cos'è la causalità

La causalità è un concetto fondamentale nell'ambito della filosofia, della scienza e della statistica che riguarda la relazione tra causa ed effetto. In generale, si riferisce all'idea che un evento, chiamato causa, sia responsabile dell'occorrenza di un altro evento, chiamato effetto.

cos'è l'inferenza causale

L'inferenza causale, d'altra parte, si riferisce al processo di trarre conclusioni sulla relazione di causalità tra due eventi o fenomeni. L'inferenza causale cerca di stabilire se un evento è la causa di un altro evento osservato, utilizzando metodi logici o empirici.

Nel contesto scientifico, l'inferenza causale è spesso basata sull'analisi di dati e sull'applicazione di metodi statistici. Ad esempio, gli esperimenti controllati sono spesso utilizzati per stabilire relazioni causali. In un esperimento controllato, vengono manipolate delle variabili indipendenti per osservare gli effetti sulle variabili dipendenti. Se gli effetti osservati sono consistenti e ripetibili, si può fare inferenza causale e affermare che la variabile indipendente ha causato gli effetti osservati sulla variabile dipendente.

Tuttavia, in molti casi, l'inferenza causale può essere più complessa a causa di fattori come la presenza di molte variabili nascoste, il tempo e l'ordine degli eventi, la casualità e le relazioni complesse tra le varie variabili. Pertanto, stabilire una relazione causale può essere un compito difficile e richiede un'analisi accurata dei dati e una corretta interpretazione dei risultati.

È importante sottolineare che l'inferenza causale non implica necessariamente una relazione di causa-effetto diretta e univoca tra due eventi. Spesso, si parla di associazioni causali o effetti mediati, dove una serie di eventi o fattori possono contribuire all'effetto osservato. Inoltre, l'inferenza causale non può dimostrare con certezza assoluta una relazione di causa-effetto, ma può fornire evidenze solide per supportare un'ipotesi causale.

attuali metodi di causalità utilizzati

Ci sono diversi metodi di causalità utilizzati attualmente per studiare le relazioni causali tra eventi o fenomeni. Alcuni dei metodi più comuni includono:

1. **Esperimenti controllati**: Gli esperimenti controllati sono spesso considerati il "gold standard" per l'inferenza causale. In questi esperimenti, i ricercatori manipolano deliberatamente una variabile indipendente per osservare gli effetti sulle variabili dipendenti, controllando al contempo altre variabili che potrebbero influenzare i risultati. Questo metodo consente di stabilire una relazione di causa-effetto in modo più definitivo, specialmente se viene applicato un randomizzazione casuale.
2. **Analisi longitudinale**: L'analisi longitudinale implica l'osservazione di una popolazione nel corso del tempo per studiare i cambiamenti nelle variabili e identificare possibili relazioni causali. Questo metodo può coinvolgere l'uso di studi di coorte (in cui un gruppo di individui viene seguito nel tempo) o di studi di panel (in cui gli stessi individui sono osservati in diverse occasioni). L'analisi longitudinale può aiutare a identificare il tempo e l'ordine degli eventi e a ridurre l'effetto di variabili nascoste.
3. **Analisi dei dati osservazionali**: Nell'analisi dei dati osservazionali, i ricercatori studiano relazioni causali utilizzando dati raccolti in modo non sperimentale, ad esempio attraverso indagini, registri o dati storici. Questi dati possono essere analizzati utilizzando metodi statistici avanzati, come i modelli di regressione, per controllare variabili di confondimento e stabilire relazioni causali plausibili.
4. **Esperimenti naturali**: Gli esperimenti naturali sfruttano eventi o situazioni naturali che agiscono come esperimenti naturali. Ad esempio, l'analisi dell'impatto di un disastro naturale su una comunità può fornire indizi sull'effetto causale del disastro su variabili come la salute o l'economia.
5. **Metodi basati su machine learning**: L'applicazione di algoritmi di machine learning può aiutare a identificare relazioni causali complesse. Ad esempio, l'uso di algoritmi di regressione causale o modelli di apprendimento causale può aiutare a identificare i fattori che influenzano un evento o un risultato specifico.

