|  |  |
| --- | --- |
| Causal Graph Identification by Large Language Models | |
|  | |
| Studente/i | Relatore |
| Piqué Gregorio | Antonucci Alessandro |
| Correlatore |
| Zaffalon Marco |
| Committente |
| - |
| Corso di laurea | Codice progetto |
| Ingegneria Informatica | C10681 |
| Anno |  |
| 2023 |  |

|  |  |
| --- | --- |
| 01/09/2023 |  |

|  |  |
| --- | --- |
|  |  |

Dedica / Ringraziamenti (opzionale)

Un’eventuale dedica o eventuali ringraziamenti vanno inseriti subito dopo la copertina, su una pagina intera e in genere con uno stile differenziato rispetto al resto del testo, usando ad es. dei caratteri particolari come il **corsivo (*italic*)** e allineando il testo a destra.

**Attenzione**: di norma **non vanno ringraziate** le persone direttamente coinvolte nel progetto, quali i Docenti, i colleghi di progetto (per i progetti con più partecipanti), il rappresentante del Committente. Per queste persone un ringraziamento può essere espresso solo in casi particolari di impegno eccezionale che si vuole riconoscere.

Esempi:

*Questo libro è dedicato a mia moglie*

*Anna e a mia figlia Lucia.*

*Senza il loro sostegno, la loro comprensione*

*e il loro aiuto non lo avrei mai scritto.*

*Vorrei qui ringraziare la mia fidanzata Paola per il suo continuo sostegno*

*durante lo svolgimento del progetto.*

Indice generale

[1.1 ZZZZZZZZZZZZZZZZZZ 12](#_Toc82498934)

[1.1.1 AAAAAAAAAAAAAAAAAA 12](#_Toc82498935)

[1.2 VVVVVVVVVVVVVVVVV 13](#_Toc82498936)

[1.2.1 Ccccccccccccccccccccc 13](#_Toc82498937)

[2.1. XXXXXXXXXXXX 14](#_Toc82498938)

[2.2 SSSSSSSSSSSSSSSS 14](#_Toc82498939)

[2.2.1 RRRRRRRRRRRRRRRRRR 14](#_Toc82498940)

[3.1 HHHHHHHHHHHHHHHH 15](#_Toc82498941)

[3.2 RRRRRRRRRRRRRRRRR 15](#_Toc82498942)

Indice delle figure (opzionale, in caso di molte figure)

Indice delle tabelle (opzionale, in caso di molte tabelle)

Indice analitico (opzionale, per parole chiave)

**Abstract**

Advances in causal inference is vital across multiple fields and contexts. A correct and complete understanding of the causal relationships behind the system of interest is a fundamental requirement for making accurate decisions. Several methods and techniques can be used to identify causal relationships with the task called causal discovery, but many of these approaches present different flaws and weaknesses. Large Language Models (LLMs) can be used as a new assistant to aid human efforts and contributing to the task of causal analysis. This project aims to evaluate the ability of LLMs in identifying causal relationships and causal graphs from natural language texts. We will also present a set of techniques that can be used to improve the accuracy of LLM results. The project required the implementation of a software infrastructure to collect textual data and interact with the GPT API to process it to conduct causal discovery. The results showed that LLMs provide non-trivial contribution in helping with the identification of causal graphs. The application of LLMs to tasks of this nature is still in its early stages and has some limitations, but it has achieved some promising results and revealed new opportunities.

**Riassunto**

I progressi nell’inferenza causale sono fondamentali in diversi campi e contesti. Una comprensione corretta e completa delle relazioni causali alla base del sistema di interesse è un requisito fondamentale per prendere decisioni accurate. Per identificare le relazioni causali si può usare l’operazione chiamata "scoperta causale", applicando diversi metodi e tecniche, ma molti di questi approcci presentano diversi difetti e debolezze. I modelli linguistici di grandi dimensioni o Large Language Models (LLM) possono essere utilizzati come un nuovo assistente per aiutare gli sforzi umani e contribuire al compito dell’analisi causale. Questo progetto mira a valutare la capacità degli LLM di identificare relazioni causali e grafi causali da testi in linguaggio naturale. Verrà inoltre presentata una serie di tecniche che possono essere utilizzate per migliorare l’accuratezza dei risultati degli LLM. Il progetto ha richiesto l’implementazione di un’infrastruttura software per raccogliere dati testuali e interagire con l’API GPT per elaborarli e condurre l’operazione di scoperta causale. I risultati hanno dimostrato che gli LLM forniscono un contributo non banale nell’identificazione di grafi causali. L’applicazione degli LLM a compiti di questa natura è ancora agli inizi e presenta varie limitazioni, ma ha ottenuto risultati promettenti e ha rivelato nuove opportunità.

