Causal Graph Identification by LLMs

Because of the issue of artificial intelligence (AI) systems relying on simple associations and the need for trustworthy machine learning (ML) tools, the scientific community has increasingly focused on studying causality as a means to address the limitations of AI and ML. Causality is seen as a potential solution to overcome these limitations. The current machine learning systems primarily rely on statistical approaches and lack the ability to reason about interventions and retrospection. This limitation hinders their potential to achieve strong AI. To attain human-level intelligence, learning machines require a model of reality, similar to those utilized in causal inference, to provide guidance and enhance their capabilities. [14]

1. Introduction

Advances in causal inference is vital for many fields and contexts, such as the medical one, where most of the asked research questions are not associational, but causal in nature; with these questions, researchers try to uncover the cause-and-effect relationships between variables (e.g., treatments, interventions, outcomes). These questions cannot be answered from observed data alone and could require specific and expert domain knowledge.

Although expert opinion remains one of if not the best tool for causal analysis (e.g., causal discovery for building causal graphs), it can be very time and resource consuming, since the amount of research data becomes larger and larger (reaching dimensions that limit the possibility of parsing through the enormity of evidence for building DAGs), and experts may eventually commit errors or miss important graph details.   
These difficulties could be partially solved by using LLMs, which have been trained on immense amounts of textual data [5].

1. Project Objective

**The primary goal of this project is to conduct an empirical study to assess the possibility of performing causal analysis using Large Language Models (LLMs).**

The project focuses on the operation of causal discovery, which is the task of learning the structure of causal relationships between variables and entities; its output is a directed graph that represents the underlying data-generation process (DGPO) and provides insight into the true causal relationships between variables. The generated graph is used as a base for many other (if not all) fundamental tasks in causal analysis (e.g. effect inference, prediction, attribution) [1].

The causal discovery task is performed starting from natural language, that is textual data, such as scientific papers and research publications.

The data is then processed to extract the main textual entities; a discovery procedure is then used to find the causal relationship between these entities.

The final operation creates the causal graph using the causal relationships found in the previous step and plots the directed acyclic graph (DAG).

State of the art

The main objective of causality discovery is understanding the underlying cause-and-effect relationships in the system of interest. The current state-of-the-art techniques for causal discovery can be divided into two main classes: constraint-based methods and score-based methods.

Constraint-based methods use the patterns of conditional independence among variables to deduce or determine the underlying causal relationships within a system, examining how the presence or absence of one variable affects the likelihood of another variable.

Score-based methods use scoring functions to assess the quality of causal structures and then search for the optimal structure that maximizes the assigned score.

Both these methods include different algorithms and techniques. While these techniques have been shown to be effective across multiple applications and scenarios, they are not without challenges. The complexities introduced by noise within data or unobserved confounders (i.e., variables correlated with both cause and effect) can obscure causal relationships. Additionally, the limited availability of data further challenges the task of causal discovery.

When dealing only with real-world observational data, it is generally not possible to perform causal discovery and find the exact causal graph. The reason relies on the Markov equivalence class property, where multiple graphs structures are equally likely to be found, given the same data distribution [1].

**Markov equivalence**: two (or more) graphs are said to be Markov equivalent (or part of the same Markov equivalence class) if they share the same skeleton (graph structure with undirected edges) and same immoralities (same collider patterns in the graph) [13].

Different approaches and effort have tried to overcome this limitation, yet the identification of causal graphs continues to pose challenges, especially when working with real-world observational data, revealing a concerning assessment of their effectiveness [1].

LLMs offer a fresh perspective on causal discovery by focusing on the *metadata* associated with variables, rather than their data values. This metadata-based reasoning is typically done by human domain experts when constructing causal graphs, who use their general or specialized domain knowledge and common sense.

# add comment that expert approach is time consuming and error prone

LLMs could infer the causal graph structure by “reasoning” on variables’ metadata and the problem context, both expressed in natural language: the LLMs would use their training knowledge (and additional input data) to determine the edges of the graph, in contrast to the Score-based causal discovery.

1. Implementation

The project can be divided in two main steps: data collection and data analysis. The former one consisted in collecting the necessary data for the latter, which can itself be divided into multiple other sub-operations.

First, the data collection process will be presented, highlighting the utilization of the National Center for Biotechnology Information (NCBI) API for requesting the necessary textual data. Subsequently, the operations of data processing and causal analysis will be discussed.

3.1 PubMed scraping

The first step of the project consisted in collecting the necessary textual data for testing the causal discovery capabilities of the GPT LLM.

The used data is taken from the PubMed database, a free search engine accessing primarily the MEDLINE database of references and abstracts on life sciences and biomedical topics [Pubmed - wikipedia].

Initially, only abstracts (and extra details) were extracted from the PubMed database as a preliminary effort.

**Scraping pipeline**

A pipeline handling the essential operations was created for extracting the necessary textual data from the PubMed database. To automate this extraction process, a python script was written using the public API provided by the NCBI as stable interface into its query and database system.

The pipeline allows the user to extract textual data from PubMed by searching for specific terms.

Immagine che contiene testo, schermata, Carattere, design

Descrizione generata automaticamente

Figure 1 - PubMed data extraction flow

The pipeline's main operations are handled by the *search\_by\_terms*, *get\_articles\_data*, and *clean\_data* procedures.