È importante notare che nessun metodo di causalità è perfetto e ogni approccio ha i suoi limiti. Spesso, è necessario combinare diversi metodi per ottenere una migliore comprensione delle relazioni causali. Inoltre, la scelta del metodo dipenderà dal contesto specifico della ricerca e dalle risorse disponibili.

modelli ML di apprendimento causale

I modelli di apprendimento causale sono un'area di ricerca nel campo del machine learning che si concentra sull'identificazione delle relazioni di causa-effetto tra variabili. A differenza degli approcci tradizionali di machine learning che mirano a predire un'etichetta o un risultato basandosi su correlazioni tra variabili, i modelli di apprendimento causale mirano a determinare se una variabile influenza causalmente un'altra variabile.

I modelli di apprendimento causale cercano di rispondere a domande come "**Quali sono le cause che portano a un determinato effetto?"** o "**Quali variabili possono essere manipolate per ottenere un risultato desiderato?".** Questi modelli si basano su principi e metodi tratti dalla teoria causale e dalle statistiche.

Uno dei principali obiettivi dei modelli di apprendimento causale è l'identificazione delle relazioni causali a partire dai dati osservazionali. Ciò può essere particolarmente complesso poiché spesso i dati osservazionali non forniscono informazioni dirette sulla causalità. Pertanto, i modelli di apprendimento causale cercano di identificare pattern, strutture e dipendenze che suggeriscono relazioni causali plausibili.

Tra i principali metodi utilizzati nei modelli di apprendimento causale ci sono:

1. **Alberi di decisione causali**: Questi modelli utilizzano alberi di decisione per rappresentare le relazioni causali tra variabili. Ogni nodo dell'albero rappresenta una variabile e le divisioni successive rappresentano relazioni causali.
2. **Grafi causali**: I grafi causali sono rappresentazioni grafiche delle relazioni causali tra variabili. In questi grafi, i nodi rappresentano le variabili e gli archi rappresentano le relazioni causali.
3. **Modelli di regressione causale**: Questi modelli utilizzano metodi di regressione per stimare l'effetto causale di una variabile su un'altra, controllando contemporaneamente altre variabili.
4. **Inference Engine di Bayes**: Questo approccio combina i principi della teoria causale con l'inferenza bayesiana per identificare le relazioni causali tra variabili.
5. **Metodi di apprendimento automatico controtestuali**: Questi metodi cercano di identificare l'effetto causale di una variabile manipolando artificialmente la variabile di interesse e confrontando i risultati con un gruppo di controllo.

I modelli di apprendimento causale sono ancora un'area di ricerca attiva e in continua evoluzione. Mentre offrono opportunità promettenti per l'identificazione delle relazioni causali da dati osservazionali, richiedono ancora attenzione e ulteriori sviluppi per affrontare le sfide associate all'identificazione e all'interpretazione delle relazioni causali.

LLMS (es. GPT) per cercare relazioni causali

È possibile utilizzare grandi modelli del linguaggio come GPT o BERT per cercare relazioni causali tra entità a partire da un testo dato in input. Questi modelli possono essere addestrati su grandi quantità di dati testuali provenienti da varie fonti, e possono imparare rappresentazioni linguistiche ricche che includono informazioni implicite sulle relazioni causali.

Se un modello come GPT o BERT è stato addestrato su un corpus di testi che contiene informazioni sulla causalità, il modello potrebbe aver appreso delle associazioni tra le entità e le espressioni linguistiche che indicano relazioni causali. Inoltre, questi modelli possono catturare relazioni implicite attraverso l'analisi dei contesti in cui le entità compaiono nel testo.

Tuttavia, è importante notare che i modelli del linguaggio non possiedono una vera comprensione concettuale o una conoscenza causale intrinseca. L'apprendimento di queste relazioni è basato sulla statistica e sulla co-occorrenza dei dati di addestramento. Pertanto, le relazioni causali identificate dai modelli del linguaggio possono essere influenzate da bias o correlazioni superficiali presenti nei dati di addestramento

Per utilizzare i modelli del linguaggio per l'identificazione delle relazioni causali, è importante valutare attentamente le loro prestazioni, testarli su dati di validazione o di test appropriati e confrontarli con altre fonti di conoscenza causale, come studi scientifici o basi di conoscenza specifiche del dominio.