**Progetto assegnato**

**Descrizione**

Il progetto consiste primariamente in uno studio empirico delle possibilità di apprendere grafi causali mediante query su LLMs (large language models). Il lavoro richiede lo sviluppo dell'infrastruttura software (codice Python che interagisce con API) per l'apprendimento del grafo da testi in linguaggio naturale. Una serie di benchmark esistenti verranno usate per la validazione e per il confronto contro altre tecniche allo stato dell'arte.

**Descrizione**

Sviluppo codice. Benchmarking. Valutazione risultati e sintesi.

**Obbiettivi**

Confronto con lo stato dell’arte.

**Tecnologie**

Python + GPT API.

Chapter 1

Introduction

Advances in causal inference is crucial for many fields and contexts. In the realm of artificial intelligence and AI systems, a key challenge lies in their predominant reliance on statistical approaches and lack the ability to reason. The need for trustworthy machine learning tools has led to a growing interest in causality as a potential solution to this problem. Causality is the study of cause and effect, and causal understanding is the foundation of sound decision-making [1]. Without it, decisions are likely to be ineffective or even harmful. Causality is a fundamental way of understanding the world, and it can be used to build more accurate and reliable AI systems. Machines and systems with a model of reality, similar to that used in causality, could have the potential to achieve strong AI and artificial general intelligence [2, 3].

Another field where causality is important is the medical one. In medicine, most of the asked research questions are not associational, i.e., modelling statistical relationships between various quantities, but causal in nature; with these questions, researchers try to uncover the cause-and-effect relationships between variables, such as treatments, interventions, and outcomes. These questions cannot be answered from observed data alone and could require specific and expert domain knowledge [2].

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, as the amount of research data becomes larger and larger reaching dimensions that limit the possibility of parsing through the

enormity of available evidence. The human factor also increases the likelihood of introducing potential errors or overlooking critical graphs details.

These difficulties could be partially solved by using large language models, which have been trained on vast volumes of textual data [2]. One remarkable example of such a language model is the Generative Pre-trained Transformer (GPT) language model. GPT LLMs are designed to understand and generate human-like text based on the given prompts and are accessible, for example, with the GPT API (see Chapter 3.3 for more details) or with the LLM-based chatbot ChatGPT (Figure 1).

Immagine che contiene testo, schermata, Carattere, design

Descrizione generata automaticamente

Figure 1: ChatGPT prompt example

## Project Objective

The primary goal of this project is to conduct an empirical study to assess the possibility of performing causal analysis using 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 and provides insight into the true causal relationships between variables. The generated graph forms the basis for many, if not all, fundamental tasks in causal analysis, such as effect inference, prediction, and causal attribution [4].

Figure 2 shows an example of a simple causal graph that represents the relationship between its entities, which are encoded as graph nodes.

## Document Overview

This section provides a succinct overview of the content and objectives of each chapter in this document.

To start, chapter 2 discusses the state of the art concerning the current causal discovery techniques.

Chapter 3 then presents the methods, approaches and tools used in the project.

Chapter 4 delves deeper into the implementation details of the developed software for the project.

The subsequent chapter 5 presents the benchmark tests and the evaluation metrics used to assess the quality of the results.

Chapter 6 then shows and analyses the results achieved and discusses them, addressing also and the limitations found.

To conclude, chapter 7 presents the final considerations on the project and topic covered, briefly indicating what was addressed, highlighting the limitations of the project, and then leaving indications of possible future work.

