Research practice III

Final Report

Causal Analysis of COVID-19 utilizing Knowledge Graphs and GPT-Driven Chat Models

Valentina Moreno[†], Elizabeth Ramírez and Juan David Martínez[‡*]
vmorenor@eafit.edu.co, eramirem@gmail.com, jdmartinev@eafit.edu.co

[†]Mathematical Engineering, Universidad EAFIT
[‡]Computing and Analytics, School of Applied Sciences and Engineering, Universidad EAFIT

November 15, 2024

Abstract

This study addresses the challenge of understanding COVID-19's complex causal relationships by combining knowledge graphs with GPT-driven chat models. Using the CORD-19 dataset, we developed a system that extracts and structures causal information through knowledge triplets, creating a comprehensive graph with 688 unique entities. The methodology integrates knowledge graphs with Retrieval Augmented Generation (RAG) to enhance GPT models' causal reasoning capabilities. The resulting system was validated through human evaluation and cross-validation with Claude 3, demonstrating improved ability to provide contextually relevant and accurate responses to COVID-19 queries. This approach offers a novel framework for analyzing complex epidemiological data, enabling more structured and reliable insights for researchers.

Keywords: Knowledge-graphs, COVID-19, Retrieval-augmented generation, Large language models, Causal analysis.

1 Introduction

The COVID-19 pandemic has caused significant global economic and social disruption, deeply affecting public health, economies, and daily life. One of the key challenges in understanding and managing such a complex phenomenon lies in deciphering the causal relationships behind the virus's spread and impact. Traditional epidemiological models, although valuable, often struggle to capture the intricate, multi-layered interactions between biological, social, and environmental factors, not to mention the large amount of time spent in the construction of these models. Recent advances in large language models (LLMs) provide potential tools for analyzing vast amounts of data, but their lack in causal reasoning abilities and often miss deeper, structured insights into

[*]Tutor

complex cause-and-effect relationships.

To address this limitation, knowledge graphs present a powerful solution by structuring information into interconnected entities and relationships, enabling a more structured representation of causal interactions. Building a knowledge graph (G = (V, E)), where V represents entities and E denotes relationships, offers a pathway to organizing COVID-19 data into a format conducive to causal inference. In this context, using tools like KnowledgeGraphIndex from llamaindex allows the extraction of knowledge triplets from unstructured COVID-19 texts. This process efficiently organizes data into a graph structure, supporting queries that expose potential causal patterns across various factors involved in the pandemic.

Integrating knowledge graphs with retrieval-augmented generation (RAG) systems, such as GPT models, enhances interpretability and scalability by transforming complex data into coherent, human-readable insights. Embedding graph structures into vector space and leveraging transformer-based models for natural language generation allows for meaningful synthesis of information, making critical insights accessible to researchers and policymakers. This project leverages COVID-19 research to extract causal data that can inform and enrich the contextual knowledge available to LLMs, aiming to improve the precision and utility of AI-driven insights in pandemic response efforts.

2 Problem definition

As we mentioned, COVID-19 has presented a uniquely complex problem, where understanding the causal mechanisms underlying its spread, impacts, and mitigative strategies is crucial for effective public health responses. Recently, large language models like GPT have gained popularity for analyzing vast amounts of data due to their versatility and accessibility to non-specialized users. These models, by design, provide a flexible framework for answering almost any question, making them appealing even to those without technical expertise. However, while GPT models can generate insightful narratives, they lack the inherent accuracy needed for causal analysis, often overlooking essential causal relationships and producing responses based solely on correlation. This presents a significant limitation in high-stakes contexts like COVID-19 research, where accurate causal inference is vital.

To address this challenge, knowledge graphs offer a powerful solution. Using KnowledgeGraphIndex from llamaindex, knowledge triplets are extracted from unstructured COVID-19-related text (subject, predicate, object) and organized into a graph structure G=(V,E), where vertices V represent entities, and edges E represent their relationships. This graph can then be queried, providing a structured, interpretable foundation that enhances the reliability of data analysis and the accuracy of causal inferences, particularly when integrated with retrieval-augmented generation (RAG) techniques to inform LLM outputs.