Inoltre, è importante notare che l'identificazione delle relazioni causali richiede spesso una comprensione più approfondita del dominio e delle variabili coinvolte. Pertanto, l'uso di modelli del linguaggio può essere un punto di partenza per generare ipotesi o scoprire possibili relazioni causali, ma richiederà una validazione e un'analisi più approfondita per ottenere conclusioni affidabili.

Methods for *Causal Discovery* and *Inference*[1]

correlation does not imply causation. In general, assuming that if there is a correlation, there is also causation is a fallacy due to omitted data or links for which (biased) human reasoning leads to erroneous conclusions. causation does not imply correlation. There are cases in which, although there is a strong causal relationship between events, there is no evidence of correlation in a specific sample (e.g.: spontaneous generation in inanimate matter 🡪 Pasteur).

The article discusses 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. [2]

Starting from a set of observational data, **causal discovery** tries to infer the causal relationship across the different variables in the dataset. **Causal inference** focuses on testing whether two variables are related and assessing the impact of one on the other. Clearly, these two tasks are antipodal: on the one hand, causal discovery does not assume any relationship among involved variables; rather, it is inferred directly from a set of data. On the other hand, causal inference assumes a relationship among variables and tries to test and quantify the actual relationship in the available data.

The Causal Discovery Problem[4]

**Causal Graph**: A causal graph G is a graphical description of a system in terms of cause-effect relationships, i.e. the causal mechanism. **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.

The causal discovery problem consists in selecting a causal graph as a possible explanation for a given data set. Formally, let G be the set of graphs defined over the variables V of a dataset D and G∗ ∈ G be the true but unknown graph from which D has been generated.

**Causal Discovery Problem**: The causal discovery problem consists in recovering the true graph G∗ from the given data set D (uncover the cause-and-effect relationships that drive the observed data), by answering questions such as "Which variables directly affect each other?" or "What is the causal directionality between variables?".   
However, inferring causality solely from observational data is challenging because correlations alone do not establish causation. Many factors, including confounding variables and feedback loops, can give rise to spurious correlations, making it difficult to determine the true causal relationships. The assumption is that the graph's relationships accurately represent the underlying conditional independence relationships. This assumption, known as d-faithfulness or "directed faithfulness," requires the graphical model to be a directed acyclic graph (DAG).

**Markov property**: In a causal graph, the Markov property refers to the conditional independence relationships that hold among the variables represented in the graph. The Markov property states that each variable in the graph is independent of its non-effects (parents in the graph) given its direct causes (children in the graph) and the knowledge of the causal structure; e.g.: given a causal graph, if we condition on the values of a variable's direct causes (parents) in the graph, then the variable becomes independent of its non-effects (other variables in the graph) that are not direct causes or direct effects. By enforcing the Markov property, causal graphs provide a compact representation of the causal relationships among variables and facilitate the identification and estimation of causal effects in the presence of observed data. The Markov property tells how variables are related to each other based on conditional probabilities. D-separation helps to identify when variables are independent of each other given a specific set of other variables. It helps understanding the relationships and dependencies between variables in a graphical model.   
To find if variables are independent, one can use the *d-separation* criterion, which is based on the concept of blocked path. This may occur with specific graphical patterns, ~~such as forks (A 🡨 B 🡪 C) and colliders (A 🡪 B 🡨 C), but not chains (A 🡪 B 🡪 C).~~  A path [A – … – B] (which does not take directions of edges into consideration) is said to be *unblocked* if the path [A – … – B] contains only chains or forks; if it contains a collider, then the path is *blocked*. It is although possible to ***condition on*** a variable (in the middle of chains, forks and colliders) to “revert the role” in terms of blocking and unblocking path: it is possible to block the paths [A 🡪 X 🡪 B] or [A 🡨 X 🡪 B] and unblock the path [A 🡨 X 🡪 B], by conditioning on X.   
**Markov Assumption**: Two sets of nodes X and Y are d-separated by a set of nodes Z if all of the paths between X and Y are blocked by Z:  **==>**  . D-separation allows to read off conditional independencies in the distribution.

