# PROBING FOR CORRELATIONS OF CAUSAL FACTS: LARGE LANGUAGE MODELS AND CAUSALITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) are subject to an ongoing heated debate, leaving open the question of progress towards AGI and dividing the community into two camps: the ones who see the arguably impressive results as evidence to the scaling hypothesis, and the others who are worried about the lack of interpretability and reasoning capabilities. By investigating to which extent causal representations might be captured by LLMs, we make a humble effort towards resolving the ongoing philosophical conflicts. We hypothesize that causal facts are part of the training data and that the LLM are capable of picking up correlations between the questions on causal relations with their expected (or "right") causal answers. We study this hypothesis two-fold, (1) by analyzing the LLM's causal question answering capabilities and (2) by probing the LLM's embeddings for correlations on the causal facts. Our analyses suggests that LLMs are somewhat capable of answering causal queries the right way through memorization of the corresponding question-answer pair. However, more importantly, the evidence suggests that LLMs do not perform causal reasoning to arrive at their answers.

## 1 THE BIG PICTURE

The two opening paragraphs provide context to recent developments in the AI/ML community as a whole, which we stress to be important as it motivates the research question being opened by the presented work.

### 1.1 AN ONGOING HEATED DEBATE ABOUT FOUNDATION MODELS

In the advent of large scale models such as BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), DALL-E (Ramesh et al., 2021), AI history suggests to repeat itself[1] as arguably impressive text generation and image synthesis results divide the community in terms of interpretation regarding the progression of the field as a whole towards the grand goal of AGI (key references involve (Marcus & Davis, 2021; Marcus, 2022) that sparked intense discussions amongst Turing awardee Yann LeCun and others via social networks). Researchers at the Institute for Human-Centered AI at Stanford recently coined said large scale models as *foundation* models to account for the "emerging paradigm" of models that provide a base from which task-specific models are derived through adaptation (Bommasani et al., 2021). The emergence of what seem to be the two different camps within the discussion around foundation models is characterized by researchers who recognize said models as significant progression towards AGI and those who do not. For the former group of "believers", the results act as corroborating evidence for the *scaling hypothesis* (Branwen, 2020; Sutton, 2019) which captures that the idea of emergent properties as a result of scaling neural network in terms of parameters and data, thereby, rooting parts of the overarching idea in results from neuroscience that suggest the human brain to "just" be a scaled up primate brain (Herculano-Houzel, 2012). An arguably similar idea, the *reward is enough* hypothesis, was recently discussed by (Silver et al., 2021). On the other side, the "non-believers" see the achieved results as a mere reflection of the sheer scale of data and parameters, put differently "the methods are old" and their lack of interpretability and reasoning capabilities will remain persistent. Turing awardee Judea Pearl who contributed seminal

[1] A short treatise that discussed patterns in the history of AI research observes: "early, dramatic success followed by sudden unexpected difficulties." (Chauvet, 2018)

work towards a rigorous formalization of causality (Pearl, 2009) announced his alliance with the latter position via social media, stating "These models are castles in the air. They have no foundations whatsoever." discrediting the models for lacking any identifiable notion to causality.

### 1.2 Foundation Models and Causality

Speaking of causality, the Pearlian counterfactual theory of causation has recently found prominent support in the AI/ML community (Schölkopf, 2022; Peters et al., 2017; Geffner et al., 2022). An increasing presence of publications at major conferences/journals concerned with the integration of causality with AI/ML (including (Janzing & Schölkopf, 2018; Lee & Bareinboim, 2019; Zečević et al., 2021) to mention a select few) suggests a growing subfield that sets a consensus on *causal* AI/ML as promising paradigm for next-generation systems. Still, as the difficulty of the integration with otherwise prominent success stories of deep learning, such as computer vision, becomes apparent, countering opinions speak out against causal AI/ML (Bishop, 2021). In this work, we take the former perspective *pro* causal AI/ML. We argue that, going back to the ongoing debate on foundation models, the questions around causality can fuel research to resolve the disagreement causing the debate to begin with. We identify the key problem of the debate to lie in exactly discussed scale of data and parameters that only further cement the inherently black-box nature of the base models. Therefore, to answer whether foundation models have made progress towards AGI and to give reason onto why causal AI/ML could be a milestone, it seems to suffice to ask and investigate the question of the extent to which *foundation models can talk causality*. For the sake of simplicity, we will take the "talking" literally and focus on LLMs in this work while leaving general foundation models (e.g. image-based models) for future work. However, for observations that we believe should also hold more generally for foundation models (like the concept of "correlations of causal facts" that we will introduce in this work) we will use the term foundation model instead of writing LLM.

**Our contribution.** We present a humble effort towards the goal of resolving the ongoing conflicts by first making an important observation on what we call "correlations of causal facts" (abbreviated CCF) which acts as the argumentative foundation (or running hypothesis) for our subsequent, second step, in which we perform a systematic analysis to grasp to which extent causal representations might be captured by the large scale language models being evaluated. We make our code publicly available at: `https://anonymous.4open.science/r/LLMsAndCausality-E220/README.md`.

