the ROC Curve (AUC), Average Precision (AP), and F1 score metrics. The AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which depicts the true positive rate versus the false positive rate at different classification thresholds (Huang and Ling, 2005). The testing case sets the threshold at a balanced 50% true positive rate and 50% false positive rate, given an equal number of positive and negative data points. A higher AUC value suggests that the model is more effective at distinguishing between positive and negative examples. The AP metric summarises the precision-recall curve as the weighted average of precision at each threshold, using the increment in recall from the previous threshold as the weight (Huang and Ling, 2005). This metric is particularly useful for evaluating model performance in terms of precision (accuracy of positive predictions) and recall (capacity to identify positive instances), especially in cases of class imbalance. The Macro F1 score, which ranges from 0 to 1, reflects the balance between precision and recall, with higher scores indicating better performance (Sai et al., 2022). We ran all the baselines and LLMs, testing one variation for each prompt example, and averaged the performance results over five runs; the numbers in brackets represent the standard deviation. The baseline graph machine learning methods are executed using the DGL (Wang et al., 2019) and Networkx (Hagberg et al., 2008) libraries in Python. The experiments are conducted on a server equipped with an A40 GPU with 48 GB of memory.

5 Results and Discussion

5.1 Main Results

As illustrated in Table 2, ComLLM significantly surpasses all existing baselines on link prediction tasks within both the human disease network and the human symptoms-disease datasets. Specifically, when integrated with SEAL, the top-performing model in our baseline, ComLLM with GPT-4 enhances the best SEAL results by +10.70% in AUC, +11.81% in AP, and +11.85% in F1 for the disease network. At the same time, our model outperformed SEAL by +6.07% in AUC, +8.74%, and +9.65% in F1 for the human symptoms-disease network. Notably, within the ComLLM framework, our newly introduced Graph Prompt with RAG achieves the highest performance consistently across both datasets, demonstrating the strong impact of these components.

In evaluating ComLLM's performance on Human Disease Network and Human Symptoms-Disease Network data using different prompting techniques, we observe significant variations across metrics, such as AUC, AP, and F1 in Table 3. In the zero-shot setting, performance is relatively modest, with an AUC of approximately 0.6537 and 0.6673 for the respective networks, improving incrementally as more context or instructions are introduced. Progressing to few-shot settings, an improvement is noticed with AUC values of 0.7817 and 0.7891. Including COT prompts further boosts the model's accuracy, demonstrating AUC values of 0.8033 and 0.8133. Adding graph prompts alongside few-shot and COT prompting strategies leads to superior outcomes, with AUCs peaking at 0.8451 and 0.8633. Lastly, integrating RAG with the previous techniques maximises performance, reaching AUCs of 0.8898 and 0.9012, showcasing a sophisticated understanding and retrieval of relevant medical knowl-