***search\_by\_terms***

The *search\_by\_terms* procedure is the first operation of the pipeline. As the name suggests, it allows the user to search for articles in the PubMed database containing the specified search terms. The search terms are joined as query parameters in the request URL. An API\_KEY is also sent in the request URL, to allow up to 10 requests per second and to ensure smooth and supported access to the desired resources.

def search\_by\_terms(terms, db, retmax, use\_history):

    terms\_string = '+AND+'.join([s.strip().replace(' ', '+') for s in terms])

    url = f'{base\_url}esearch.fcgi?db={db}&term={terms\_string}&retmax={retmax}&api\_key={api\_key}'

    if use\_history:

        url += '&usehistory=y'

    response = requests.get(url)

    ids = re.findall(r"<Id>(\d+)</Id>", response.text)

    if use\_history:

        web\_match = re.search(r"<WebEnv>(\S+)</WebEnv>", response.text)

        web = web\_match.group(1) if web\_match else None

        key\_match = re.search(r"<QueryKey>(\d+)</QueryKey>", response.text)

        key = key\_match.group(1) if key\_match else None

        return ids, web, key

    return ids

The response is in a xml format, and it is processed to extract all ID numbers of the articles found in the specified NCBI database, which in this case is PubMed.

This simplified version of the first pipeline operation only returns the extracted IDs. However, the script allows users to utilize the NCBI *Entrez History* feature, which proves to be significantly more efficient when dealing with tasks that involve searching for or downloading a substantial number of records. This approach helps streamline the process and optimize the retrieval of records in a more efficient manner, making it possible to upload many IDs or download several hundred records at once.

***get\_articles\_data***

The *get\_articles\_data* procedure is the second step of the data acquisition pipeline. It queries the NCBI for the actual content of the articles with the specified ID.

def get\_articles\_data(ids=[], web\_env='', query\_key='', db='pubmed', retmax=1000):

    url = f'{base\_url}efetch.fcgi?db={db}'

    if use\_web\_env:

        url += f'&query\_key={query\_key}&WebEnv={web\_env}'

    else:

        ids\_string = [str(id) for id in ids]

        url += '&id=' + ','.join(ids\_string)

    url += f'&rettype=abstract&retmode=xml&api\_key={api\_key}&retmax={retmax}'

    response = requests.get(url)

    soup = BeautifulSoup(response.text, features="xml")

    articles = soup.find\_all('PubmedArticle')

    if not articles:

        print('ERROR: No articles found')

        return None

    data = pd.DataFrame(columns=['id', 'title', 'abstract', 'keywords', 'pub\_date'])

    for article in articles:

        id = article.find('PMID').get\_text()

        date = article.find('PubMedPubDate', {'PubStatus': 'received'})

        pub\_date = datetime.strptime(f'{date.find("Day").get\_text()} {date.find("Month").get\_text()} {date.find("Year").get\_text()}', "%d %m %Y")

        title = article.find('ArticleTitle').get\_text()

        abstract = ''.join([a.get\_text() for a in article.find\_all('AbstractText')])

        keywords = [k.get\_text() for k in article.find\_all('Keyword')]

        data = pd.concat([data, pd.DataFrame({'id': id, 'title': title, 'abstract': abstract, 'keywords': [keywords], 'pub\_date': pub\_date})]).reset\_index(drop=True)

    return data

Just like before, the shown function is a simplified version of the second step of the pipeline: the NCBI API allows users to query article data with the article ID or by using the *Entrez History* feature, which can provide a more efficient data retrieval. A URL parameter of the request defines the main data content requested, which, in this case, are the abstracts (rettype=abstract).

The returned data is in a xml format, and it is processed and parsed to extract the necessary information. The recovered data include the abstract of the article and some additional information about the article itself. The additional information are the article ID number, the title, the keywords, and the publication date.

***clean\_data***

The *clean\_data* procedure is the third and last step of the data acquisition pipeline. It performs cleaning operations on the obtained data, e.g., by removing null abstract values, duplicates, and eventually removing data of articles published in a particular date range.

def clean\_data(data, drop\_id\_duplicates, drop\_empty\_abstracts, drop\_nan\_abstracts, drop\_date\_nan, drop\_date\_before, drop\_date\_after, search\_terms):

    if data is None or data.empty:

        print('ERROR: No data provided')

        return None

    if drop\_id\_duplicates:

        data = data.drop\_duplicates(subset=['id']).reset\_index(drop=True)

    if drop\_empty\_abstracts:

        data = data.loc[data['abstract'] != ''].reset\_index(drop=True)

    if drop\_nan\_abstracts:

        data = data.dropna(subset=['abstract']).reset\_index(drop=True)

    if drop\_date\_nan:

        data = data.dropna(subset=['pub\_date']).reset\_index(drop=True)

    if drop\_date\_before:

        data = data.loc[data['pub\_date'] > drop\_date\_before].reset\_index(drop=True)

    if drop\_date\_after:

        data = data.loc[data['pub\_date'] < drop\_date\_after].reset\_index(drop=True)

    if search\_terms:

        data['search\_terms'] = [search\_terms]\*len(data)

    return data

3.2 Causal Analysis

After completing the preliminary phase of data collection, the main focus of the project shifted towards the actual analysis operations. The primary objective was to investigate the causal capabilities of LLMs, specifically focusing on causal discovery.

This causal analysis operation consists of learning the causal graphs from a given dataset and context, by uncovering the cause-and-effect relationships and dependencies between the variables and entities of the system of interest: this is done by answering questions such as "Which variables directly affect each other?" or "What is the causal directionality between variables?".

To introduce the causal analysis operation, the next section provides a set of definitions for the main concepts and assumptions on graph theory and causality.

**Definitions** [6]

Definition 1 (Graph). A graph G = (V, E) is a mathematical object represented by a tuple of two sets: a finite set of vertices V and a finite set of edges E ⊆ V × V.

Definition 2 (Directed Graph). A directed graph (DG) G is a graph where the edge (X, Y) is distinct from the edge (Y, X).

Definition 3 (Path). A path π = (X − · · · − Y) is a tuple of non-repeating vertices, where each vertex is connected to the next in the sequence with an edge.

Definition 4 (Directed Path). A directed path π = (X → · · · → Y) is a tuple of non-repeating vertices, where each vertex is connected to the next in the sequence with a directed edge.