**Mixed graphs**: In a real-world scenario, the collected data and variables are rarely sufficient to find the causes of the system of interest. This causes a set of unobserved variables *U* to influence the system’s variables relationships and make the set of considered variables *V* to be *casually insufficient*. In this situation, a ***partially-directed graph*** allows to distinguish cause-effect relationships (A 🡪 B) and yet unknow ones (A -- B), where the direction of the edge is still uncertain. Moreover, a ***mixed graph*** allows to define *undirected* (--), *directed* (🡪) and *bidirected* (<->) edges; this graph model allows to represent relationships not yet fully understood, and more specifically distinguish between undirected and bidirected edges, where the former indicates the presence of a dependency or correlation with an unknown direction of causality (it is unknown if one influences the other or if there is a causal relationship at all), and the latter indicates that there is a direct causal influence but the direction of the relationship is unknown (it is uncertain which variable is the cause and which is the effect).  
Another model for representing causal relationships between a system’s variables is an ***ancestral graph***. This allows for a more comprehensive representation of these relationships, acknowledging the possible presence of underlying factors and variables that have an influence on the observed ones.

Causal discovery evaluation metrics[1]

|  |  |
| --- | --- |
| **Metric** | **Description** |
| Missing edges | Number of edges that are present in the original model but not in the generated one |
| Extra edges | Number of edges that are present in the generated model but not in the original one |
| Incorrect adjacencies (undirected edges) | Number of undirected edges that are present in the generated model but not in the original one |
| Correct directed edges | Number directed edges present in the generated model that were correctly directed |
| Incorrect directed edges | Number directed edges present in the generated model that were incorrectly directed |
| Structural hamming distance | Sum of missing edges, extra edges, and incorrectly directed edges |
| Structural intervention distance | For each pair *X* and *Y* checks whether the parents of *X*  in the generated model are a valid adjustment set (Pearl, [2009](https://wires.onlinelibrary.wiley.com/doi/full/10.1002/widm.1449#widm1449-bib-0107)) in the true model. If it is, it is counted as a correct procedure. If it is not, it is counted as a mistake. |
| Adjacency precision | Adj Precision = |
| Adjacency recall | Adj Recall = |
| Arrowhead precision | Arrhd Precision = |
| Arrowhead recall | Arrhd Recall = |

LLMs for causal analysis [7]

The quality of causal world models of LLMs matters for AI safety and alignment. More capable LLMs are supposed to assist humans in important decisions and solve scientific questions. For all of these applications, it is important that their causal world model is accurate, i.e. that they accurately reflect the causal relationships of the real world [5].

LLMs bring significant new capabilities which are complementary to existing causal methods. They do so by capturing the human domain knowledge relevant to the task, which forms an essential part of any causal analysis. As a result, LLMs have the capability of transforming how causal analysis is done, with the potential to automate or assist in each step of a causal reasoning process (causal question 🡪 causal assumption 🡪 identification 🡪 estimation 🡪 refutation and validation 🡪 iterate on question).

Conventional causal discovery and effect inference rely strongly on prior domain knowledge of potential causal mechanisms in a system. Current best practice is to rely on human domain experts to provide this knowledge, yet correctly capturing domain knowledge in a formal representation suitable for analysis remains a challenge and is often a primary point of weakness for the validity of causal analyses. LLM capabilities now open the possibility of programmatic access to an array of (memorized or inferred) causal mechanisms, capturing general and domain-specific knowledge, and may augment human domain experts by aiding in bootstrapping, critiquing, etc.  
Many of the tasks where we previously relied on human experts alone now can be partially or entirely automated with human supervision.

**Evaluating LLMs behaviors [7]**Multiple probing strategies are needed to understand the causal capabilities of LLMs. We may be tempted to ascribe a particular capability to a LLM if it answers correctly to a set of questions, but the answers may not necessarily be due to its causal capability: they may, in fact, be due to other factors such as exploiting some structures in the questions or memorization of similar questions encountered in its pre-training phase.