Method	Human disease Network			Human symptoms-disease network		
	AUC	AP	Macro F1	AUC	AP	Macro F1
CN	0.5198	0.5198	0.6667	0.5015	0.5015	0.6601
AA	0.5232	0.5232	0.6667	0.5003	0.5003	0.6599
MF	0.7559 (0.0609)	0.6631 (0.0504)	0.7758 (0.0323)	0.6898 (0.0048)	0.6708 (0.0116)	0.6755 (0.0061)
Node2Vec	0.5678 (0.0405)	0.5814 (0.0289)	0.6833 (0.0169)	0.5207 (0.0345)	0.5266 (0.0227)	0.6631 (0.0232)
GCN	0.7553 (0.0103)	0.7402 (0.0133)	0.7218 (0.0096)	0.7804 (0.0748)	0.7924 (0.0823)	0.7002 (0.0922)
GraphSAGE	0.7947 (0.0233)	0.7337 (0.0283)	0.7619 (0.0298)	0.8343 (0.0892)	0.8230 (0.0775)	0.7557 (0.0687)
MVGRL	0.8077 (0.0205)	0.7729 (0.0199)	0.7728 (0.0186)	0.8414 (0.0992)	0.8335 (0.0892)	0.7669 (0.0823)
SEAL	0.8038 (0.0199)	0.8010 (0.0178)	0.8035 (0.0187)	0.8496 (0.0328)	0.8366 (0.0339)	0.8238 (0.0429)
CPLLM	0.6588 (0.0945)	0.6423 (0.0889)	0.6789 (0.0956)	0.6557 (0.0898)	0.6455 (0.0789)	0.6776 (0.0698)
Health-LLM	0.7898 (0.0456)	0.7813 (0.0552)	0.7623 (0.0569)	0.8127 (0.0456)	0.8099 (0.0599)	0.8051 (0.0398)
GraphPrompter	0.6427 (0.0899)	0.6498 (0.0832)	0.6991 (0.0851)	0.6485 (0.1189)	0.6411 (0.1185)	0.6877 (0.1288)
ComLLM - Llama 2	0.6238 (0.0776)	0.6277 (0.0687)	0.6876 (0.0448)	0.6325 (0.1276)	0.6377 (0.0887)	0.6671 (0.1048)
ComLLM - Llama 3	0.8037 (0.0438)	0.8151 (0.0534)	0.8029 (0.0448)	0.8273 (0.1031)	0.8261 (0.0994)	0.7792 (0.1442)
ComLLM - Llama 3.1	0.8149 (0.0196)	0.8199 (0.0231)	0.8109 (0.0259)	0.8499 (0.0989)	0.8421 (0.0883)	0.7873 (0.0.0989)
ComLLM - GPT 3.5	0.6477 (0.0334)	0.6907 (0.0332)	0.6924 (0.0422)	0.6768 (0.0984)	0.6905 (0.1232)	0.6824 (0.0922)
ComLLM - GPT 4	**0.8898** (0.0343)	**0.8956** (0.0387)	**0.8987** (0.0399)	**0.9012** (0.0886)	**0.9098** (0.0889)	**0.9033** (0.0889)

Table 2: Performance metrics in Human disease Network and Human symptoms-disease network, We report the average performance and standard deviation (in brackets) of each model over five runs. Bold denotes the best performances for the proposed method. Underline denotes the best-performed baseline.

edge. We observe a similar pattern for AP and F1 from Table 3. Below, we discuss the experimental findings of our study.

- LLMs, especially GPT4, demonstrate a profound ability in foundational graph reasoning, substantially outperforming the state-of-the-art model, SEAL. This significant advancement in performance underscores the robust computational power and adaptability of LLMs to complex network analysis tasks, marking a notable shift in the landscape of graph-based predictive modelling. These experiments highlighted LLMs superior performance, showcasing its robust capabilities in link prediction task on graph and predictive modelling in the healthcare.
- Incorporating few-shot demonstrations improves prediction performance compared to their zero-shot counterparts. This observation highlights how even a few labelled examples can guide language models towards more accurate predictions. LLMs can swiftly adjust to the unique features of the disease prediction by using a select group of representative samples.
- While in-context learning has been widely successful in enabling LLMs to learn directly from examples as demonstrated in foundational re-

Method	Human disease Network			Human symptoms-disease network		
	AUC	AP	Macro F1	AUC	AP	Macro F1
Zero-shot	0.6537 (0.0323)	0.6607 (0.0467)	0.6614 (0.0367)	0.6673 (0.0886)	0.6695 (0.1011)	0.6810 (0.0823)
Few-shot	0.7817 (0.0499)	0.7909 (0.0627)	0.7832 (0.0422)	0.7891 (0.1102)	0.7961 (0.1398)	0.7810 (0.1331)
Few-shot + COT	0.8033 (0.0309)	0.8111 (0.0327)	0.8001 (0.0229)	0.8133 (0.0582)	0.8145 (0.0698)	0.8332 (0.0531)
Zero-shot + Graph Prompt	0.8245 (0.0299)	0.8317 (0.0304)	0.8222 (0.0288)	0.8408 (0.0592)	0.8424 (0.0678)	0.8411 (0.0593)
Few-shot + Graph Prompt	0.8356 (0.0276)	0.8417 (0.0341)	0.8322 (0.0298)	0.8412 (0.0602)	0.8454 (0.0668)	0.8423 (0.0582)
Few-shot + COT + Graph Prompt	0.8451 (0.0409)	0.8576 (0.0401)	0.8539 (0.0443)	0.8633 (0.0382)	0.8645 (0.0489)	0.8632 (0.0531)
Few-shot + COT + Graph Prompt + RAG	**0.8898** (0.0343)	**0.8956** (0.0387)	**0.8987** (0.0399)	**0.9012** (0.0886)	**0.9098** (0.0889)	**0.9033** (0.0889)

Table 3: Performance metrics for GPT-4 in Human disease Network and Human symptoms-disease network. We report the average performance and standard deviation (in brackets) of each model over five runs. Bold denotes the best performances for the proposed method.

search (Brown et al., 2020), its impact appears to be less pronounced in more sophisticated graph reasoning tasks. Few-shot in-context learning, in particular, shows fewer performance gains in larger graph link prediction tasks, indicating that the learning strategies may need further adaptation to fully exploit the potential of LLMs in handling more complex and larger-scale graph analytical challenges.