Definition 5 (Cycle). A cycle is a path that starts and ends at the same vertex.

Definition 6 (Directed Acyclic Graph). A directed acyclic graph (DAG) is a directed graph G that has no cycles.

Definition 7 (Causal Graph). A causal graph G is a graphical description of a system in terms of cause-effect relationships, i.e., the causal mechanism. Causal graphs, usually in the form of Directed Acyclic Graphs (DAGs), encode contextual knowledge of variables (both observable and unobservable) and their causal dependency.

As mentioned, DAGs are graphical representations of causal relationships among variables, where the arrows indicate the direction of causality.

Definition 8 (Direct and Indirect Cause). For each directed edge (X, Y) ∈ E, X is a direct cause of Y and Y is a direct effect of X. Recursively, every cause of X that is not a direct cause of Y, is an indirect cause of Y.

In a causal graph, the nodes represent the context entities and variables (e.g., in a medical context they would be symptoms, illnesses, diseases, treatments, medications, outcomes, etc. …) while the edges represent the causal relationship between said entities (e.g., a medical treatment can cause a particular outcome or side effect).

The graph encodes the causal relationships between entities and variables. These relationships are represented by different types of edges: directed edges indicate direct causes, bi-directed edges represent entities that are causally correlated with the two variables having an unobserved or latent common cause, and non-existent edges indicate the absence of a causal relationship between the variables.

The acyclic property of DAGs is crucial for ensuring their interpretability and for preserving the causal relationships they represent. This property is fundamental for many reasons: it ensures logical consistency, temporal ordering, identifiability of causal effects, facilitates counterfactual reasoning, and aids in prediction and intervention tasks.

**Causal Analysis Pipeline**

This step of the project consisted in working with the collected data from the PubMed database to extract information from the abstracts. The necessary information consisted of the main named entities of the textual data, which were then used to perform the actual causal analysis.

This second part of the project was implemented in a python script, in the form of a single pipeline of multiple sub-steps, called *causal\_discovery\_pipeline*: these operations include extracting entities from the textual data, performing the actual causal analysis on the found entities, and ultimately generate the resulting causal graph.

Immagine che contiene testo, schermata, Carattere, design

Descrizione generata automaticamente

Figure 2 - Causal analysis flow

Part of the sub-steps of this second part of the project, which involves causal analysis, relies on the GPT API.

**GPT API** [2]

The GPT API is a tool that provides access to OpenAI's GPT models, allowing the integration of natural language processing capabilities into applications. It works by sending requests to the API endpoint with a given prompt, and in return, it generates quality, context-aware text based on the provided input.

The GPT API is used by sending HTTP requests to the API endpoint, specifying the model to use, and providing the necessary parameters such as the prompt and optional additional messages to contextualize the use and behavior of the model, i.e., how the model should answer to requests. The API then processes the request and returns the generated text as a response.

The GPT API can be employed in a variety of applications and use cases. It can be used to generate conversational agents, draft emails or other pieces of writing, provide language translation, answer questions, or assist with content creation.

**Using the GPT API**

The GTP API is used by specifying the model to use (e.g., gpt-4) and additional messages, such as the *system* message and the *user* message. These act as instructions to the model, with the former being a system level instruction to guide the model's behavior throughout the conversation (e.g., asking the model to answer or to act in a specific way), and the latter functioning as the actual request the model is required to answer.

In particular, the system message is used to contextualize the model and its behavior, to make it more useful and accurate for the required operation: for the project’s causal analysis tasks, for example, the prompt was prepended with the message “You are a helpful assistant for causal reasoning”, to try steering the output space to more causally consistent answers. This was shown being an effective prompt-engineering technique that result in more accurate answers [1].

The system message helps set the behavior of the assistant, by modifying the personality of the assistant or providing specific instructions about how it should behave throughout the conversation. However, the system message is optional and the model’s behavior without a system message is likely to be like using a generic message such as "You are a helpful assistant." [2].

messages=[{"role": "system", "content": system\_msg},

          {"role": "user", "content": user\_msg}])

Another parameter that can be optionally set is the *temperature*: this represents the degree of exploration or randomness of the model’s output. A higher value (e.g., 0.8) increases creativity and diversity but might be less focused. A lower value (e.g., 0.2) produces more deterministic output following patterns.

The following is a complete example of a GPT API chat completion request that uses the *gpt-4* model, and specifies system, assistant, and user messages.

openai.ChatCompletion.create(

  model="gpt-4",

  messages=[

        {"role": "system", "content": "You are a helpful assistant."},

        {"role": "user", "content": "Who won the world series in 2020?"},

        {"role": "assistant", "content": "The Los Angeles Dodgers won the World Series in 2020."},

        {"role": "user", "content": "Where was it played?"}

    ],

  temperature=0.3,

)

An example of response looks as follows:

{

  "choices": [

    {

      "finish\_reason": "stop",

      "index": 0,

      "message": {

        "content": "The 2020 World Series was played in Texas at Globe Life Field in Arlington.",

        "role": "assistant"

      }

    }

  ],

  "created": 1677664795,

  "id": "chatcmpl-7QyqpwdfhqwajicIEznoc6Q47XAyW",

  "model": "gpt-3.5-turbo-0613",

  "object": "chat.completion",

  "usage": {

    "completion\_tokens": 17,

    "prompt\_tokens": 57,

    "total\_tokens": 74

  }

}

In Python, the assistant’s reply can be extracted as follows:

response['choices'][0]['message']['content']

**GPT Prompt Engineering** [3][4]

Depending on the task, the GPT LLM can produce satisfactory results when asked to answer a question. However, there are some expedients that have shown to be beneficial and to increase the results accuracy when querying the LLM. These techniques are part of the discipline known as prompt-engineering, which comprises a set of rules and instructions that serve as guidelines to enhance the capabilities of LLMs on a wide range of common and complex tasks.