* **Question-answer evaluation and Benchmark tests**: ask questions and evaluate the model’s answers using standardized exams and tests written to assess human capabilities
* **Memorization tests**: used to determine whether the LLM was exposed to the benchmark answers and has memorized them (done by giving the LLM part of a sentence and ask it to complete the remaining part)
* **Redaction tests**: redaction and perturbation of the input (words, one by one) to see if and how the answers change each time (changes in the answer indicate the LLM is probably attending to the redacted input/word)

**LLMs and Causal Discovery [7]**Causal discovery 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 (if not all) other fundamental tasks in causal analysis (e.g. effect inference, prediction, attribution).

With real-world observational data, it is generally not possible to perform causal identification and discovery of the correct 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.

**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) [8].

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. 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 covariance-based causal discovery.

LLMs enable knowledge-based causal discovery and achieve competitive performance in determining pairwise causal relationships between variables, across datasets from multiple domains, including medicine and climate science [7].

~~The tests of LLMs capabilities in causal discovery reported are based on pairwise causal relationship discovery task, using the~~ ~~Tubingen cause-effect pairs dataset [9] (which contains commonly known variables that an average non-field expert could answer correctly using common sense and base knowledge), and the Neuropathic pain dataset [] (constructed by medical experts and containing variables and causal relationships whose understanding require specialized medical knowledge).~~   
The evaluation of LLMs' causal discovery abilities involves tasks that focus on identifying pairwise causal relationships, using both commonly known subjects that an average non-field expert could answer correctly using common sense and base field knowledge (Tubingen cause-effect pairs dataset [9]), and more specialized domains that require expertise in a specific field for accurate understanding and interpretation (Neuropathic pain dataset [6]).  
Using LLMs, it has been observed that prompt engineering increases considerably the results accuracy when querying the LLM on causal dependencies and edge directions [10]. Moreover, using gpt-4 with these prompt engineering techniques results in a weighted accuracy of 96% [7].  
The main prompt engineering techniques applied are prepending the prompt with the message “You are a helpful assistant for causal reasoning”, to try steering the output space to more causally consistent answers, asking a single question regarding the direction of the causal dependency (whether A 🡪 B or A 🡨 B) and to answer with a step-by-step explanation.

Although these results on the benchmark datasets are significantly higher than the best covariance-based causal discovery algorithms, cases occur where the LLM answers with an argument that ultimately ends with different answers from the ground-truth.   
~~In some cases, after a question where the LLM performs causal reasoning to arrive to a correct answer, a similar question (in terms of requiring the same causal knowledge) makes the LLM answer with an incoherent argument and provide an incorrect answer: this happens, for example, from the asking if the age of an abalone may cause its length to increase to asking if the age of an abalone may cause the its diameter to increase.~~  
In some cases (like expected), the LLM exhibits inconsistent behavior in causal reasoning. For instance, when asked a question that requires the model to perform causal reasoning and provide a correct answer, a similar question that requires the same causal knowledge may result in an incoherent argument and an incorrect answer. This inconsistency can be observed when comparing questions such as whether the age of an abalone can cause its length to increase versus whether the age of an abalone can cause its diameter to increase: in this example, the LLM correctly identifies the direction of the causal relation of the first question (that age may cause the length to increase), but for the second one (requiring arguably the same causal reasoning abilities) provides an incoherent argument and incorrect answer (stating that changing the diameter of an abalone causes a change in its age).  
In other instances, using its training knowledge, the LLM ~~determines relations between entities~~ makes points that were not necessarily taken in consideration in the benchmark’s ground-truth; this happens with a question regarding stock returns of related companies, where the LLM correctly identifies the subsidiary relation between the two companies, but argues (correctly) that since stock markets do not follow fixed patterns, major changes in one company may cause different consequences based on the situation, e.g. people investing more or less in the other company.  
Another case of incorrect final answer happens when the LLM performs a correct and coherent step-by-step explanation of its reasoning, but ultimately chooses a wrong final answer inconsistent with its own argument. Nevertheless, this kind of error is easier to verify, by using the LLM itself: this can be done by adding a consistency check layer to verify the coherence of the final answers, using GPT-4. The more advanced LLM is in fact capable of finding the logical inconsistency of the output’s reasoning and provide the correct answer.