Chapter 2

State of the Art

This chapter presents the current state-of-the-art methods and techniques used for causal analysis and causal graph identification from data, which have limitations and challenges. A different approach is the knowledge-based method, which focuses on the variables' metadata, rather than their data values. This is the same method adopted by LLMs, who are trained on large amounts of textual data.

## Causal discovery from data

The primary goal of this project is to conduct an empirical study to assess the possibility of performing causal analysis using LLMs.

### 1.1.1 AAAAAAAAAAAAAAAAAA

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel vulputate lobortis, orci purus consectetur felis, sed luctus neque dui eu odio. Aliquam risus nisl, egestas eu accumsan non, luctus in nibh. Nulla consectetur lacus eu enim imperdiet rhoncus. Proin lobortis tincidunt eros eu feugiat.

Immagine che contiene strumento, cilindro, macchina

Descrizione generata automaticamente

Figura 1: Martinetto meccanico

* **intercambiabilità**: Lorem ipsum dolor sit amet, consectetur adipiscing elit.
* **intercambiabilità**: Lorem ipsum dolor sit amet, consectetur adipiscing elit.
* **intercambiabilità**: Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus;

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit.;
2. Lorem ipsum dolor sit amet, consectetur adipiscing elit..

## 1.2 VVVVVVVVVVVVVVVVV

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus,

### 1.2.1 Ccccccccccccccccccccc

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus,

Capitolo 1

## ZZZZZZZZZZZZZZZZZZ

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel

### 1.1.1 AAAAAAAAAAAAAAAAAA

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel vulputate lobortis, orci purus consectetur felis, sed luctus neque dui eu odio. Aliquam risus nisl, egestas eu accumsan non, luctus in nibh. Nulla consectetur lacus eu enim imperdiet rhoncus. Proin lobortis tincidunt eros eu feugiat.