Prompt engineering has emerged as a powerful technique to enhance the performance and control the behavior of Language Models (LMs), particularly Large Language Models (LLMs), such as the used *gpt-3.5* and *gpt-4* models. Prompt engineering involves crafting system and user messages that guide the model's responses and shape its output to meet specific requirements.

The goal of prompt engineering is to provide contextual cues and instructions to the language model, enabling it to generate more accurate, relevant, and desired responses. By designing prompts, it is possible to tailor the behavior of LLMs, making them more suitable for various tasks, domains, and user needs.

Among the many prompt engineering techniques, one of the most important strategies is improving the clarity and precision of the prompt text by providing clear and specific instructions.   
Delimiters like brackets, tags or quotes can segregate sections within the prompt, aiding in a more organized interpretation of the input. Furthermore, prompting for structured output by specifying the desired response format guides the model in generating well-organized results.  
Checking task conditions also ensures the necessary assumptions are met. For instance, the prompt can verify whether essential information is available to complete the task and provide alternative instructions if this information is missing.

The application of few-shot prompting involves showing successful task examples to the model before requesting similar ones. This helps the model understand the context better, preparing it to deliver pertinent and accurate responses.

In line with another principle of prompt engineering a methodical approach is used to provide the model with enough time to “think”. This involves requesting the model to answer with a step-by-step explanation of its thought process before providing the final answer. Additionally, the model is directed to work out its own solution rather than rushing to conclusions. For instance, if the model needs to check a given solution's accuracy, it's prompted to come up with its solution and then compare it to the provided one.

**NER: Extracting Medical Entities from Text**

As previously mentioned, the second part of the project consisted in working with the collected data. The first step of the operation involved performing Named Entity Recognition on the abstracts, a fundamental procedure to extract and classify named entities. This step was essential for further processing and analysis.

Named Entity Recognition (NER) is a crucial natural language processing (NLP) task that aims to identify and classify named entities within text. In the context of medical texts, NER plays a vital role in extracting specific medical entities such as diseases, symptoms, treatments, drugs, anatomical terms, and medical procedures. Medical texts pose challenges for NER due to their specialized terminology, which often includes abbreviations and multiple names and entities used as synonyms. Additionally, the complex language structures found in medical texts, along with the diverse sources from which they originate, further complicate the NER process.

The NER operation was performed using the GPT LLM.

**NER gpt prompt messages**

To enhance the performance of the Language Model (LLM) for the NER task, both the system and user messages were designed accordingly.

The system message employed was "You are a helpful assistant for Named Entity Recognition of medical texts" to provide guidance to the model and improve its understanding of the task at hand.

To further aid the model's comprehension, the user message was crafted using the abstract of the medical text, complemented with additional information about the types of entities to be extracted. In this case, since the texts were focused on medical literature and research publications, the model was explicitly instructed to identify entities, with a particular emphasis on diseases, medications, treatments, symptoms, etc….

The intention of customizing the user message by providing relevant context and specific entity requirements, was to guide the LLM towards producing more accurate and relevant results for the ongoing NER operation.

def gpt\_ner(text):

    system\_msg = ' You are a helpful assistant for Named Entity Recognition of medical texts'

    user\_msg = f'Given the following text, please identify the named entities, especially diseases, medications, treatments, symptoms. <Text>{text}</Text>. Answer within the tags <Answer><Entity>...</Entity</Answer>.'

    response = gpt\_request(system\_msg, user\_msg)

    if not response:

        return []

    soup = BeautifulSoup(response, 'xml')

    entities = [entity.text for entity in soup.find\_all('Entity')]

    return entities

The result of the *gpt\_ner* function is an array containing all the found entities; it is then used for the subsequent causal analysis.

The *causal\_discovery\_pipeline* also provides the option to include an entity optimization step through the *optimize\_entities* procedure: by using the GPT api, the pipeline operation focuses on identifying synonyms, redundant entities, or entities and names that can be used interchangeably. In the generated output, entities with synonymous or similar meanings are matched together.

**Causal discovery**

With the NER operation completed and the entities extracted begins the main step of the pipeline, which is the causal discovery operation.

This step consists of the *gpt\_causal\_discovery* function, which processes the text passed as input to the main pipeline to perform causal discovery.

The approach for this operation uses a naïve discovery procedure, as it tries to infer the causal relationship across the different variables by querying the LLM on the direction of the pairwise causal relationships for each possible pair combination. The type of causal relationship between a pair of entities corresponds to the edge orientation of the causal graph.

**Causal edge direction**

In a causal Directed Acyclic Graph (DAG), the relationship between a pair of entities can be formalized as either a directed, a bidirected, or a non-existing edge.

The **directed edge** (*A 🡪 B*) denotes a direct causal dependence between the two features A and B, where A is a direct cause of B, without excluding the possible presence of a common cause of both A and B.

The **bi-directed edge** (*A <-> B* or both *A 🡪 B* and *A 🡨 B*) represents a causal relationship where A and B are causally correlated, and the two variables have an unobserved or latent common cause.

A **non-existent** edge denotes that no causal relationship exists between the two variables.

**Possible answers**

To infer the causal edge direction, the pipeline function queries the LLM on which cause-and-effect relationship is more likely given the two entities.