**Full graph discovery [7]**To extend the task from pairwise causal relationships discovery to full graph discovery, a basic strategy is to enumerate all possible variable pairs and repeat the pairwise test for each pair combination. This full graph discovery task poses additional challenges that do not occur in the simple pairwise discovery.  
~~These include distinction of edge direction (whether A 🡪 B or A 🡨 B), direct and indirect causes (if A🡪B🡪C, it might be correct in pairwise discovery to output A🡪B and A🡪C, but in the full graph outputting the A🡪C edge would be wrong).~~   
These include not adding edges between non-related variables, distinguishing the edge direction (whether A 🡪 B or A 🡨 B), distinguishing direct and indirect causes (if A🡪B🡪C, it might be correct in pairwise discovery to output A🡪B and A🡪C, but in the full graph outputting the A🡪C edge would be wrong) all considering the set of the task’s input variables (if B is not present in the input variable set, A🡪C is a valid edge and should be included).

Again, correct prompt engineering [6][7][10] can lead to substantial increase in the model’s accuracy.   
~~For this full graph discovery task, another option is given to the LLM to find non existing relationships between variables [7].~~   
For this full graph discovery task, the LLM has an additional answer option (“Option C: *No causal relationship exists*”) to find non existing relationships between variables [7]. For the Arctic Sea ice dataset [11], the gpt-4 LLM-based algorithm results in being competitive for causal discovery with state-of-the-art covariance-based discovery algorithms [7].

4 LLMs for Actual Causality and Causal Judgments

The LLM’s ability to apply trained domain knowledge and “common sense” makes it a valuable real-world application for causality tasks. However, it may still struggle and fail in unpredictable ways, so additional input context may help for having more accurate answers.  
*Actual causality* concerns problems of responsibility attribution in real world applications, such as legal reasoning, machine failure debugging, and many more…

To define causality in a way that is consistent with how humans naturally attribute cause and concepts of responsibility and event outcomes, researchers have attempted to use a Structural Causal Model (SCM). The task of finding this causal model has proven very difficult, because human causality judgements depend on background elements, context, and factors that are difficult to formalize in a SCM. Sidestepping the challenge of formalizing many concepts in a causal model, LLMs can capture and extract the background context of events directly from natural language descriptions of the event.

Among the key elements of Actual causality is counterfactual reasoning (*if X had not occurred, Y would not have occurred*): this involves simulating alternative scenarios where events may not have happened, and then reason about the possible consequences. Counterfactual reasoning is a desirable skill for LLMs, as it would assist humans in decision making and planning. A benchmark containing 275 multiple-answer instances was used to evaluate counterfactual reasoning of the LLMs; GPT-4 reaches an accuracy of over 92% (against the human accuracy of 98%). These results are achieved by giving the LLM the concatenation of a premise and a counterfactual question, and the list of possible answers (the system message of “You are a helpful assistant for counterfactual reasoning” is additionally provided). The results show that the LLM (GPT-4) can simulate different scenarios and answer what-if questions. In some circumstances, when the LLM’s answers diverge from the benchmark, the LLM reveals ambiguity in the input text given to describe the scenario.   
In general, minimizing ambiguity (by giving additional details) when interacting with the LLM, can increase the chances of a correct answer.

**Inferring actual causality**LLMs made possible causally reasoning directly over natural language. Another analysis investigates whether the LLM can infer if events are necessary or sufficient: *necessity* and *sufficiency* are two key concepts of causality. Necessary causality refers to an event that, if it had not occurred, the outcome event would not have happened. Sufficient causality refers to an event that, if it had occurred, the outcome event would have occurred.  
The results show how LLMs (especially GPT-4) are more accurate in deciding necessity of cause (over 92% accurate) than sufficiency (78% accurate); this may be because necessity always involves comparing the output under a counterfactual scenario (where sufficiency may be nuanced).  
LLM reasoning represent a potential tool for causality, by understanding scenarios described using natural language, but also show the lack of robustness with unexpected and unpredictable failures.