3 State of the art

Knowledge graphs (KG) are a powerful tool widely used in the industry, with applications dating back to at least the 70s, Sowa (1976). The modern usage of the term originates from the 2012 announcement of the Google Knowledge Graph (GNG) Singhal (2012). This was followed by subsequent implementations of knowledge graphs by companies such as Airbnb, Amazon, eBay, Facebook, IBM, LinkedIn, Microsoft, Uber, and others. Nowadays, implementations such as the GNG improve search results by providing contextual information. Amazon uses them for refined product recommendations and Alexa's functionality. Facebook enhances its Graph Search for complex user queries, while IBM's Watson uses them for advanced data analysis. These applications demonstrate the versatility and power of knowledge graphs in driving innovation and improving various services across industries.

With the rise of Artificial Intelligence (AI), knowledge graphs have unlocked new possibilities for enhancing the development of various tools by integrating, managing, and extracting value from diverse, large-scale data sources. For instance, Paulheim (2017) demonstrated the use of these technologies integrated with AI, significantly improving the performance of AI systems by providing a structured and annotated representation of knowledge. This integration allows AI systems to better understand and interpret data, leading to more informed and accurate decision-making. The structured nature of knowledge graphs enables AI models to leverage context and relationships between data points, which is crucial for tasks such as natural language processing, recommendation systems, and predictive analytics Ehrlinger & Wöß (2016). As a result, the combination of AI and knowledge graphs is driving advancements in various domains, from healthcare and finance to e-commerce and social networking, by enabling more sophisticated and reliable insights.

The integration of RAG systems with knowledge graphs has further advanced the field, enabling large language models like GPT to leverage structured knowledge effectively. RAG systems combine retrieval mechanisms with generative models to enhance accuracy, relevance, and interpretability in language generation. Hussien et al. (2024) proposed an explainable road users behavior prediction system that integrates the reasoning abilities of Knowledge Graphs and the expressiveness capabilities of LLM. This approach has demonstrated significant improvements in LLM performance by anchoring responses in external knowledge, helping models produce more grounded and contextually accurate answers Hussien *et al.* (2024).

Moreover, recent research has explored the application of knowledge graphs in the medical field, particularly to improve the understanding and treatment of diseases. For example, knowledge graphs have been used to integrate and analyze biomedical data, leading to the discovery of new drug targets and the repurposing of existing drugs Rajabi & Kafaie (2022). This approach has proven especially useful during the COVID-19 pandemic, where rapid and comprehensive analysis of biomedical literature is crucial. By organizing large amounts of data in a structured format, knowledge graphs facilitate the identification of potential therapeutic agents and the understanding of disease mechanisms, thereby accelerating the research process and supporting public health efforts. The continued development and application of knowledge graphs in various fields highlight their potential to transform data management and analysis, paving the way for more efficient and effective solutions to complex problems.

4 Methodology

The first step in developing the proposed architecture was identifying an appropriate dataset to construct the knowledge graph, which would later be integrated into the RAG framework for the GPT model. We selected the CORD-19 dataset, a comprehensive resource created in response to the COVID-19 pandemic. CORD-19 contains over one million scholarly articles, including more than 400,000 with full text, covering COVID-19, SARS-CoV-2, and related coronaviruses. This open-access dataset is provided to the global research community, enabling the application of recent natural language processing (NLP) and AI techniques to uncover new insights and support ongoing efforts to combat this public health crisis. Given the rapid growth of coronavirus-related literature, there is an increasing need for AI-powered methods to help researchers keep pace with new findings. Due to computational constraints in this project, we utilized only the abstracts from the dataset on a local machine Wang *et al.* (2020).

We utilized the knowledge graph query engine from LlamaIndex to efficiently extract knowledge triplets—composed of a subject, predicate, and object—from the CORD-19 dataset. These triplets form the primal relationships and entities within the knowledge graph, encapsulating the structured information needed for downstream applications. The query engine streamlines the complex process of identifying and extracting meaningful data from large text corpora, automating the construction of a knowledge graph that represents relationships between entities and key concepts in the dataset.

To ensure that our knowledge graph captured relevant and accurate information, we designed a validation schema to define and filter the types of entities and relationships needed. This schema enabled us to specify both the permissible entities (such as "DISEASE" or "PERSON") and the allowable relationships between them (e.g., "CAUSES" or "TRIGGERS"). This approach allowed us to focus on interactions critical to understanding COVID-19 and its impacts.