- Although advanced prompting methods, such as COT, are less effective compared to graph prompting, they still contribute positively by marginally enhancing the graph reasoning skills of LLMs for disease research. This suggests that while some prompting techniques may not fully leverage the inherent strengths of LLMs in graph reasoning, they still play a supportive role in refining the models' analytical processes.
- In comparing the performance of LLMs with traditional machine learning methods, we observe that LLMs exhibit higher AP but lower AUC. This result shows that LLMs are particularly adept at accurately identifying positive cases (i.e., there is a link between two diseases). Typically, LLMs are more likely to classify diseases as connected, even when they are not. This tendency indicates that LLMs, particularly advanced models, adopt a more cautious strategy, probably driven by a design preference to avoid missing true positive cases. This conservative approach in LLMs may be deliberately employed to reduce the risk of overlooking critical connections in disease progression.
- The integration of RAG with LLMs enhances their functionality significantly. This approach enables LLMs to generate accurate and logically coherent intermediate steps, which are crucial for complex reasoning tasks. Such integration showcases the LLMs' ability to synthesise and utilise relevant information effectively, elevating their problem-solving capabilities in dynamic environments.

5.2 Ablation study

We conducted an ablation study for Llama 3 8B and Llama 3.1 405B as part of our evaluation for ComLLM. Table 4 present the performance impact of various prompt engineering strategies, including zero-shot, few-shot, COT, and graph prompting, both independently and in combination for Llama 3 8B and Llama 3.1 405B, respectively. The results showed how different configurations contribute to the model's overall effectiveness, highlighting the

Method	Human Disease Network						Human Symptoms-Disease Network					
	Llama 3 8B			Llama 3.1 405B			Llama 3 8B			Llama 3.1 405B		
	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1	AUC	AP	F1
Zero-shot	0.6037 (0.0379)	0.6005 (0.0248)	0.6008 (0.0302)	0.6142 (0.0299)	0.6105 (0.0218)	0.6101 (0.0299)	0.6413 (0.0279)	0.6418 (0.0411)	0.6227 (0.0118)	0.6533 (0.0284)	0.6511 (0.0423)	0.6328 (0.0123)
Few-shot	0.7421 (0.0191)	0.7365 (0.0326)	0.7312 (0.0129)	0.7522 (0.0201)	0.7485 (0.0326)	0.7400 (0.0133)	0.7354 (0.0792)	0.7307 (0.0588)	0.7064 (0.0688)	0.7625 (0.0802)	0.7607 (0.0598)	0.7564 (0.0671)
Few-shot + COT	0.7553 (0.0188)	0.7538 (0.0191)	0.7423 (0.0111)	0.7635 (0.0139)	0.7631 (0.0177)	0.7523 (0.0189)	0.7198 (0.0282)	0.7167 (0.0319)	0.7012 (0.0381)	0.7998 (0.0222)	0.7900 (0.0323)	0.7812 (0.0338)
Zero-shot + Graph Prompt	0.7717 (0.0197)	0.7681 (0.0132)	0.7654 (0.0191)	0.7881 (0.0139)	0.7801 (0.0121)	0.7754 (0.0201)	0.8019 (0.0281)	0.8001 (0.0335)	0.7577 (0.0313)	0.8219 (0.0291)	0.8201 (0.0345)	0.8177 (0.0323)
Few-shot + Graph Prompt	0.7877 (0.0198)	0.7865 (0.0182)	0.7782 (0.0198)	0.7993 (0.0201)	0.7912 (0.0191)	0.7901 (0.0178)	0.8092 (0.0442)	0.8085 (0.0605)	0.7623 (0.0596)	0.8412 (0.0602)	0.8454 (0.0668)	0.8423 (0.0582)
Few-shot + COT + Graph Prompt	0.7966 (0.0126)	0.7921 (0.0213)	0.7888 (0.0211)	0.8012 (0.0109)	0.8007 (0.0213)	0.7998 (0.0249)	0.8133 (0.0306)	0.8145 (0.0479)	0.7688 (0.0331)	0.8633 (0.0382)	0.8645 (0.0489)	0.8632 (0.0531)
Few-shot + COT + Graph Prompt + RAG	**0.8037** (0.0438)	**0.8151** (0.0534)	**0.8029** (0.0448)	**0.8149** (0.0196)	**0.8199** (0.0231)	**0.8109** (0.0259)	**0.8273** (0.1031)	**0.8261** (0.0994)	**0.7792** (0.1142)	**0.8499** (0.0989)	**0.8421** (0.0883)	**0.7873** (0.0989)