It is then tested whether LLMs provide sufficient performance in inferring *normality*: normality questions if an event or agent caused an outcome (uncaring if it was intentional or not).

[Github - ChatGPT causality pairs](https://github.com/amit-sharma/chatgpt-causality-pairs)

Causal Parrots [12] (published April 2023)

This article discusses whether LLMs, even when succeeding in doing causal inference and causal reasoning, are actually reproposing a *meta*-SCM (Structural Causal Model) built on data the LLMs was trained on; this would make the LLM some kind of “causal parrots”, reciting learned knowledge embedded in the data.

It is also debated whether “just scaling” large models (BERT, GPT, DALL-E) is a sufficient condition for progression towards AGI (Artificial General Intelligence). The paper questions the extent to which the LLMs (called *foundation models*) *can* talk causality.

Of all things, the authors of [12] focus on analyzing whether the LLMs give causal facts as answers to a causal question.

The LLMs (GPT-3, Luminous, OPT) causal abilities are first tested with simple causal-chains and intuitive physics questions.   
In some cases, when the LLMs answer incorrectly to a well-known logic question (“Does a kilogram of steel weigh more than a kilogram of feathers?”), after reproposing the same question in an extended formulation (saying “[’A kilogram of steel is heavier than a kilogram of feathers’] *is what most people say*”) that adds no new factual information, GPT-3 catches this hint given on a meta level (usually used in human communication to indicate the presence of a trick question) and answers correctly.

**Causal discovery**Causal discovery differs from the previous tasks (causal-chains and intuitive physics questions) because LLMs are not required to reason with facts, but it become sufficient to recall the correct ones.

The first approach for causal discovery uses a naïve discovery procedure. This method does not use actual data as input to perform inferences: in this sense, it resembles how humans recall facts by relying on memorized knowledge rather than actively referencing actual data.

**Conclusions**One of the paper’s main objections is that LLMs are parrots, meaning they only mimic what was seen in the textual data during training. The authors state that if a LLM correctly answers a causal query, this is because the model might have encountered causal facts during training (like “reading” an encyclopedia page on the subject). Stating that understanding something only happens when learning from physical measurements (e.g. of the relationship between altitude and temperature) and not when simply reading up on a textual article, the authors then open a “philosophical” argument, where the LLMs can be compared to Plato’s cave: it is questioned to which extent one can learn about the real world’s functioning by just observing the shadows of its objects (the LLM is the cave, where one can observe some shadows, i.e. some correct causal answers, but the question is raised whether this would be in fact actual causality, i.e. the real world’s objects).

The authors conclude that for causal analysis, it is not possible to rely only on the capabilities of LLMs (since they cannot process physical measurements to extract actual information to ground their text-based knowledge, i.e. they do not *understand* but might only *know*, so are not capable of both generalization and justification); LLMs can though be used as a head-start to causal learning (discovery and inference).

Causality with LLMs [13] (published April 2023)

The paper tries assessing the ability of LLMs to answer causal questions. The results show that LLMs can answer correctly to causal questions with existing causal knowledge but lack the ability to provide satisfactory answers for discovering new knowledge from data or for high-stake decision-making tasks.

Causal knowledge is fundamental for human nature, as it helps us understanding real world phenomena and to try predicting them. Causal knowledge and the ability to understand causal queries is also a key skill/element for AGI.

**Causal questions**The authors identify 3 types of questions:

* **Type 1**: identifying causal relationships using domain knowledge (e.g. “*Will my minor spine injury cause numbness in my shoulder?*”)
* **Type 2**: discovering new knowledge from data (e.g. find relationships between variables by studying real-world measurements)
* **Type 3**: quantitative estimating of the consequences of actions (e.g. (i) best sales program for highest revenues or (ii) medical counsel on injection dosage)

LLMs appear to be able answering **Type 1** questions, thanks to their large textual knowledge. For the other 2 types of causal questions, the LLM needs to understand the underlying fundamental causal mechanisms to discover new knowledge from data (**Type 2** questions), understand the effect of actions on a large population (**Type 3 (i)** questions) or on a personalized level for high-stake decisions (**Type 3 (ii)** questions).