The following table provides an overview of the main entity types and their potential relationships within the knowledge graph:

Entity Type	**Possible Relations**
PERSON	CAUSES, INDUCES, AFFECTS, IMPACTS, TRIGGERS, IMPROVES
PLACE	CAUSES, AFFECTS, IMPACTS, TRIGGERS, IMPROVES
ORGANIZATION	CAUSES, INDUCES, AFFECTS, IMPACTS, TRIGGERS, IMPROVES
INDUSTRY	CAUSES, INDUCES, AFFECTS, IMPACTS, TRIGGERS, IMPROVES
DISEASE	CAUSES, INDUCES, AFFECTS, IMPACTS, TRIGGERS, IMPROVES

As shown in the table, each entity type (e.g., "DISEASE") can be connected through various relationships that reflect possible causal or impact pathways, such as "CAUSES" or "AFFECTS." These relationships help us map how different factors, such as organizations or diseases, interact within the COVID-19 context, offering insights into causal patterns or chains of influence within the pandemic data.

In practice, the validation schema allowed us to refine the initial knowledge graph generated by the query engine, reducing noise and ensuring a coherent, relevant structure. This selective process was especially valuable given the large dataset size and the need for a concise, interpretable graph that could support causal analysis in our RAG framework.

The next steps in our methodology involved structuring the extracted triplets into a graph data structure capable of supporting further analysis. Once the knowledge graph was built, we applied an algorithm to identify the six closest neighboring paths within the graph, allowing us to detect indirect or non-explicit relationships between entities and concepts across different research papers. These paths reveal connections that might not be immediately apparent, enriching our understanding of how topics interrelate within the COVID-19 literature.

Using these identified paths, we generated descriptive phrases capturing the essence of each indirect relationship. These phrases were then embedded and stored in a vector database. This tool is critical here, as it allows us to efficiently store, retrieve, and query high-dimensional embeddings—numerical representations of the relational information—based on semantic similarity. By using Chroma, we can swiftly retrieve relevant data based on context, enhancing the precision and responsiveness of the RAG framework when queried. This setup enables the GPT-4 model to pull in pertinent context from our knowledge graph, allowing it to generate responses grounded in the data and providing users with coherent, contextually enriched answers that leverage both explicit and inferred relationships in the literature.

To evaluate the effectiveness of the RAG system, we implemented a two-pronged testing approach: human validation and cross-verification with a larger language model. For human validation, we reviewed the answers generated by the RAG-augmented GPT-4 model, assessing whether the responses were accurate, relevant, and grounded in the relationships defined in our knowledge graph. This step was essential to ensure that the system reliably answered queries with correct and contextually appropriate information.

Additionally, we employed cross-validation with a larger language model. By inputting the generated responses into this larger model, we evaluated whether it recognized the answers as plausible and correct, simulating a form of model-based verification. This dual approach of human and model validation provided robust testing for the RAG system, allowing us to iteratively refine its accuracy and relevance in response to complex COVID-19-related queries.

5 Results

5.1 Knowledge graph creation

In the results section, we present different visual representations of the knowledge graph, illustrating its structural features and key relationships. The first view focuses on direct relationships between core entities identified in the COVID-19 literature, such as “DISEASE,” “PERSON,” and “ORGANIZATION.” This view highlights primary causal and impact relationships, showing straightforward connections, like how COVID-19 (DISEASE) impacts healthcare organizations or affects various populations. This visualization provides an intuitive overview of the foundational structure and immediate interactions within the knowledge graph, allowing a clear, high-level understanding of the data.

Figure 1: Direct relationships

The second view zooms in on the indirect, or non-explicit, connections revealed by the six closest neighboring paths between entities, which our methodology identified using path-finding algorithms. These pathways expose secondary relationships, such as how a PLACE, like a hospital, might indirectly influence disease transmission through associated factors, revealing otherwise hidden connections. These indirect pathways add depth to the analysis, providing insights into the layered relationships between entities in different articles and capturing insights that would otherwise go unrecognized.