**Related Work.** This present work takes inspiration from various recent results. Yet, to the best of our knowledge, it is the first to investigate the question in its presented form. For instance, Wang et al. (2021) leveraged BERT as underlying foundation model to perform inferences according to the rules of Pearl's $do$-calculus (Pearl, 2009). This allows for causal inference with the foundation model as engine, one of the things we will elaborate on further, but it misses out on the question of how causal the models themselves might be to begin with. Another work, by Khetan et al. (2022), is closer to our work in the sense that causal relations are queried by natural language directly, however, the subject of interest is orthogonal to both the ongoing debate and the investigation presented in this work. On another note, McMilin (2022) investigated selection bias within LLMs by first arguing about reasonable causal modelling assumptions and then validating them empirically. Discarding causality for a moment but with the arguably identical goal of understanding what LLMs are capable of, Zhang et al. (2022a) investigated an approach using propositional logic that concluded that LLMs only learn statistical features that inherently exist in logical reasoning problems. Also noteworthy are works such as conducted by Talmor et al. (2022) where the goal is to create a benchmark that makes explicit the deficiencies (if existent) of LLMs, which can be understood as a similar goal to understanding how the models work in the first place.

## 2 Correlations of Causal Facts

"Correlation does not imply causation," goes the famous saying (see Aldrich (1995); Pearl (2009)), that accounts for the fact that following Reichenbach's *common cause principle* a correlation between two variables might be either because one is causing the other, or because there is a third variable causing both (Reichenbach, 1956). To infer the actual causation within the system of interest, we might resort to *manipulating* the system, as another fomous saying suggests "No causation without manipulation" (Holland, 1986). A celebrated victory of Pearl's notion to causality is the *causal hiearchy theorem* (CHT) which guarantees that purely observational data collected from a

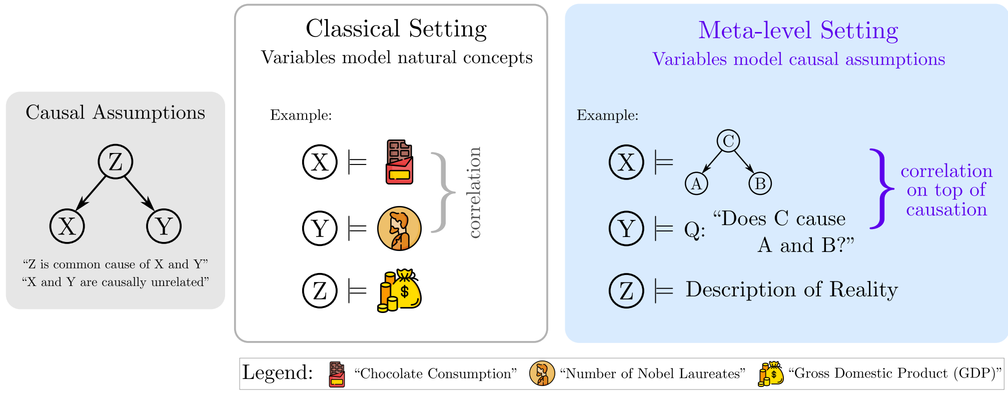



Figure 1: **Correlations of Causal Facts.** See Sec.2 for a detailed description. Causal assumptions refer to statements on the causal relationships between the variables being studied. In the classical setting, the variables denote "natural" concepts, whereas in the meta-level setting they denote causal assumptions. A foundation model will encounter data that of both kinds, thus legitimately raising the question to which extent they might be considered causal. (Best viewed in color.)

system can not be used to uniquely determine causal statements, when no other causal assumptions are available (Bareinboim et al., 2020). The CHT certainly seems to imply that *no matter how much* we scale our foundation models (in terms of data and parameters), we will never be able to perform causal inference. In a nutshell, the CHT seems to disprove the scaling hypothesis. Or does it? In this work, we argue that foundation models might be exploiting a "loop hole" in the CHT[2]. Namely, what happens if the *causal assumptions* (which are required, by the CHT, for causal inference) are represented in observational data itself? In essence, a Structural Causal Model (SCM) (Pearl, 2009; Peters et al., 2017), which is commonly recognized as the data-generating process, is not restricted to modelling "natural" concepts such as "temperature" or "chocolate consumption per capita" only. Rather, the formalism seems to allow for data that in some sense *talks about data generating processes themselves*[3] Fig.1 illustrates this idea schematically alongside an example. Essentially, we intend on asking a *philosophically* fundamental question that (as we will show) implies other interesting questions of practical interest to the AI/ML community. Namely, to which extent does *understanding* causality differ from *knowing* causality? Such a question is certainly reminiscent of the Chinese Room Argument by Searle (2009). Therefore, if one could blur "understanding" and "knowing" causality, then this would imply that foundation models are already to an extent causal. Independent of the philosophical question (which by the way is beyond AI/ML systems an unresolved question also of human cognition), knowing to which extent we can rely on our foundation models to simply know the right causal answer for a causal query has important applications in AI/ML. The foundation model could be used (i) to head start learning with rough estimates, (ii) could serve as a recognition system for hidden variables that would require an increased computational complexity, and (iii) be used as interactive modules with human-in-the-loop.