Table 4: Performance metrics for Llama 3 8B and Llama 3.1 405B in the Human Disease Network and Human Symptoms-Disease Network. The average performance and standard deviation (in brackets) are reported, with bold denoting the best performances.

improvements achieved by integrating advanced techniques like COT and RAG[2].

5.3 Discussion

Disease networks are invaluable for disease prediction, yet traditional methods relying on supervised learning require extensive, costly labeled datasets. Our study used various LLMs with different prompting techniques and RAG to predict disease comorbidity in two differently sized disease networks. Results show that ComLLM excels in disease progression prediction, surpassing SEAL, with few-shot and RAG demonstrations enhancing accuracy. While gains in complex graph tasks were less pronounced, advanced prompts still improved LLM capabilities. ComLLMs displayed higher AP but lower AUC than traditional methods, indicating a focus on avoiding false negatives. Further, integrating RAG into our method enhances logical coherence and problem-solving.

Our findings suggest several avenues for future research: enhancing LLM performance through different input designs and advanced prompting techniques, such as Least-To-Most and Automatic Reasoning; further exploring RAG integration's impact on accuracy and coherence; and developing new techniques for graph augmentation to improve semantic understanding in disease progression tasks. Additionally, our experiments highlight the potential of LLMs in tasks beyond traditional graph learning methods, advocating for the integration of graph-based information and the application of LLMs to other graph-related tasks.

6 Conclusion

In this study, we introduce ComLLM, a framework that merges graph prompts and RAG with LLMs for improved disease progression predictions in disease networks. Empirical evidence demonstrates its superiority over baseline models, such as GNN-based models, across two datasets with different graph sizes. Our framework can predict potential disease comorbidities effectively. Beyond its practical improvements, ComLLM also acts as a proof-of-concept for LLM-based systems in healthcare, underscoring the significant potential for further developing AI applications in the health sector.

Limitation

Despite the promising directions indicated by our research, there are several limitations to consider. The COT prompting technique was less effective in our experiments, indicating a need for alternative

[2]We have also provided the computational costs and inference time of different models on both datasets in Appendix B.

strategies. Although RAG integration showed potential, it requires more robust methods to ensure reliable accuracy and coherence. Our results, based on specific disease networks, may not generalise to other networks or diseases, which could limit the applicability of our findings. Additionally, using LLMs, especially for complex tasks, demands significant computational resources, a crucial consideration for practical applications. Addressing these challenges will enhance the application of LLMs in predicting disease progression and expand their utility in healthcare more broadly.

Ethics Statement

Our research strictly adheres to ethical standards, utilising only open-source datasets to ensure transparency and reproducibility. These datasets are freely available under licenses suitable for academic and non-commercial use, supporting an open and collaborative research environment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Furqan Aziz, Victor Roth Cardoso, Laura Bravo-Merodio, Dominic Russ, Samantha C Pendleton, John A Williams, Animesh Acharjee, and Georgios V Gkoutos. 2021. Multimorbidity prediction using link prediction. *Scientific Reports*, 11(1):16392.

Karen Barnett, Stewart W Mercer, Michael Norbury, Graham Watt, Sally Wyke, and Bruce Guthrie. 2012. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. *The Lancet*, 380(9836):37–43.

Mathias Carl Blom, Awais Ashfaq, Anita Sant'Anna, Philip D Anderson, and Markus Lingman. 2019. Training machine learning models to predict 30-day mortality in patients discharged from the emergency department: a retrospective, population-based registry study. *BMJ open*, 9(8):e028015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine learning for healthcare conference*, pages 301–318. PMLR.