As previously stated, the *system message* used for this operation is *'You are a helpful assistant for causal reasoning and cause-and-effect relationship discovery'*, to try steering the output space to more causally consistent answers.   
On the other hand, the *user message* presents the current pair of entities of interest, asking a single question about the direction of the causal dependency, and requesting a step-by-step explanation in response. The possible answers the LLM is requested to choose from are also listed in the user message, and these are:

1. "X" causes "Y";
2. "Y" causes "X";
3. "X" and "Y" are not causally related;
4. there is a common factor that is the cause for both "X" and "Y"

To then enhance the accuracy and exploration of cause-and-effect relationships, the prompt uses random verbs of causation when querying the GPT LLM. This approach can be beneficial in terms of coverage of language patterns and potential causal relationships, can reduce the risk of bias that may come from consistently relying on a specific verb, and can encourage the model to explore different relationships between variables, allowing for a more comprehensive analysis of the data.

def gpt\_causal\_discovery(entities, text, use\_pretrained\_knowledge, reverse\_variable\_check):

    graph\_edges = []

    system\_msg = 'You are a helpful assistant for causal reasoning and cause-and-effect relationship discovery.'

    text\_msg = ''

    text\_msg += ''

    if text:

        text\_msg += f'the following medical text <Text>{text}</Text> '

        if use\_pretrained\_knowledge:

            text\_msg += 'and '

        else:

            text\_msg += ', '

    if use\_pretrained\_knowledge:

        text\_msg += 'your pre-trained knowledge, '

    for i1, e1 in enumerate(entities):

        for i2, e2 in enumerate(entities):

            if i1 == i2:

                continue

            if not reverse\_variable\_check and i1 >= i2:

                continue

            user\_msg = f'Given {text\_msg}the entities "{e1}" and "{e2}", Which cause-and-effect relationship is more likely? A. "{e1}" causes "{e2}"; B. "{e2}" causes "{e1}"; C: "{e1}" and "{e2}" are not causally related; D: there is a common factor that is the cause for both "{e1}" and "{e2}";  Lets work this out in a step by step way to be sure that we have the right answer. Then provide your final answer within the tags <Answer>[answer]</Answer>, (e.g. <Answer>C</Answer>).'

            response = gpt\_request(system\_msg, user\_msg)

            if response:

                graph\_edges.append(((e1, e2), response))

    return graph\_edges

In a default execution mode, the pipeline examines combinations (without repetition) of all pairs of entities identified. The total number of queries (one for each pair) is

in the case of ten entities, the total queries are .

**Double variables edge test**

The pipeline also allows for a double test for each variable pair, checking all potential variations without repetition (i.e., relationship “X” - “Y” and “Y” - “X”). As a result, the LLM is queried twice for each pair of entities, with a total number of variations of

With ten entities, the total queries are . In this case, the pipeline additionally verifies answer compatibility: the answer to the query about the causal relationship between "X" and "Y" must be consistent with the answer to the query about the relationship between "Y" and "X". This validation is managed by the *check\_invalid\_answers* function, which distinguishes between valid and invalid edge directions based on response consistency.

The edge direction and causal relationship between entities associated with "invalid" answers are then re-queried using the *correct\_invalid\_edges* function; this process involves querying the LLM with the inconsistent answers obtained earlier and asking again for the most likely relationship between the given variables. The newly acquired answers are then added to the previously identified "valid" edges.

The output of this step of the pipeline is an array containing the type of causal relationship between each pair of entities.

3.3 Plotting the causal graph

The next and final operations of the project involve plotting the resulting causal graph.

**Graph preprocessing**

Before plotting the graph, an intermediate operation of edge and node preprocessing is performed. This preprocessing step aids in the next ones by decoding the LLM answers and converting them into sets of nodes and normalized directed edges (i.e., in the form of *X → Y*).

The main procedure responsible for this operation is the *preprocess\_edges* function. It generates a set consisting of all nodes (previously extracted entities), an array containing the normalized directed edges, one containing all bidirected edges, and a dictionary representing the graph, where its keys are the nodes in the graph, and its values are list of the nodes that the given node points to: this dictionary encodes only directed edges that represent a direct causal relationship between entities.  
The normalization of edges is performed by the *normalize\_edge\_direction* function, which takes the nodes and the LLM's answer regarding their causal relationship as input, as well as the dictionary representing the graph. This function processes the output of the LLM to add the resulting node-to-node adjacency to the *graph* dictionary. It then returns the corresponding edge, in the form of X → Y, which represents the causal dependency between the nodes.

def preprocess\_edges(edges):

    graph = {}

    directed\_edges = []

    bidirected\_edges = []

    for (n1, n2), answer in edges:

        if n1 not in graph:

            graph[n1] = []

        if n2 not in graph:

            graph[n2] = []

        direction = normalize\_edge\_direction(n1, n2, answer, graph)

        if direction:

            if len(direction) == 2:

                bidirected\_edges.extend(direction)

            else:

                processed\_edges.extend(direction)

    nodes = list(graph.keys())

    return nodes, processed\_edges, bidirected\_edges, graph

def normalize\_edge\_direction(e1, e2, answer, graph):

    if answer in arrows:

        if arrows[answer] == forward\_arrow:

            graph[e1].append(e2)

            return [(e1, e2)]

        elif arrows[answer] == backward\_arrow:

            graph[e2].append(e1)

            return [(e2, e1)]

        elif arrows[answer] == bidirectional\_arrow:

            return [(e2, e1), (e1, e2)]

        else:

            return None

    else:

        return None

An additional step is taken before plotting the graph, where a check is performed to determine whether the resulting graph is acyclic: this is done using the *graph* dictionary previously constructed.

**Cycle check**

As previously mentioned, the resulting graph should be a DAG (Directed Acyclic Graph), which is a directed graph without cycles. In the context of causal analysis and causality in general, the acyclic property of graphs is crucial to maintain the logical meaning and coherence encoded within the graph. The absence of cycles ensures that there are no circular dependencies or contradictory relationships, allowing for a clear and meaningful representation of causality in the system.

The *find\_cycles* function is a dedicated procedure designed to determine whether the constructed graph contains any cycles. It accepts an array of nodes and an array of edges as input parameters, which together represent the graph. The function makes use of the *graph-tool* Python package, a highly efficient module for graph manipulation and analysis. The underlying components of this Python package are primarily implemented in C++ to optimize performance.