**Example from Fig.1.** In this setting, *causal assumptions* refer to things like "$X$ and $Y$ are unrelated" or "$Z$ is a common cause". Collectively this set of assumptions might be depicted as a causal graph. They are assumptions since they constrain the data generating process i.e., they live outside the data on a meta-level. Let's assume the given trivariate graph (Fig.1, left), where $X, Y, Z$ are interpreted as given in "Classical Setting" (Fig.1, middle). Note how the variables denote "natural" (possibly low-level, physical, quantifiable) concepts. Maurage et al. (2013) show how $X$ and $Y$ are correlated, yet, no causation is expected as surely $Z$ could act as a common cause. Now, consider a different encoding of the variables following the "Meta-level Setting" (Fig.1, right). With big data in the natural language settings, we can certainly expect statements such as "The GDP explains both the increased research facilities, leading to more Nobel laureates, and increased chocolate

[2]Or rather, it is a *subtle* detail that might easily be forgotten.

[3]Self-referencing systems are at the core of seminal arguments dating back to the origins of computer science. See for instance Turing's Halting problem proof (Turing et al., 1936) or Gödel's incompleteness proofs (Gödel, 1931).

production, leading to more consumption" and the corresponding graph depiction (which would also be in natural language) to occur together. Thus, we'd observe a "correlation of causal facts."

**A Causal View on CCF.** While we naturally think of observational data as simply capturing some phenomenological properties of a population and not about meta properties, these meta properties have a natural meaning in causality namely that of *causal facts* (hence the naming "correlations of causal facts"). Causal facts can be understood as data that lives on rung 2 or 3 of the Pearl causal hierarchy, that is, interventions and counterfactuals. Fascinatingly, these causal facts have been acquired somehow (e.g. through experiments) *and are now simply observed*. To give an example, physicists conducted experiments to find out about how an increase in altitude ($A$) will cause temperatures ($T$) to generally decrease due to some chemical processes. Knowledge of $A \rightarrow T$ is a causal fact which required interventions for the fact to be discovered. However, the LLM will be capable of learning this intervention information *without performing interventions* since it will simply "read a book" which explains the causal mechanism that relates altitude and temperature.

## 3 What Structures do Large Language Models Find?

**Structure Discovery via LLMs.** Since we want to query the large language model for what it has learned, while unfortunately being unable to directly measure the expected "correlations of causal facts"[4], we resort to indirect measurements by simply querying the black box system systematically. In a sense, it is the analogue procedure of what the community around explainable AI in computer vision, that is, "opening the black box" by attributing to the input the changes it caused on the output[5]. Such attributions offer a way to "understand what the neural net sees" (Linardatos et al., 2020). Another reason to focus on LLMs first is that they provide a natural way of expressing causal assumptions (e.g. "does $X$ cause $Y$?"), which is not clear for unimodal, vision foundation models. Fig.2 provides a schematic overview of the naïve structure discovery algorithm that we propose. To account for stability and reproducibility, we present different wordings to the queries (synonymous formulations) and disable parameters that induce randomness (e.g. temperature), respectively. It is important to note that the proposed naïve structure discovery procedure is not a proper induction method in the classical sense as it does not use actual data as input to perform the inferences (all the possible inferences are established upon training completion). In that sense, LLMs behave much like humans, who simply recall that "a higher altitude means a lower temperature" than to look at actual recordings of altitude and temperature (and other variables) to perform the causal inference. As anticipated, the LLM thereby also inherits natural language ambiguities. To give an example, even if the LLM is prompted with an additional "Answer with Yes or No" the LLM is not constrained to oblige. To cope with this issue, we introduce different answer types such as "Yes/No, probably", "Yes/No, indirectly", "Yes/No, other factors", "Yes/No, through explanation", "Inconclusive" and "No answer / General Statement" to classify the LLM's answers. To further ensure stability of the results, manual proof reading is conducted[6].

**Overview of Experimental Setup.** We evaluate three publicly accessible LLMs: AlephAlpha's Luminous (LLM-L; AlephAlpha (2022)), OpenAI's GPT-3 (LLM-G; Brown et al. (2020)), and Meta's OPT (LLM-O; Zhang et al. (2022b)). All models are transformer based architectures (Vaswani et al., 2017), trained at scale qualifying them as LLMs (see Bommasani et al. (2021)). Our analysis primarily investigates three different questions:

(**Q1**) In setting where the causal graph is (partially) known how do LLMs perform?
(**Q2**) In "common sense" settings like reasoning or intuitive physics how do LLMs perform?
(**Q3**) How do synonyms or more general variable name altercations affect the LLM's estimate?

### 3.1 Discussion of Results for **Q1**

In this experiment we consider publicly available data sets that propose a "ground truth" causal graph (which depicts the data generating process). We consider six data sets: altitude (A; Mooij et al.

[4]For this, access to the training data would be required. Furthermore, the sheer amount of data would require a similar compute resources as training the LLM in the first place.

[5]Which is an analogue to causal inference in general.

[6]While this required human labour is suboptimal, it poses a first step towards an automated analysis aiming at discovering the right research directions for future work.