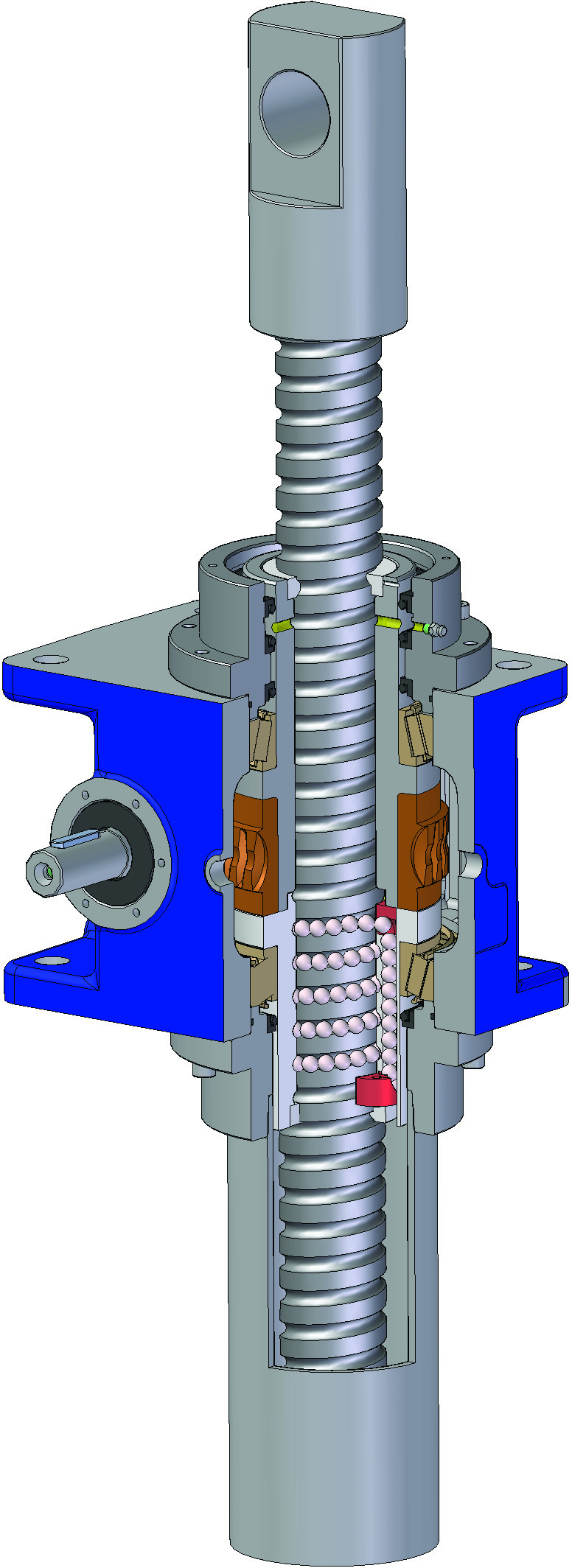


Figura 1: Martinetto meccanico

* **intercambiabilità**: Lorem ipsum dolor sit amet, consectetur adipiscing elit.
* **intercambiabilità**: Lorem ipsum dolor sit amet, consectetur adipiscing elit.
* **intercambiabilità**: Lorem ipsum dolor sit amet, consectetur adipiscing elit.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus;

1. Lorem ipsum dolor sit amet, consectetur adipiscing elit.;
2. Lorem ipsum dolor sit amet, consectetur adipiscing elit..

## 1.2 VVVVVVVVVVVVVVVVV

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus,

### 1.2.1 Ccccccccccccccccccccc

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus,

Capitolo 2

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel vulputate lobortis, orci purus consectetur felis, sed luctus neque dui eu odio. Aliquam risus nisl, egestas eu accumsan non, luctus in nibh. Nulla consectetur lacus eu enim imperdiet rhoncus. Proin lobortis tincidunt eros eu feugiat.

## 2.1. XXXXXXXXXXXX

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel vulputate lobortis, orci purus consectetur felis, sed luctus neque dui eu odio. Aliquam risus nisl, egestas eu accumsan non, luctus in nibh. Nulla consectetur lacus eu enim imperdiet rhoncus. Proin lobortis tincidunt eros eu feugiat.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evento** | **Esempio** | **Obiettivo** | **Criteri** |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |

Tabella 1: Esempio di informazioni

## 2.2 SSSSSSSSSSSSSSSS

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet.



Figura 2: Turbina

### 2.2.1 RRRRRRRRRRRRRRRRRR

Generalmente le 6 viste permettono di rappresentare in maniera esaustiva la forma esterna di praticamente tutti gli oggetti. Di regola il particolare viene disegnato nelle viste sufficienti per una sua rappresentazione esaustiva.Capitolo 3

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel

## 3.1 HHHHHHHHHHHHHHHH

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet. Mauris nec tellus purus, vitae sollicitudin nunc. Aenean suscipit, ipsum vel vulputate lobortis, orci purus consectetur felis, sed luctus neque dui eu odio. Aliquam risus nisl, egestas eu accumsan non, luctus in nibh. Nulla consectetur lacus eu enim imperdiet rhoncus. Proin lobortis tincidunt eros eu feugiat.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evento** | **Esempio** | **Obiettivo** | **Criteri** |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |

Tabella 2: Coefficienti di calcolo

## 3.2 RRRRRRRRRRRRRRRRR

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut rutrum fringilla tortor, id ullamcorper mauris elementum sit amet.

|  |  |  |  |
| --- | --- | --- | --- |
| **Evento** | **Esempio** | **Obiettivo** | **Criteri** |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |
| XXXXXXXXX | XXXXXXXXX | XXXXXXXXX | XXXXXXXXX |

Tabella 3: Coefficienti di calcolo

Bibliography

[1] Long, S. and Piché, A. and Zantedeschi, V. and Schuster, T. and and Drouin, A., “Causal Discovery with Language Models as Imperfect Experts”, *arXiv e-prints*, 2023. doi:10.48550/arXiv.2307.02390.

[2] Kıcıman, E. and Ness, R. and Sharma, A., and Tan, C., “Causal Reasoning and Large Language Models: Opening a New Frontier for Causality”, *arXiv e-prints*, 2023. doi:10.48550/arXiv.2305.00050.