The cycles are identified using the *all\_circuits* function from the *graph-tool* package.   
In case cycles exist, they are represented as lists of graph nodes and returned as an output parameter of the *find\_cycles* function.

def find\_cycles(nodes=[], edges=[]):

    if not nodes or not edges:

        return []

    g = gt.Graph(directed=True)

    nodes\_ids = {}

    v\_prop = g.new\_vertex\_property("string")

    for n in nodes:

        v = g.add\_vertex()

        v\_prop[v] = n

        nodes\_ids[n] = v

    for (n1, n2) in edges:

        e = g.add\_edge(nodes\_ids[n1], nodes\_ids[n2])

    cycles = []

    for c in gt.all\_circuits(g):

        cycles.append([v\_prop[v] for v in c])

    return cycles

**Plotting the graph**

The operation of plotting the causal graph is the last step of the casual discovery pipeline, and it is processed by the *build\_graph* procedure. This function is designed to construct and visualize a directed graph using libraries for graph and network creation, manipulation, and analysis, such as the NetworkX and Pyvis libraries. NetworkX is dedicated to general-purpose graph operations, while Pyvis serves as a visualization library suitable for generating interactive network graphs.

The function allows cycle highlighting and the creation of interactive plots.

Cycles within the graph can be highlighted by coloring relevant edges in red.

The function supports both static and interactive graph presentation modes, simplifying the visualization and analysis of entity relationships. The resulting interactive graph is then exported as an .html file.

def build\_graph(nodes, edges, bidirected\_edges, cycles, plot\_static\_graph, directory\_name, graph\_name):

    if plot\_static\_graph:

        plt.figure()

    G = nx.DiGraph()

    G.add\_nodes\_from(nodes)

    for e1, e2 in edges:

        G.add\_edge(e1, e2, color='black', style='solid')

    for cycle in cycles:

        for i in range(len(cycle) - 1):

            G[cycle[i]][cycle[i + 1]]['color'] = 'red'

        G[cycle[-1]][cycle[0]]['color'] = 'red'

    for e1, e2 in bidirected\_edges:

        G.add\_edge(e1, e2, color='grey', style='dashed')

    if plot\_static\_graph:

        pos = nx.spring\_layout(G)

        nx.draw\_networkx\_nodes(G, pos)

        nx.draw\_networkx\_labels(G, pos)

        edge\_colors = [G.edges[edge]['color'] for edge in G.edges()]

        edge\_styles = [G.edges[edge]['style'] for edge in G.edges()]

        nx.draw(G, pos, node\_color='skyblue', node\_size=1500,

                font\_size=10, font\_weight='bold', arrowsize=20, edge\_color=edge\_colors, style=edge\_styles,

                width=2)

        plt.title(graph\_name)

        plt.show()

    net = Network(directed=True, notebook=True)

    net.from\_nx(G)

    net.force\_atlas\_2based()

    net.show\_buttons(filter\_=['physics'])

    os.makedirs(directory\_name, exist\_ok=True)

    net.save\_graph(f'{directory\_name}/{graph\_name}.html')

**Immagine che contiene linea, diagramma, cerchio, design

Descrizione generata automaticamente**

Figure 3: Causal graph example

1. Benchmarks

Because of its relatively recent popularity, this subject has witnessed many contributions over the past few months. Many publications and research papers have tested the general capabilities of LLMs using standardized exams and tests written to assess human aptitude and knowledge across various domains.   
For causal reasoning capabilities, researchers have used widely known and established benchmarks with datasets from multiple domains, including medicine and climate science [1].

**4.1 Pairwise Causal Relationship Discovery using LLMs**

In the context of causal discovery abilities, these benchmark datasets mainly consist of lists of variable pairs, where each pair represents a causal relationship that can be encoded as a directed edge in a causal DAG.

The assessment of LLMs' causal discovery capabilities involves tasks that focus on identifying pairwise causal relationships, determining whether variable *A* causes variable *B* or vice-versa. These tasks involve both well-known scenarios that an average non-expert correctly address using common sense and basic field knowledge (e.g., Tubingen cause-effect pairs dataset [9]), as well as more specialized domains that require expertise in a specific field to ensure accurate understanding and interpretation (e.g., Neuropathic pain dataset [10]).  
As previously mentioned, it has been noted that prompt engineering significantly increases the accuracy of results when querying the LLM for causal dependencies and edge directions [4]. Furthermore, using advanced Language Models, such as GPT-4, along with prompt engineering techniques results even in higher accuracy.

**4.2 Full Causal Graph Identification using LLMs**

Since the primary objective of this project is to evaluate the abilities of LLMs in identifying complete causal graphs, the LLM was tested using slightly different benchmarks compared to the ones mentioned earlier.

Extending the task from simple identification of pairwise causal relationships to full graph discovery introduces additional challenges that are not present in the former task. These include, for example, the need to avoid introducing edges between unrelated variables and distinguishing between direct and indirect causes [1].  
The adopted strategy, as discussed in the previous chapters, involves enumerating all possible pairs of variables and performing the pairwise test for each pair combination.

For this project, the LLM was tested against existing causal graphs, which served as benchmarks representing the ground truth. The graphs used as ground truth [5][11] predominantly revolve around medical and health-related subjects, as the project focused on identifying causal relationships and uncovering causal graph structures within the medical context.

**4.3 Evaluation metrics [12]**

Various evaluation metrics are used to assess the quality of the obtained causal discovery results. These metrics aim to identify shared patterns between the ground truth model and the one generated from the process. Given that the ground truth when dealing with causality and causal discovery is commonly represented in a graph form (e.g., DAGs), these metrics are also related with network metrics.