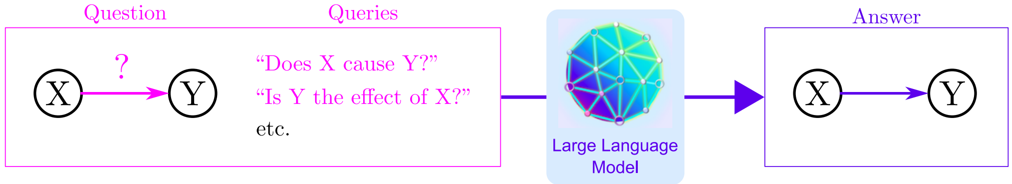



Figure 2: **Structure Discovery via Large Language Models.** Schematic overview. By iteratively querying differently worded natural language queries that aim at questioning the existence of a causal relationship for all variable combinations of interest, we construct a graph prediction from the LLM. (Best viewed in color.)

(2016)), health (H; Zečević et al. (2021)), recovery (R; Charig et al. (1986)), driving (D; synthetic), cancer (C) and earthquake (E) both (Korb & Nicholson, 2010). We use five different query wordings (or formulations), namely "Are $X$ and $Y$ causally related?", "Is there a causal connection between $X$ and $Y$?", "Is there a causality between $X$ and $Y$?", "Does $X$ cause $Y$?", and "Does $X$ influence $Y$?" of which the first three are classified as *symmetric* queries since the expected answer is a mere association $X$–$Y$ and the last two wordings classify as *asymmetric* accordingly i.e., we expect either $X \rightarrow Y$ or $X \leftarrow Y$ (in the case of an existing relation). Furthermore, we note that some of the wordings make "causal" explicit. For the different variable pairings, multiplying with the number of formulations, we have 10 questions for A, 100 for C, 60 for CH, 30 for D, 100 for E and 30 for R respectively. Three key observations were made:

**Asymmetric queries prefer unique direction.** Consider Fig.4 for the predictions of LLM-O when switching from the symmetric query (top row) to the asymmetric query (bottom row), which shows how the LLM starts settling on a unique direction (single edge) for multiple previously undecided relations, thereby, significantly improving prediction quality across all graph predictions (i.e., the false positive rate is being reduced). While this observation is consistent with the natural interpretation that an asymmetric query like "Does $X$ cause $Y$?" only accepts the answers $X$ $Y$ or $X \rightarrow Y$, the observation is still surprising as there are no formal guarantees to the query that this should be the case. It might suggest that the LLM indeed learned the difference between the two types of questions on a causal level.

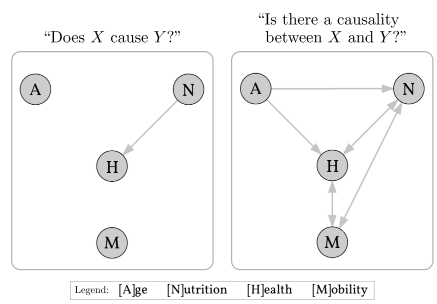



Figure 3: **LLM Sensitivity to Query Wording.** LLM naïve graph predictions for two different query wordings/formulations (LLM-G shown). A significant change in output is being observed.

**Over- or underestimation.** Comparing the predicted to the ground truth graphs reveals that the models either tend to overestimate the number of connections leaning towards a fully connected graph (LLM-L,-O), whereas others underestimate/hesitate to predict (LLM-G).

**Sensitivity to wording.** While some models remain overall stable in their prediction across data sets and wordings (LLM-L,-O), others react with unsmooth change to alternate wordings. Consider Fig.3 where a significant change in the predicted graph is observed simply by changing the query wording. A possible interpretation for this observation is that a keyword such as "causality" might be embedded further away from an alternate keyword (here for instance "cause") within the LLM's latent space, thereby, accessing (in this case) correct answers.

### 3.2 Discussion of Results for **Q2**

The key idea of this work, "correlations of causal facts", discussed in Sec.2 only works if these correlations actually exist (that is, we can expect the models to have encountered those). While we are unable to assess this rigorously, we can make the case that *common sense* reasoning tasks of either abstract nature or on an intuitive physics level (see for instance Tenenbaum et al. (2011)) are reasonable settings in which we can expect said correlations i.e., we can expect big data to cover relevant