Francesco Folino and Clara Pizzuti. 2012. Link prediction approaches for disease networks. In *International Conference on Information Technology in Bio- and Medical Informatics*, pages 99–108. Springer.

Kwang-Il Goh, Michael E Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Will Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

Kaveh Hassani and Amir Hosein Khasahmadi. 2020. Contrastive multi-view representation learning on graphs. In *International conference on machine learning*, pages 4116–4126. PMLR.

César A Hidalgo, Nicholas Blumm, Albert-László Barabási, and Nicholas A Christakis. 2009. A dynamic network approach for the study of human phenotypes. *PLoS computational biology*, 5(4):e1000353.

Jin Huang and Charles X Ling. 2005. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310.

Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. 2023. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. *arXiv preprint arXiv:2402.00746*.

YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, and Daniel Shu Wei Ting. 2024. Development and testing of retrieval augmented generation in large language models–a case study report. *arXiv preprint arXiv:2402.01733*.

Yubin Kim, Xuhai Xu, Daniel McDuff, Cynthia Breazeal, and Hae Won Park. 2024. Health-llm: Large language models for health prediction via wearable sensor data. *arXiv preprint arXiv:2401.06866*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

LangChain. 2024. *LangChain Documentation*. Accessed: 2024-05-16.

Scott Levin, Sean Barnes, Matthew Toerper, Arnaud Debraine, Anthony DeAngelo, Eric Hamrock, Jeremiah Hinson, Erik Hoyer, Trushar Dungarani, and Eric Howell. 2021. Machine-learning-based hospital discharge predictions can support multidisciplinary rounds and decrease hospital length-of-stay. *BMJ Innovations*, 7(2).

Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. Pre-trained language models for text generation: A survey. *ACM Computing Surveys*, 56(9):1–39.

Yang Li, Buyue Qian, Xianli Zhang, and Hui Liu. 2020. Knowledge guided diagnosis prediction via graph spatial-temporal network. In *Proceedings of the 2020 SIAM International Conference on Data Mining*, pages 19–27. SIAM.

David Liben-Nowell and Jon Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559.

Jiaqi Liu, James Ma, Jiaojiao Wang, Daniel Dajun Zeng, Hongbin Song, Ligui Wang, and Zhidong Cao. 2016. Comorbidity analysis according to sex and age in hypertension patients in china. *International journal of medical sciences*, 13(2):99.

Zheyuan Liu, Xiaoxin He, Yijun Tian, and Nitesh V Chawla. 2024. Can we soft prompt llms for graph learning tasks? In *Companion Proceedings of the ACM on Web Conference 2024*, pages 481–484.

Haohui Lu and Shahadat Uddin. 2022. Embedding-based link predictions to explore latent comorbidity of chronic diseases. *Health information science and systems*, 11(1):2.

Haohui Lu and Shahadat Uddin. 2023. Disease prediction using graph machine learning based on electronic health data: a review of approaches and trends. In *Healthcare*, volume 11, page 1031. MDPI.

Haohui Lu, Shahadat Uddin, Farshid Hajati, Mohammad Ali Moni, and Matloob Khushi. 2022. A patient network-based machine learning model for disease prediction: The case of type 2 diabetes mellitus. *Applied Intelligence*, 52(3):2411–2422.

Li Ma, Haoyu Han, Juanhui Li, Harry Shomer, Hui Liu, Xiaofeng Gao, and Jiliang Tang. 2024. Mixture of link predictors. *arXiv preprint arXiv:2402.08583*.

Aditya Krishna Menon and Charles Elkan. 2011. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part II 22*, pages 437–452. Springer.

Meta. 2024. *Meta Llama*. Accessed: 2024-05-16.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10.

Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava. 2019. Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, 7:81542–81554.

National Library of Medicine. 2024. *PubMed*. Accessed: 2024-05-16.

OpenAI. 2024a. *Embeddings*. Accessed: 2024-05-16.

OpenAI. 2024b. *Models*. Accessed: 2024-05-16.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Ofir Ben Shoham and Nadav Rappoport. 2023. Cpllm: Clinical prediction with large language models. *arXiv preprint arXiv:2309.11295*.