[3] Long, S., “Can large language models build causal graphs?”, *arXiv e-prints*, 2023. doi:10.48550/arXiv.2303.05279.

[4] Willig, M. and Zečević, M. and Dhami, D. S. and Kersting, K. (2023, April), “Causal Parrots: Large Language Models May Talk Causality But Are Not Causal”, <https://openreview.net/forum?id=tv46tCzs83>

[5] Zhang, C., “Understanding Causality with Large Language Models: Feasibility and Opportunities”, *arXiv e-prints*, 2023. doi:10.48550/arXiv.2304.05524.

[6] Pearl, J. (2018, February). “Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution”. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining (pp. 3-3). ACM. doi:10.1145/3159652.3176182

[7] Zhang, Y. and Zhang, X. and & Zhang, W. (2019, January). “Review of Causal Discovery Methods Based on Graphical Models”. Frontiers in Genetics, 10, 524. doi:10.3389/fgene.2019.00524

[8] Verma, T. and Pearl, J. (1990). “Equivalence and Synthesis of Causal Models”. Probabilistic and Causal Inference.

[9] Zanga, A. and Stella, F., “A Survey on Causal Discovery: Theory and Practice”, *arXiv e-prints*, 2023. doi:10.48550/arXiv.2305.10032.

[10] OpenAI, “GPT – OpenAI API” [Online], 29.06.2023, <https://platform.openai.com/docs/guides/gpt>

[11] “Prompt Engineering Guide” [Online], 30.06.2023, <https://www.promptingguide.ai/>

[12] Hobbhahn, M. and Lieberum, T., “Investigating causal understanding in LLMS” [Online], 02.06.2023, <https://www.lesswrong.com/posts/yZb5eFvDoaqB337X5/investigating-causal-understanding-in-llms>

[13] The PyCoach, “OpenAI and Andrew Ng Just Released a FREE ChatGPT Prompt Engineering Course” [Online], 10.08.2023, <https://artificialcorner.com/openai-and-andrew-ng-just-released-a-free-chatgpt-prompt-engineering-course-b0884c03e946>

[14] Fulford, I. and Ng, A., “ChatGPT Prompt Engineering for Developers” [Online], 10.08.2023, <https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/>

[15] Shieh, J., “Best practices for prompt engineering with OpenAI API” [Online], 10.08.2023, <https://help.openai.com/en/articles/6654000-best-practices-for-prompt-engineering-with-openai-api>

[16] “PubMed”, Wikipedia, The Free Encyclopedia [Online], 30.06.2023, <https://en.wikipedia.org/wiki/PubMed>

[17] Sayers, E., “A General Introduction to the E-utilities”, National Center for Biotechnology Information [Online], 02.06.2023, <https://www.ncbi.nlm.nih.gov/books/NBK25497/>

[18] Peixoto, P. T., “The graph-tool python library”, (2014) DOI: 10.6084/m9.figshare.1164194

[19] Hagberg, A. A. and Schult, D. A. and Swart, P. J., “Exploring network structure, dynamics, and function using NetworkX”, in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), Aug 2008

[20] Tran, K., “Pyvis: Visualize Interactive Network Graphs in Python” [Online], 01.07.2023, <https://towardsdatascience.com/pyvis-visualize-interactive-network-graphs-in-python-77e059791f01>

[21] Mooij, J. M. and Peters, J. and Janzing, D. and Zscheischler, J. and Schölkopf, B., “Distinguishing cause from effect using observational data: methods and benchmarks”, arXiv e-prints, 2014. doi:10.48550/arXiv.1412.3773.

[22] Tu, R. and Ma, C. and Zhang, C., “Causal-Discovery Performance of ChatGPT in the context of Neuropathic Pain Diagnosis”, arXiv e-prints, 2023. doi:10.48550/arXiv.2301.13819.

[23] Scutari, M., “Bayesian Network Repository” [Online], 20.07.2023, <https://www.bnlearn.com/bnrepository/>

[24] Kalainathan, D. and Goudet, O., “Causal Discovery Toolbox: Uncover causal relationships in Python”, *arXiv e-prints*, 2019. doi:10.48550/arXiv.1903.02278.