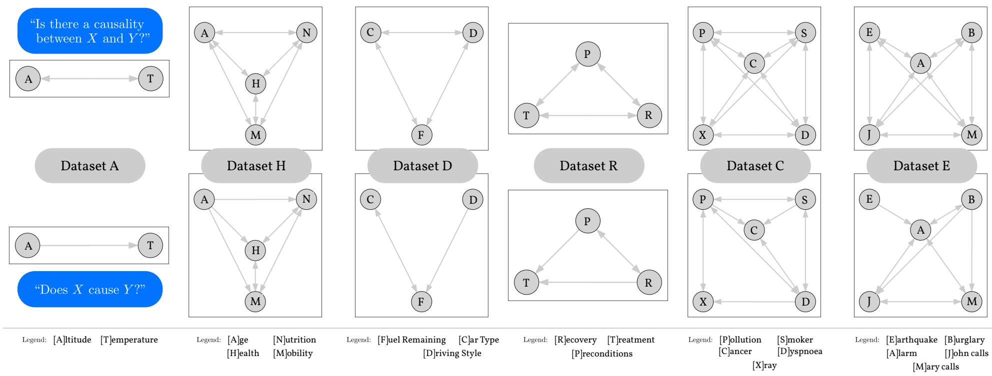



Figure 4: **Asymmetric Query Wording Implies Unidirectedness.** LLM naïve graph predictions on data sets that provide a causal graph (LLM-O is shown). Top row, predictions with a symmetric query wording, bottom row, predictions with an asymmetric query wording. Surprisingly, the LLM is capable of deciding multiple edges uniquely (and correctly) when switching to the asymmetric formulation without explicit guarantees to such behavior.

literature. In this experiment, we consider for the abstract reasoning (AR) 15 different questions (an example of a transitive[7] chain, "If $A$ causes $B$ and $B$ causes $C$ does $A$ cause $C$?") and for intuitive physics (IP) 36 questions (an example, "A ball is placed on a table and rolls off. What does this tell us about the table?"). Interestingly, both LLM-L and LLM-O either fail to provide sensible answers or provide answers that are ambiguous, for instance, the LLM might *loop* indefinitely (repeating the first predicted sentence over and over again) or it might produce a "multiple-choice quiz" like output for which it will also choose an answer itself (see appendix for details on this case). LLM-G was well behaved, providing sensible output throughout. Two key observations were made:

**Remarkable accuracy.** LLM-G was able to answer most of the queries correctly 11 correct, 3 wrong, 1 unanswered for AR, and 21 correct, 9 wrong, 6 indecisive for IP. For AR, when extending the causal $n$-chain argument (that is, "If $X_1$ causes $X_2$ ... and $X_{n-1}$ causes $X_n$") with $n = 6$ the model started to fail answering correctly. Also, replacing the variable letters with alternate letters did not harm the prediction. For IP, some of the examples are compelling such as "Mary can not move a heavy stone by herself. However, she brought a small object and a metal rod with her. How can Mary move the stone?" to which the model answers "Mary can use the metal rod as a lever to move the stone".

**Inconsistency in knowledge.** From a human perspective arguably "equivalent" situations, are not handled consistently, e.g. to the question "What is heavier: A kilogram of metal or a kilogram of feathers?" the model answers wrongly "A kilogram of metal is heavier than a kilogram of feathers" but when asked "Most people say 'A kilogram of metal is heavier than a kilogram of feathers', but in reality?" the model correctly answers "They weigh the same."

### 3.3 Discussion of Results for **Q3**

In this setting, we fix a single graph (e.g. here the graph from data set H, which involves the variables "age", "nutrition", "health", and "mobility") and alternate the variable names. We either choose words recognized as *synonyms* of the original or words that might appear in a similar context but have an identifiable difference to the original word. A single key observation was made.

**Variable renaming might cause unsmooth change.** LLM-L reacted with increased sparsity in graph prediction when changing the variable "mobility" to mean "fitness". On the other hand, LLM-O conversely reacted with decreased sparsity in graph prediction when changing the variable "age" to "aging." Arguably, the former change is more drastic than the second since fitness as a concept might refer to a superset that includes mobility but also other things like conditioning etc.,

[7] We assume the default case of transitivity but as shown by (Halpern, 2016) this not universally true.

whereas aging "just" refers to the process of increasing the age. The pattern seems overall arbitrary, but we believe the observation that "similar" words might cause drastic change is noteworthy.

# 4 EXPERIMENTAL EVIDENCE FOR CCF

We investigate to which extent LLM embed features of higher abstraction levels into their text embeddings. The simplest form of embedding might be a bag-of-words approach which simply encodes the presence of a word in a given sentence. This approach can be performed by a purely syntactical analysis of the sentence and doesn't require any learned knowledge. As an extension to this, LLMs could learn to express synonymous expressions as similar vector by encoding semantically related words in the same position of the embedding. This abstraction requires learned background knowledge about word semantics and is more abstract as it is invariant to exact sentence wording. Following this train of thought one could imagine to include highly abstract features into the embedding. The 'truthfulness' of a sentence could be one such example. Doing so requires semantic understanding of the sentence and knowledge about the real world to check whether the given statement is true or false. Embeddings that represent such high level features therefore lead to increased correlation between correct statements and in turn of CCF as a special case about causal facts.

**Why abstraction matters.** Here we want to motivate why answering causal questions using CCF requires the models to encode higher abstract features into the text embeddings. We imagine three possible cases. Case 1: A low-level abstraction encoding as a list of facts. Think of encoding the presence of individual words in a sentence. Each fact is encoded at a separate position and no positions are shared between the features. When exchanging some words for synonyms, the synonyms will be encoded at a different position in the vector. As a consequence the correlation between synonymous sentences will be close to zero although their meaning stays the same. Case 2: The LLM encodes synonymous words at the same position and as a consequence correlation between similar sentences will be high. Case 3: The model embeds the abstract concept of 'truth of a statement' into the embedding vector. As a consequence all true statements will feature a higher correlation. This extends especially to the causal case, where we can query the believed truthfulness of a causal statement by comparing it to a prototypical 'truth' vector (e.g. average over embeddings of true sentences).