These include commonly used metrics like precision, recall, F1 score, accuracy, Structural Hamming Distance (SHD) and more. The following table # lists the evaluation metrics used for the benchmark tests.

|  |  |
| --- | --- |
| **Metric** | **Description** |
| Missing edges | Number of edges that are present in the ground truth graph but not in the generated one |
| Extra edges | Number of edges that are present in the generated graph but not in the ground truth one |
| Correctly directed edges | Number of edges present in the generated graph that were correctly directed |
| Incorrectly directed edges | Number of edges present in the generated graph that were incorrectly directed |
| Structural hamming distance | Sum of missing edges, extra edges, and incorrectly directed edges |
| Precision | Measure of how many of the identified causal relationships are correct out of the total relationships identified. |
| Recall | Measures the ability to identify all actual causal relationships. |
| F1 score | Harmonic mean of precision and recall |
| Precision-Recall Curve | Depicts the trade-off between the precision and recall of the identified causal relationships. |
| Area Under PR Curve | Quantifies the overall performance by summarizing the precision-recall trade-off across different thresholds. |

**4.4 Results**

The benchmark evaluation allows to quantify the capabilities of the LLM by comparing its predictions against a ground-truth reference, with commonly used evaluation metrics, like accuracy, precision, and recall.

The benchmarks are run with different LLM-based methods and algorithms. An algorithm returning either edge at random for each pair was used as a baseline.

Table # shows the results of the different methods and algorithms applied to the benchmark tests.

As expected, the baseline algorithm (indicated by the "Random" row in the table) presents the poorest performance, with an average structural hamming distance exceeding 8 errors, a precision of 0.33, a recall of 0.38 and a resulting F1 score of approximately 0.355.

The *gpt-3.5-turbo* model instead shows better results, with an average SHD that is less than half of the distance achieved by the random baseline. Both precision and recall values show notable improvements compared to the baseline method, with a F1 score of about 0.66.

The highest performance is achieved by the most advanced LLM, the *gpt-4* model. The results show an important reduction in SHD to an average of 1.6 errors, while precision rises to 0.89 and recall achieves an impressive 0.98, leading to a F1 score of 0.93. This score nearly triples the F1 score achieved by the random baseline, showing the non-trivial contribution of LLM outputs in facilitating the identification of causal graphs.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | SHD | Precision | Recall | F1 score |
| Random | 8.3 | 0.334 | 0.379 | 0.355 |
| gpt-3.5-turbo | 4.166 | 0.710 | 0.622 | 0.663 |
| gpt-4 | **1.666** | **0.888** | **0.976** | **0.930** |

* 1. **Causal discovery from medical abstract**

The project

**Implementazione**

**Benchmarks**

**Risultati (benchmark + testi veri)**

**Discussione risultati**

**Diagramma flusso aggiornato**

**In flow pipeline analisi causale aggiungere che si parte con un testo 🡪 NER 🡪 …**

You will be provided with !!![ an abstract of a medical research paper delimited by the <Text></Text> xml tags, and ]!!! a pair of

            entities delimited by the <Entity></Entity> xml tags representing medical entities !!![ extracted from the given abstract ]!!!, such

            as medications, treatments, symptoms, diseases, outcomes, side effects, or other medical factors.

            !!![

            Text:

            <Text>{text}</Text>

            ]!!!

            Entities:

            <Entity>{e1}</Entity>

            <Entity>{e2}</Entity>

            !!![ Please read the provided abstract carefully to comprehend the context and content. ]!!!

            Examine the roles, interactions, and details surrounding the entities !!![ within the abstract. ]!!!

            Based !!![ only ]!!! on !!![the information in the abstract]!!! !!![your pretrained knowledge]!!!, determine the most likely cause-and-effect

            relationship between the entities from the following options (A, B, C, D):

            A. "{e1}" causes "{e2}";

            B. "{e2}" causes "{e1}";

            C: "{e1}" and "{e2}" are not causally related;

            D: there is a common factor that is the cause for both "{e1}" and "{e2}";

            Your response should accurately reflect the likely causal connection between the two entities based on the

            information !!![ presented in the abstract ]!!! !!![ you are aware of ]!!!.

            If no clear causal relationship is apparent, select the appropriate option accordingly.

            Then provide your final answer within the tags <Answer>[answer]</Answer>, (e.g. <Answer>C</Answer>).

You will be provided with {"an abstract of a medical research paper delimited by the <Text></Text> xml tags, and " if text else ""} a pair of entities delimited by the <Entity></Entity> xml tags representing medical entities {"extracted from the given abstract" if text else ""}, such as medications, treatments, symptoms, diseases, outcomes, side effects, or other medical factors.

  Text:

      <Text>{text}</Text>

Entities:

<Entity>{e1}</Entity>

      <Entity>{e2}</Entity>

{"Please read the provided abstract carefully to comprehend the context and content." if text else ""}

Examine the roles, interactions, and details surrounding the entities {"within the abstract" if text else ""}.

Based {"only " if text and not use\_pretrained\_knowledge else ""}on {"the information in the text " if text else ""}{"and " if text and use\_pretrained\_knowledge else ""}{"your pretrained knowledge" if use\_pretrained\_knowledge or not text else ""}, determine the most likely cause-and-effect relationship between the entities from the following listed options (A, B, C, D):

Options:

A: "{e1}" {pick\_random\_causal\_verb()} "{e2}";

B: "{e2}" {pick\_random\_causal\_verb()} "{e1}";

         C: "{e1}" and "{e2}" are not directly causally related;

         D: there is a common factor that is the cause for both "{e1}" and "{e2}";

Your response should analyze the situation in a step-by-step manner, ensuring the correctness of the ultimate conclusion, which should accurately reflect the likely causal connection between the two entities based on the information {"presented in the text" if text else ""} {"and any additional knowledge" if text and use\_pretrained\_knowledge else ""} {"you are aware of" if use\_pretrained\_knowledge or not text else ""}.