[25] Nogueira, A. and Pugnana, A. and Ruggieri, S. and Pedreschi, D. and Gama, J. (2022). “Methods and tools for causal discovery and causal inference”. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 12. 10.1002/widm.1449.

[26] Huang P. (2023). “Research progress on the protective mechanism of a novel soluble epoxide hydrolase inhibitor TPPU on ischemic stroke”. Frontiers in neurology, 14, 1083972. https://doi.org/10.3389/fneur.2023.1083972

****Appendix****

1. Implementation details

The source code used for this project is available at **repository\_url**.

A.1. Scraping

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

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

    url\_params = {

            'db': db,

            'term': terms\_string,

            'retmax': retmax,

            'api\_key': api\_key,

        }

    if use\_history:

        url\_params['usehistory'] = 'y'

    url = f'{base\_url}esearch.fcgi'

    response = requests.get(url, params=url\_params)

    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

def get\_articles\_data(ids, web\_env, query\_key, db, retmax):

    use\_web\_env = not ids

    url\_params = {

        'db': db,

        'rettype': 'abstract',

        'retmode': 'xml',

        'api\_key': api\_key,

        'retmax': retmax,

    }

    if use\_web\_env:

        url\_params['query\_key'] = query\_key

        url\_params['WebEnv'] = web\_env

    else:

        ids\_string = ','.join(map(str, ids))

        url\_params['id'] = ids\_string

    url = f'{base\_url}efetch.fcgi'

    response = requests.get(url, params=url\_params)

    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:

        article\_data = {

            'id': article.find('PMID').get\_text(),

            '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')]],

        }

        pub\_date = article.find('PubMedPubDate', {'PubStatus': 'received'})

        if pub\_date:

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

        data = pd.concat([data, pd.DataFrame(article\_data)]).reset\_index(drop=True)

    return data

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 drop\_id\_duplicates:

        data = data.drop\_duplicates(subset=['id'], inplace=False)

    if drop\_empty\_abstracts:

        data = data[data['abstract'] != '']

    if drop\_nan\_abstracts:

        data = data.dropna(subset=['abstract'])

    if drop\_abstracts\_with\_matches and drop\_abstracts\_matches:

        data = data[~data['abstract'].str.startswith(tuple(drop\_abstracts\_matches))]

    if drop\_date\_nan:

        data = data.dropna(subset=['pub\_date'])

    if drop\_date\_before:

        data = data[data['pub\_date'] > drop\_date\_before]

    if drop\_date\_after:

        data = data[data['pub\_date'] < drop\_date\_after]

    if search\_terms:

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

    return data.reset\_index(drop=True)

A.2. Causal discovery

The complete GPT user messages can be found at Appendix B.

Sono inserite direttamente nel fascicolo della documentazione.

Vanno identificate singolarmente con A1, A2, A3, ecc. e i relativi titoli, oppure a gruppi con A1, A2, ecc., B1, B2, ecc., C1, C2, ecc., nel caso si vogliano evidenziare dei blocchi di appendici dello stesso tipo, come ad es. schemi, diagrammi, listati, ecc.

Nel caso di appendici su cui sia impossibile o troppo complicato inserire la numerazione delle pagine, come ad esempio listati già formattati, va inserita una prima pagina numerata che riporti il contenuto e il numero di pagine dell’appendice stessa, facendo poi seguire le pagine originali (con numerazione autonoma o senza numerazione).

****Allegati****

Fanno parte della documentazione, ma non del relativo fascicolo, trattandosi di materiale separato, anche se riferito alla documentazione stessa.

Si tratta del CD contenente la documentazione stessa e altro materiale riferito al progetto, ev. fascicoli separati (come ad es. un Manuale d’uso), ev. materiale sperimentale del progetto.

L’identificazione avviene come per le Appendici, usando però lettere iniziali diverse, in modo da non confondere Appendici e Allegati.

Importante: gli Allegati, essendo separati, vanno sempre corredati di scritte o etichette che li identifichino come relativi al progetto e alla documentazione cui si riferiscono (titolo, codice, ecc.).