To tell apart true from false causal statements requires both, text understanding and knowledge about the real world. To test the abstraction level of the embedded features we investigate to which level LLMs might encode abstract features in their embeddings. For querying the embeddings we use OpenAI GPT-3 model ('text-similarity-davinci-001') which returns an embedding vector of 12,288 dimensions.

**Prototype prediction.** As a result of the previous discussion we investigate whether it is possible to extract vectors that reflect a prototypical truth/falsity embedding from the models by averaging multiple embeddings from sentences stemming from different domains. By varying the domain we reduce the bias causes by any such single domain to end up with a prototype of truthfulness. Now, if the LLM indeed encodes believed truth about the given statement at a specific position in its embedding space (meaning in one or more embedding locations), those specific positions will be a high in the prototype vector. We query the embeddings of true causal relations (Appendix A.2) for all of our sentence templates (Appendix A.1) resulting in 50 embeddings of causally true or false sentences. To give an example, we compare the to the embedding of say "Altitude causes tempature" against both prototype vectors, and choose the existence of an edge based on the cosine similarity:

$$\text{decision}(\boldsymbol{E}) = \begin{cases} \text{True} & \text{if cosineSimilarity}(\boldsymbol{E}, \mathbf{PT}_{\boldsymbol{T}}) \geq \text{cosineSimilarity}(\boldsymbol{E}, \mathbf{PT}_{\boldsymbol{F}}) \\ \text{False} & \text{otherwise} \end{cases},$$

where $\boldsymbol{E}$ is the embedding vector of the statement to test, $\mathbf{PT}_{\boldsymbol{T},\boldsymbol{F}}$ are the prototype vectors over the known true/false causal statements. The cosine similarity is defined as:

$$\text{cosineSimilarity}(\boldsymbol{E_1}, \boldsymbol{E_2}) := \frac{\boldsymbol{E_1} \cdot \boldsymbol{E_2}}{\|\boldsymbol{E_1}\| \|\boldsymbol{E_2}\|}.$$

We can use cosine similarity as a measure of (Pearson) correlation since they are equal up to mean centered versions of $\boldsymbol{E}_i$ and since the means are identical in this case, our cosine similarity is guaranteed to be proportional to correlation. Figure 5 shows the resulting graphs for the CH dataset

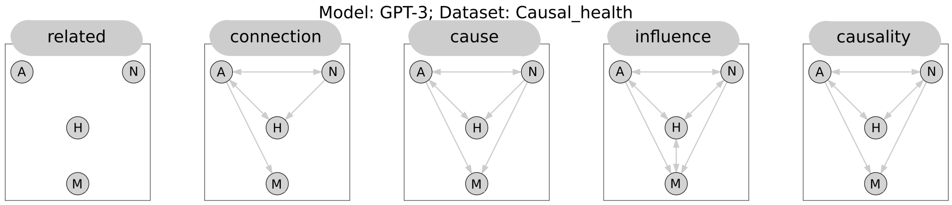



Figure 5: **Causal structure prediction using prototype vectors.** Prototype vectors are computed from the average embeddings of a set of correct (or incorrect) causal statements. The embeddings over the individual edges (e.g."Altitude causes temperature.") are compared to the 'true' or 'false' prototype vectors. The presence of an edge is decided based on whether its embedding correlates stronger with the 'true' or 'false' prototype. The remaining dataset plots are presented in the Appendix A.3.

queried with different statement templates. Results for the remaining datasets can be found in the Appendix A.3. Like in the previous experiments the number of predicted edges varies strongly depending with the used sentence template and the domain of the dataset. For example for the E and R datasets no edges can be predicted. For the A dataset prediction results for three out of the five are correct. For C, CH and D results are mixed, with an overall tendency to decide more edges on one or the alternate direction.

Given the varying prediction quality, ranging from robust prediction of e.g. altitude $\rightarrow$ temperature or smoking $\rightarrow$ cancer edges to almost no edge predictions in the E and R datasets, we conclude that LLMs do not seem to encode a general truth vector in the embeddings.

**Naïve Testing of CCF.** Given that we can not find a general truth vector, we examine our CCF hypothesis by pairing embeddings of causal questions and their corresponding answer statements. That is, we examine the correlation of a questions's embedding "Does A cause B?" (and the inverse "Does B cause A?") to the statements "A causes B." and "B causes A". Figure A.4 shows our findings. In particular the column to the right indicates whether the causal question "A $\rightarrow$ B?" correlates stronger to its equivalent statement A $\rightarrow$ B or to its inverted statement B $\rightarrow$ A. The results indicate that, regardless of asking for A $\rightarrow$ B or B $\rightarrow$ A, the questions almost always correlate strongest with their corresponding, syntactically closest statement. This indicates, that the encoded lower level features like the syntactic structure dominate over the semantic meaning of the sentence in the embedding encoding.