If no clear causal relationship is apparent, select the appropriate option accordingly.

Then provide your final answer within the tags <Answer>[answer]</Answer>, (e.g. <Answer>C</Answer>).

Table 1: Baseline results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Benchmark Name | Missing edges | Extra edges | SHD | Correct edge direction | Incorrect edge direction | Precision | Recall | F1 score | PRC Area |
| Asia | 6 | 15 | 21 | 2 | 15 | 0.118 | 0.25 | 0.16 | 0.245 |
| Smoking | 4 | 1 | 5 | 2 | 2 | 0.75 | 0.428 | 0.545 | 0.714 |
| Alcohol | 3 | 1 | 4 | 0 | 1 | 0 | 0 | NaN | 0.166 |
| Cancer | 2 | 4 | 6 | 2 | 5 | 0.428 | 0.6 | 0.5 | 0.576 |
| Diabetes | 2 | 5 | 7 | 1 | 7 | 0.375 | 0.6 | 0.461 | 0.55 |
| Obesity | 3 | 4 | 7 | 1 | 5 | 0.333 | 0.4 | 0.363 | 0.46 |
| AVG | **3.333** | **5** | **8.333** | **1.333** | **5.833** | **0.334** | **0.379** | **0.355** | **0.452** |

Table 2 GPT-3.5 results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Benchmark Name | Missing edges | Extra edges | SHD | Correct edge direction | Incorrect edge direction | Precision | Recall | F1 score | PRC Area |
| Asia | 3 | 5 | 8 | 5 | 5 | 0.5 | 0.625 | 0.555 | 0.593 |
| Smoking | 3 | 0 | 3 | 4 | 0 | 1 | 0.571 | 0.727 | 0.879 |
| Alcohol | 2 | 0 | 2 | 1 | 0 | 1 | 0.333 | 0.5 | 0.778 |
| Cancer | 0 | 1 | 1 | 4 | 2 | 0.833 | 1 | 0.909 | 0.917 |
| Diabetes | 2 | 4 | 6 | 1 | 6 | 0.429 | 0.6 | 0.5 | 0.577 |
| Obesity | 2 | 3 | 5 | 2 | 4 | 0.5 | 0.6 | 0.545 | 0.612 |
| AVG | **2** | **2.166** | **4.166** | **2.833** | **2.833** | **0.710** | **0.622** | **0.663** | **0.726** |

Table 3 GPT-4 results

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Benchmark Name | Missing edges | Extra edges | SHD | Correct edge direction | Incorrect edge direction | Precision | Recall | F1 score | PRC Area |
| Asia | 0 | 8 | 8 | 8 | 8 | 0.5 | 1 | 0.666 | 0.75 |
| Smoking | 1 | 0 | 1 | 5 | 1 | 1 | 0.857 | 0.923 | 0.96 |
| Alcohol | 0 | 0 | 0 | 3 | 0 | 1 | 1 | 1 | 1 |
| Cancer | 0 | 1 | 1 | 4 | 2 | 0.8333 | 1 | 0.909 | 0.917 |
| Diabetes | 0 | 0 | 0 | 5 | 0 | 1 | 1 | 1 | 1 |
| Obesity | 0 | 0 | 0 | 5 | 0 | 1 | 1 | 1 | 1 |
| AVG | **0.166** | **1.5** | **1.666** | **5** | **1.833** | **0.888** | **0.976** | **0.930** | **0.937** |

Sources:

[1] Causal Reasoning and Large Language Models: Opening a New Frontier for Causality, 02/06 (published May, 2023)

[2] [GPT - OpenAI API](https://platform.openai.com/docs/guides/gpt), 29/06

[3] [Prompt Engineering Guide | Prompt Engineering Guide (promptingguide.ai)](https://www.promptingguide.ai/), 30/06

[4] [Investigating causal understanding in LLMs](https://www.lesswrong.com/posts/yZb5eFvDoaqB337X5/investigating-causal-understanding-in-llms), 02/06

[5] Can Large Language Models Build Causal Graphs?, 12/06 (published March 2023)

[6] [[PDF] A Survey on Causal Discovery: Theory and Practice | Semantic Scholar](https://www.semanticscholar.org/reader/8733ea46dc4909017c7ad186d9c6202933c67b29), 01/06

[7] [Pyvis: Visualize Interactive Network Graphs in Python | by Khuyen Tran | Towards Data Science](https://towardsdatascience.com/pyvis-visualize-interactive-network-graphs-in-python-77e059791f01), 01/07

[8] [Causal Discovery with Language Models as Imperfect Experts (arxiv.org)](https://arxiv.org/abs/2307.02390), 14/07

[9] [Tubingen benchmark - Database with cause-effect pairs (mpg.de)](https://webdav.tuebingen.mpg.de/cause-effect/), 02/06

[10] [Causal-Discovery Performance of ChatGPT in the context of Neuropathic Pain Diagnosis](https://arxiv.org/pdf/2301.13819.pdf), 02/06

[11] [Bayesian Network Repository](https://www.bnlearn.com/bnrepository/), 20/07

[12] [Methods and tools for causal discovery and causal inference - Nogueira - 2022 - WIREs Data Mining and Knowledge Discovery - Wiley Online Library](https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1449), 01/06

[13] [On the Equivalence of Causal Models, Verma & Pearl](https://arxiv.org/ftp/arxiv/papers/1304/1304.1108.pdf), 02/06

[14] [Theoretical Impediments to Machine Learning, Pearl](https://arxiv.org/pdf/1801.04016.pdf)

[Causal Discovery Toolbox Documentation — Causal Discovery Toolbox 0.6.0 documentation (fentechsolutions.github.io)](https://fentechsolutions.github.io/CausalDiscoveryToolbox/html/index.html), SHD package