ML techniques are usually used to discover unknown causal relationships from data. Current LLMs seem to lack the ability to perform advanced symbolic and mathematical reasoning and reasoning purely on measurement data.

**Trust in assumptions** plays a vital role in ML models. In LLMs, a model that provides a step-by-step explanation regarding how the conclusions are reached from the assumptions is more trustworthy. Currently, LLMs often fail in their output answers by making mistakes in intermediate steps, or by skipping fundamental steps of their reasoning. This is another reason why high-stake decisions should not be taken by only considering LLMs’ answers.

**Precision and in-context requirement** are other key elements for LLMs to perform correct and reliable causal reasoning. LLMs should be able to understand and estimate effects of actions in both a large-scale scenario and a personal/specific/custom one, knowing that effects on a certain type of patient/subject can indicate different things compared to the same effects in another.

**Opportunities and impacts**LLMs with causal skills can have a big impact on multiple levels: first, the use of LLMs (with the ability to answer causal questions) will enlarge in multiple new domains; new training paradigms (RLHF) will improve the LLMs performance; LLMs can help in the machine-human communication process of answers and domain knowledge. The authors say there either will be an introduction of causal modules within the LLMs, or their causal reasoning capabilities will improve with new training paradigms.

**Causal modules** are thought to be a good and efficient way to introduce causality in LLMs (i.e. allowing LLMs to use “external” causal modules), rather than training LLMs to understand causality intrinsically.

Can LLMs Build Causal Graphs? [14]

The article questions whether LLMs can be used as a complementary tool in graph discovery tasks, by scoring edges between variables, since they encode great amount of common and domain-specific knowledge.

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.   
Causal graphs, in the form of Directed Acyclic Graphs (DAGs), encode contextual knowledge of variables (both observable and unobservable) and their causal dependency. Although expert opinion remains the best tool 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; LLMs are though sensitive to prompt engineering (e.g. asking a step-by-step explanation of the reasoning of the answer seem to significantly improve the LLM’s performance).

Sources:

[1] [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

[2] [Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution](https://dl.acm.org/doi/10.1145/3159652.3176182) , 01/06

[3] [Causal discovery in machine learning: Theories and applications](https://www.aimsciences.org/article/doi/10.3934/jdg.2021008) , 01/06

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

[5] [Investigating causal understanding in LLMs - LessWrong](https://www.lesswrong.com/posts/yZb5eFvDoaqB337X5/investigating-causal-understanding-in-llms), 01/06

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

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

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

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

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

[11] [Benchmarking of Data-Driven Causality Discovery Approaches in the Interactions of Arctic Sea Ice and Atmosphere](https://www.frontiersin.org/articles/10.3389/fdata.2021.642182/full), 02/06

[12] Causal Parrots: Large Language Models May Talk Causality But Are Not Causal, 07/06 (published April 2023)

[13] Understanding Causality with Large Language Models: Feasibility and Opportunities, 08/06 (published April 2023)

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

Results

**Performances**

The *causal discovery pipeline* is a Python script used for the (almost) complete causal analysis required operation, starting from textual data (pre-extracted from PubMed with another Python web-crawling script).   
It consists of 3 main steps/operations: entity extraction, pair-wise causal relationship discovery (both with gpt-3.5), and graph plotting.

It takes the *causal discovery pipeline* about **8’10’’** to complete the operations starting from the following dummy text:

“Smoking is a major cause of lung cancer. Tobacco smoke contains harmful substances that can lead to tumor formation in the lungs. Quitting smoking reduces the risk of lung cancer and improves overall health.”

The text contains about **30 words** (and approximately the same number of tokens), and the extracted entities are 8: [‘Smoking’, ‘lung cancer’, ‘Tobacco smoke’, ‘harmful substances’, ‘tumor formation’, ‘lungs’, ‘Quitting smoking’, ‘improves overall health’].