**Alternate Testing of CCF.** While we were not able to prove/falsify CCF in the previous setting, we are next interested in whether LLM embeddings are only encoding the syntactic sentence structure or if they include some sort of understanding. Therefore, we look at simple yes/no answer correlations such as "Does altitude cause temperature?" $\rightarrow$ "Yes.". Results are presented in section A.5 of the appendix. While we find no consistent pattern between (in)correct questions and the (dis)agreeing answer statements, we observe a *strong correlation* when switching from positive answer "Yes" to its negative counterpart "No". While the average correlation of all answer statements varies to some degree when switching from a positive to negative answer, the correlation is consistently strongest with the "There is a causal relation." answer, which specifically mentions 'causality'—which is the domain of our questions! Interestingly, this strong correlation even holds to for the 'Does A influence B?' question template which contains no word related to 'causality'. We conclude that LLMs are aware of the domain they are tasked to operate in. While no answers to the posed questions can be inferred from the question-answer pairs, we observe a strong correlation when both, the question and answer embeddings which talk about causal relations. Therefore, being in favor of CCF.

## 5 Conclusive Discussion

While this systematic analysis of LLMs incorporated validation strategies to ensure robustness of the results and measurements of correlations within the embedding space of LLMs, much of the presented analysis still depends on thorough manual labor and taking scientific/educated guesses in hope of arriving at the correct interpretation to reach sound conclusions. Therefore, also the following discussion is based on such an informal procedure based on the collected evidence.

As we started exploring in Sec.2, our physical reality creates the model/graph descriptions (which we called *causal assumptions*) and corresponding questions we could answer regarding the underlying variables, separately. That is, the graph that captures the idea of "altitude causes temperature" and a related scientific question like "Does altitude cause temperature?" have are confounded but not causally linked. In our reality, the given example is the truth, that is, there exists a *physical mechanism* that decreases temperature with increasing altitude. Therefore, we expect to see a correlation in the number of times each of those descriptions appears e.g. in some standard literature on physics or in articles that discuss global warming. Obviously, changing the description of either by intervention would not change the other, giving further reason to believe that there is no direct causation. Subsequently, changing our actual physical reality (such that temperature were counterfactually to actually cause altitude) would create an interventional distribution (on the aforementioned descriptions) that change the correlation towards this alternate pair. The LLM learns any of those correlations, and since in our physical reality the former is true, this causal relation can be expected to be learned by the LLM. This was the key idea behind the discussion of the "correlations of causal facts" that we proposed. In classical causality literature our data usually expresses low-level (physical) quantities and what makes the model causal are actually the causal assumptions e.g. a Causal Bayesian Network (CBN; see Pearl (2009)) assumes a certain graphical structure where the edges denote causation. However, there is no restriction on what the variables might denote. We might have a "big" SCM (that might be considered as *nature itself*[8]) and it generates other SCMs so to say i.e., the data talks about causal assumptions (*meta-level* abstractions, see Fig.1). We can make the CCF idea even agree with classical causality when we view the data as a collection of experimental data that is now simply being provided as an observation. The LLM obviously does not use any causal assumptions explicitly, and the CHT (recall discussion in Sec.1, Bommasani et al. (2021)) restricts causal inference from observational data, making us ultimately believe that the LLM is not causal. However, since the correlation is on data that talks about causality, could the LLM in fact be causal in some other (implicit) notion? This was the key question of this paper. However, the presented analysis is in support of the fact that there is "something causal" going on implicitly, which might be the key reason for the difficulty of resolving the ongoing heated debate in absolute terms. These LLM might only be somewhat "smart dictionaries", but for a downstream task that does not involve generalization/transfer capabilities, whether the model truly "understands" the causality of the problem or whether it just "knows" it seems to be irrelevant. This observation is reminiscent of the philosophical arguments given by (Searle, 2009). *Testing for real understanding would require to query explanations from the models.* This would allow us to test whether they accurately capture the underlying causal connections or just memorize inseparable bits of information. For the latter case being incapable of linking those bits together into consistent causal chains. We observe these shortcomings of LLMs in the abstract reasoning setting where they are only able to correctly answer for standard causal chains of alphabetically ordered nodes, but fail for deviating setups. This observation also agrees with the logical perspective provided by (Zhang et al., 2022a). One could summarize said argument's conclusion as capturing the inadequacy of the "Turing Test", that is, programming a digital computer may make it *only appear* to understand language.

**Takeaway and Societal Implications.** We believe the take-away message of this humble, initial effort in hope of a resolution of the debate is two-fold. The negative message is that we can *not* rely solely on LLMs as we cannot expect a generalization (in causal terms, and further the implicit nature of the causal assumptions consumed by the LLM raises other issues of trustworthiness etc. all discussed within explainable AI). Further, current LLMs are unable to process actual data observations to ground the available evidence for doing inference like classical (causal) structure discovery methods do. However, the positive message is that we can use the LLMs as a head start to learning and inference which thereby helps in developing new methods for more robust inference. In that sense, they might very well serve as stepping stones towards progress in AI/ML research. Also, our analysis suggests that they are rather good with common sense knowledge, that is, training data of the LLM where CCF holds. Since LLMs and generally foundation models have spread at an incredible rate not just through the AI/ML community but also to industry and laymen, it is important to discuss safety critical settings but also more generally ethical aspects relating to biases. We hope that our work makes a positive contribution to this by providing an increased understanding of what these models are and are not capable of.