Claudio Stamile, Aldo Marzullo, and Enrico Deusebio. 2021. *Graph Machine Learning: Take graph data to the next level by applying machine learning techniques and algorithms*. Packt Publishing Ltd.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Shahadat Uddin, Arif Khan, Md Ekramul Hossain, and Mohammad Ali Moni. 2019. Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, 19(1):1–16.

Shahadat Uddin, Shangzhou Wang, Arif Khan, and Haohui Lu. 2023. Comorbidity progression patterns of major chronic diseases: the impact of age, gender and time-window. *Chronic Illness*, 19(2):304–313.

Jose M Valderas, Barbara Starfield, Bonnie Sibbald, Chris Salisbury, and Martin Roland. 2009. Defining comorbidity: implications for understanding health and health services. *The Annals of Family Medicine*, 7(4):357–363.

Huimin Wang, Wai Chung Kwan, Kam-Fai Wong, and Yefeng Zheng. 2023. CoAD: Automatic diagnosis through symptom and disease collaborative generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6348–6361, Toronto, Canada. Association for Computational Linguistics.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. 2019. Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

Qianqian Xie, Edward J Schenck, He S Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful ai in medicine: A systematic review with large language models and beyond. *medRxiv*.

Muhan Zhang and Yixin Chen. 2018. Link prediction based on graph neural networks. *Advances in neural information processing systems*, 31.

XueZhong Zhou, Jörg Menche, Albert-László Barabási, and Amitabh Sharma. 2014. Human symptoms–disease network. *Nature communications*, 5(1):4212.

Yuesong Zou, Ahmad Pesaranghader, Ziyang Song, Aman Verma, David L Buckeridge, and Yue Li. 2022. Modeling electronic health record data using an end-to-end knowledge-graph-informed topic model. *Scientific Reports*, 12(1):17868.

Appendix

A Prompt Example

Zero-shot

Is there a potential relationship between Mitral Valve Prolapse and Hyperthyroidism. That indicates one might lead to or be associated with the other. Evaluate and respond with '1' for a strong link and '0' for a weak or no link. Do Not provide your reasoning.

Few-shot

For example: There are links in the following diseases: Node ID: 765, Disease: Genital Neoplasms, Female has relationship with Node ID: 845, Disease: Female Urogenital Diseases; Node ID: 717, Disease: Urinary Bladder Neoplasms has relationship with Node ID: 1230, Disease: Situs Inversus.
Is there a potential relationship between Mitral Valve Prolapse and Hyperthyroidism. That indicates one might lead to or be associated with the other. Evaluate and respond with '1' for a strong link and '0' for a weak or no link. Do Not provide your reasoning.

Chain-of-Thought

Step-by-step, analyse whether there is a potential relationship between Mitral Valve Prolapse and Hyperthyroidism. That indicates one might lead to or be associated with the other. Evaluate and respond with '1' for a strong link and '0' for a weak or no link. Do Not provide your reasoning.

Graph Prompt

In a disease network, Disease Mitral Valve prolapse has 35 connections, Disease Hyperthyroidism has 27 connections. They share 18 common diseases. Is there a potential relationship between Mitral Valve Prolapse and Hyperthyroidism. That indicates one might lead to or be associated with the other. Evaluate and respond with '1' for a strong link and '0' for a weak or no link. Do Not provide your reasoning.

B Computational Costs and Inference Time

Model	Human Disease Network		Human Symptoms-Disease Network	
	Time (Approx)	Cost (Approx)	Time (Approx)	Cost (Approx)
CN	10 s	N/A	20 s	N/A
AA	10 s	N/A	20 s	N/A
MF	5 min	N/A	10 min	N/A
Node2Vec	5 min	N/A	10 min	N/A
GCN	10 min	N/A	10 min	N/A
GraphSAGE	10 min	N/A	10 min	N/A
MVGRL	10 min	N/A	10 min	N/A
SEAL	20 min	N/A	20 min	N/A
CPLLM	20 min	$3	25 min	$10
Health-LLM	20 min	$3	25 min	$10
GraphPrompter	20 min	$0.15	25 min	$0.5
Llama 2 7B	20 min	$0.15	25 min	$0.5
Llama 3 8B	20 min	$0.3	25 min	$1
Llama 3.1 405B	20 min	$1.5	25 min	$5
GPT3.5	20 min	$0.3	25 min	$1
GPT4	20 min	$3	25 min	$10

Table 5: Execution time and cost for different models on the Human Disease Network and Human Symptoms-Disease Network.