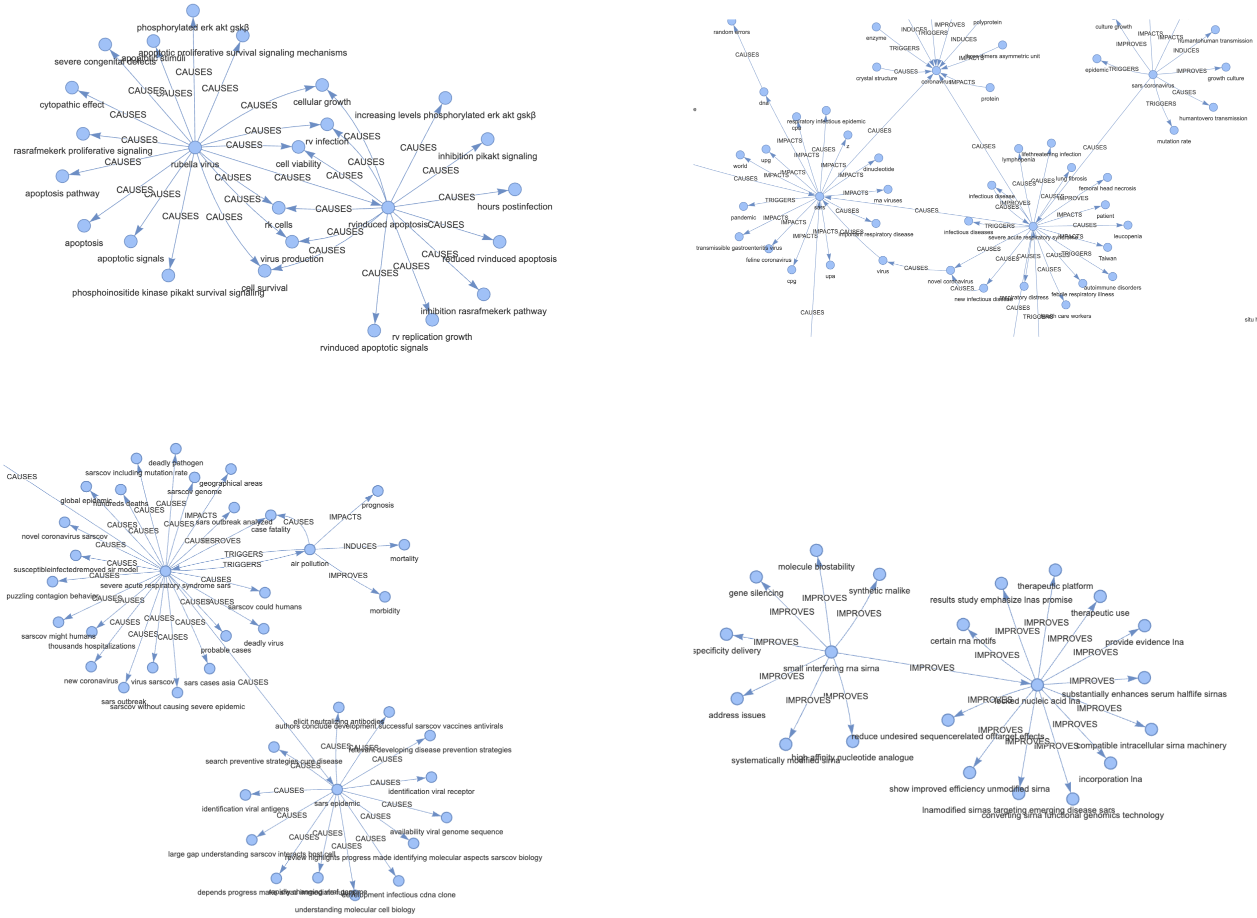

Figure 2: Indirect relationships

A total of 688 unique entities were extracted from the CORD-19 dataset, forming the backbone of the knowledge graph. These entities span various types, including “DISEASE,” “PERSON,” “PLACE,” “ORGANIZATION,” and “INDUSTRY,” capturing a broad range of topics and participants within COVID-19 research. The inclusion of 688 entities enabled a robust and detailed graph, capable of revealing both direct and indirect relationships within the data.

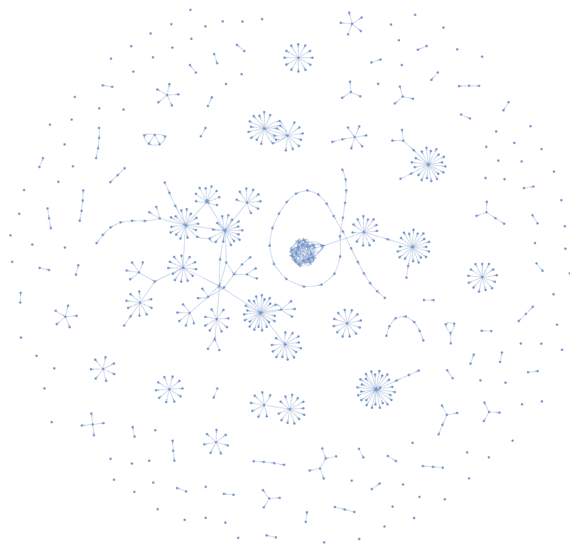

Figure 3: Full view of the knowledge graph

5.2 Implementation of RAG architecture

After constructing the knowledge graph and identifying each entity's closest six neighbors, we generated descriptive phrases capturing all possible relationships between these entities. These phrases encapsulate not only direct but also inferred relationships, drawing on patterns that may connect different aspects of the COVID-19 research across various entities and articles. A total of 118.916 phrases were created.

Each phrase was embedded into high-dimensional vectors and stored in Chroma, our chosen vector database, which plays a crucial role in the RAG architecture. By embedding and storing these relationship-rich phrases, Chroma allows for efficient similarity-based retrieval, enabling the RAG-augmented GPT-4 to access contextually relevant information dynamically.

To evaluate the accuracy and reliability of our RAG system as we mentioned before, we conducted two types of testing: human validation and automated cross-validation using a larger language model, Claude 3, which has been noted for its high token capacity and contextual understanding. This two-pronged approach allowed us to assess the RAG system's ability to provide correct, contextually relevant, and informative answers based on the constructed knowledge graph.

5.3 Validation

For this evaluation, a total of 5 questions were asked to the GPT model within the RAG framework. These questions addressed causal aspects of COVID-19, allowing for a robust assessment of the RAG system's performance in providing accurate, relevant, and well-structured responses.

5.3.1 Questions Asked

1. How do environmental factors influence the spread of COVID-19?

2. What are the consequences of pneumonia?
3. How does COVID-19 affect the cardiovascular system?
4. What is the role of ACE2 in COVID-19 infection?
5. Do air pollution levels causally influence COVID-19 mortality rates?

Human Validation

The human validation process was conducted by manually evaluating the responses provided by the RAG-augmented GPT model. I reviewed each response by verifying the information within our knowledge graph and conducting additional research online to confirm answer accuracy.

Cross-Validation with Claude 3

In addition to human validation, we implemented an automated cross-validation approach using a larger language model, Claude 3, to simulate expert review using prompt engineering.

The following table presents a summary of the answers generated by the RAG-augmented GPT-4 model, along with their validation through Claude 3. The human validation was successfully conducted for all the questions, ensuring that the responses were accurate and relevant. Each answer was cross-checked for correctness and alignment with established knowledge, with Claude 3 providing an additional layer of verification.

Used Prompt:

Instruction: You are given a response generated by a RAG-augmented GPT-4 model based on a query. Your task is to evaluate the accuracy, consistency, and relevance of the response. Analyze if the response logically incorporates known relationships and entities typically associated with the question topic.

> *Example Format:*
>
> **Question:** [Insert query posed to the RAG model]
>
> **GPT-4 RAG Response:** [Insert RAG model response]
>
> **Evaluation Questions:** Is the response accurate and appropriate for the question? (Yes/No) Does the response logically and clearly connect relevant entities or relationships? (Yes/No) What modifications, if any, would make the response more accurate or informative?

Question	GPT-4 RAG Response	Claude 3 Validation
How do environmental factors influence the spread of COVID-19?	Environmental factors play a significant role in the spread of COVID-19. Here are a few ways: • Temperature and Humidity: Some studies suggest that COVID-19 may spread less efficiently in higher temperature and humidity, although more research is needed. • Indoor vs Outdoor Spaces: The virus is more likely to spread in indoor spaces, especially those with poor ventilation. Outdoor spaces typically have better ventilation and allow for physical distancing, reducing the risk of transmission. • Surfaces: The coronavirus can survive on surfaces for hours to days, depending on the material. The risk of transmission from touching contaminated surfaces and then touching the face is lower than direct person-to-person contact, but it's still a risk. • Population Density: Areas with higher population density can have more rapid spread due to increased person-to-person contact. • Air Pollution: Some studies suggest that high levels of air pollution may exacerbate the severity of COVID-19 symptoms and increase mortality rates.	**Is the response accurate and appropriate? Yes**, the response is generally accurate and addresses how environmental factors influence COVID-19 transmission. It covers multiple relevant aspects and presents information that aligns with established scientific understanding. **Does the response logically connect entities? Yes**, the response makes clear logical connections between environmental factors and their impact on transmission. **Suggested modifications:** • Temperature and Humidity could be more specific. • Indoor vs Outdoor Spaces could include specific ventilation metrics. • Surfaces section could be improved with survival times on different surfaces. • Additional factors like UV radiation and wind patterns could be added.
What are the consequences of pneumonia?	Pneumonia can result in serious complications including: • Difficulty in breathing, lung abscess, sepsis, pleural effusion, ARDS.	**Is the response accurate and appropriate? Yes**, the response is accurate and aligns with medical knowledge. Each complication is correctly described. **Does the response logically connect entities? Yes**, causal relationships are well-established between pneumonia and its complications. **Suggested modifications:** • Add frequency/risk information. • Enhance clinical context with early warning signs. • Include treatment and prevention measures. • Add long-term effects and economic impact.
How does COVID-19 affect the cardiovascular system?	COVID-19 can cause damage to the heart muscle, lead to heart failure, blood clotting disorders, and inflammation that affects blood vessels.	**Is the response accurate and appropriate? Yes**, but incomplete. The response lacks detail and specificity. **Does the response logically connect entities? Partially**, the response is basic but does not explain the mechanisms linking COVID-19 to cardiovascular complications. **Suggested modifications:** • Add specific mechanisms of cardiac injury. • Expand on blood clotting complications and cardiovascular manifestations.
What is the role of ACE2 in COVID-19 infection?	ACE2 is the main entry point for the SARS-CoV-2 virus, allowing it to enter cells. It also has protective effects against lung injury.	**Is the response accurate and appropriate? Yes**, the response is accurate and explains ACE2's role in infection. **Does the response logically connect entities? Yes**, it clearly links ACE2, viral entry, and disease mechanisms. **Suggested modifications:** • Expand on molecular mechanisms and tissue distribution of ACE2.
Do air pollution levels causally influence COVID-19 mortality rates?	Studies suggest a potential association between air pollution and COVID-19 mortality, but causality has not been established.	**Is the response accurate and appropriate? Partially**, the response emphasizes caution but lacks specific evidence. **Does the response logically connect entities? Partially**, it acknowledges confounding factors but lacks biological mechanism explanations. **Suggested modifications:** • Include specific evidence from epidemiological studies. • Explain biological mechanisms supporting causality. • Discuss public health policy and preventive strategies.

Table 2: RAG-augmented GPT-4 Responses and Claude 3 Validation

To see the full responses explore the following document .

6 Conclusions and future research

The project underscores the power of knowledge graphs in enhancing large language models (LLMs), showcasing the significant scope of RAG (Retrieval-Augmented Generation) architectures within the industry and their potential applications. Manual verification of outputs provided deeper insights into the model's performance, enabling rapid identification of accuracy patterns and adjustment opportunities to boost relevance and alignment with expected responses.

Claude 3's validation assessments proved invaluable, confirming that responses were both contextually accurate and factually sound, while also highlighting areas for further refinement. This dual testing approach strengthened the validation process, affirming the RAG system's capacity for producing reliable and contextually rich responses based on the knowledge graph's structure and interrelations.

These results affirm the model's effectiveness and suggest an avenue for future research—enhancing the knowledge graph by incorporating information from Wikipedia. This augmentation could expand the graph's knowledge base, providing even deeper context and supporting the generation of more nuanced and informed responses.

Acknowledgements

Thanks to EAFIT Univeristy to provide spaces for students to develop their research and writing habilities.

References

Ehrlinger, Lisa, & Wöß, Wolfram. 2016. Towards a definition of knowledge graphs.

Hussien, Mohamed Manzour, Melo, Angie Nataly, Ballardini, Augusto Luis, Maldonado, Carlota Salinas, Izquierdo, Rubén, & Sotelo, Miguel Ángel. 2024. RAG-based Explainable Prediction of Road Users Behaviors for Automated Driving using Knowledge Graphs and Large Language Models. *arXiv preprint arXiv:2405.00449*.

Paulheim, Heiko. 2017. Knowledge graph refinement: A survey of approaches and evaluation methods.

Rajabi, Enayat, & Kafaie, Somayeh. 2022. Knowledge graphs and explainable ai in healthcare.

Singhal, Amit. 2012. Introducing the Knowledge Graph: things, not strings. *Official Google Blog*.

Sowa, John F. 1976. *Conceptual graphs for a data base interface*.

Wang, Lucy Lu, Lo, Kyle, Chandrasekhar, Yoganand, Reas, Russell, Yang, Jiangjiang, Burdick, Douglas, Eide, Darrin, Funk, Kathryn, Katsis, Yannis, Kinney, Rodney, *et al.* 2020. Cord-19: The covid-19 open research dataset. *ArXiv*.