[8]This idea might also be linked to the concept of a *Universal* Turing Machine (Turing et al., 1936).

# References

John Aldrich. Correlations genuine and spurious in pearson and yule. *Statistical science*, pp. 364–376, 1995.

AlephAlpha. Luminous language model. 2022.

Elias Bareinboim, Juan D Correa, Duligur Ibeling, and Thomas Icard. 1on pearl's hierarchy and. Technical report, Technical Report, 2020.

J Mark Bishop. Artificial intelligence is stupid and causal reasoning will not fix it. *Frontiers in Psychology*, 11:2603, 2021.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Gwern Branwen. The scaling hypothesis. *gwern.net*, 2020.

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

Clive R Charig, David R Webb, Stephen Richard Payne, and John E Wickham. Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithotomy, and extracorporeal shockwave lithotripsy. *Br Med J (Clin Res Ed)*, 1986.

Jean-Marie Chauvet. The 30-year cycle in the ai debate. *arXiv preprint arXiv:1810.04053*, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Hector Geffner, Rina Dechter, and Joseph Y Halpern. Probabilistic and causal inference: The works of judea pearl, 2022.

Kurt Gödel. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme i. *Monatshefte für mathematik und physik*, 38(1):173–198, 1931.

Joseph Y Halpern. *Actual causality*. MiT Press, 2016.

Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 109 (Supplement 1):10661–10668, 2012.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 1986.

Dominik Janzing and Bernhard Schölkopf. Detecting non-causal artifacts in multivariate linear regression models. In *International Conference on Machine Learning*, pp. 2245–2253. PMLR, 2018.

Vivek Khetan, Roshni Ramnani, Mayuresh Anand, Subhashis Sengupta, and Andrew E Fano. Causal bert: Language models for causality detection between events expressed in text. In *Intelligent Computing*, pp. 965–980. Springer, 2022.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. CRC press, 2010.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4164–4172, 2019.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020.

Gary Marcus. Deep learning is hitting a wall. *Nautilus, Accessed*, pp. 03–11, 2022.

Gary Marcus and Ernest Davis. Has ai found a new foundation? *The Gradient*, 2021.

Pierre Maurage, Alexandre Heeren, and Mauro Pesenti. Does chocolate consumption really boost nobel award chances? the peril of over-interpreting correlations in health studies. *The Journal of Nutrition*, 143(6):931–933, 2013.

Emily McMilin. Selection bias induced spurious correlations in large language models. *arXiv preprint arXiv:2207.08982*, 2022.

Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pp. 8821–8831. PMLR, 2021.

Hans Reichenbach. *The direction of time*, volume 65. Univ of California Press, 1956.

Bernhard Schölkopf. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

John Searle. Chinese room argument. *Scholarpedia*, 4(8):3100, 2009.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021.

Richard Sutton. The bitter lesson. *Incomplete Ideas (blog)*, 13:12, 2019.

Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. Commonsenseqa 2.0: Exposing the limits of ai through gamification. *arXiv preprint arXiv:2201.05320*, 2022.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

Alan Mathison Turing et al. On computable numbers, with an application to the entscheidungsproblem. *J. of Math*, 58(345-363):5, 1936.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Xingqiao Wang, Xiaowei Xu, Weida Tong, Ruth Roberts, and Zhichao Liu. Inferbert: A transformer-based causal inference framework for enhancing pharmacovigilance. *Frontiers in Artificial Intelligence*, 4, 2021.

Matej Zečević, Devendra Dhami, Athresh Karanam, Sriraam Natarajan, and Kristian Kersting. Interventional sum-product networks: Causal inference with tractable probabilistic models. *Advances in Neural Information Processing Systems*, 34, 2021.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502*, 2022a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022b.

# A Appendix to "Probing for Correlations of Causal Facts: Large Language Models and Causality"

This supplementary material provides plots that were out of scope in terms of presentation for the main paper but which the reader might find interesting in addition to technical details of the conducted experimental analysis.

**Technical details.** Our method was executed on one NVIDIA A100-SXM4-80GB GPU with 80 GB of RAM and it takes 10 GPU minutes to query the OPT model. For the Luminous and GPT-3, we use the provided APIs.

Fig.6 presents stability results on the LLM predictions upon querying with different formulations, Fig.7 discusses stability results for different variable namings, and Fig.8 presents all single graph predictions.

Subsections A.6 and A.7 contain the LLM answers to the Intuitive Physics and Causal Chain questions. While querying the foundation models we observed two reoccurring behaviours. First, the models tend to produce multiple-choice-style answers of the form: "A: ..., B: ..., C: ...". Additionally, we observe that the models tend to start repeating sentences. To improve readability, we manually formatted the presented answers by adding/removing line breaks, white spaces and punctuation. We also shortened the texts in case of the models starting to repeat sentences. The exact response texts are available within the code repository.

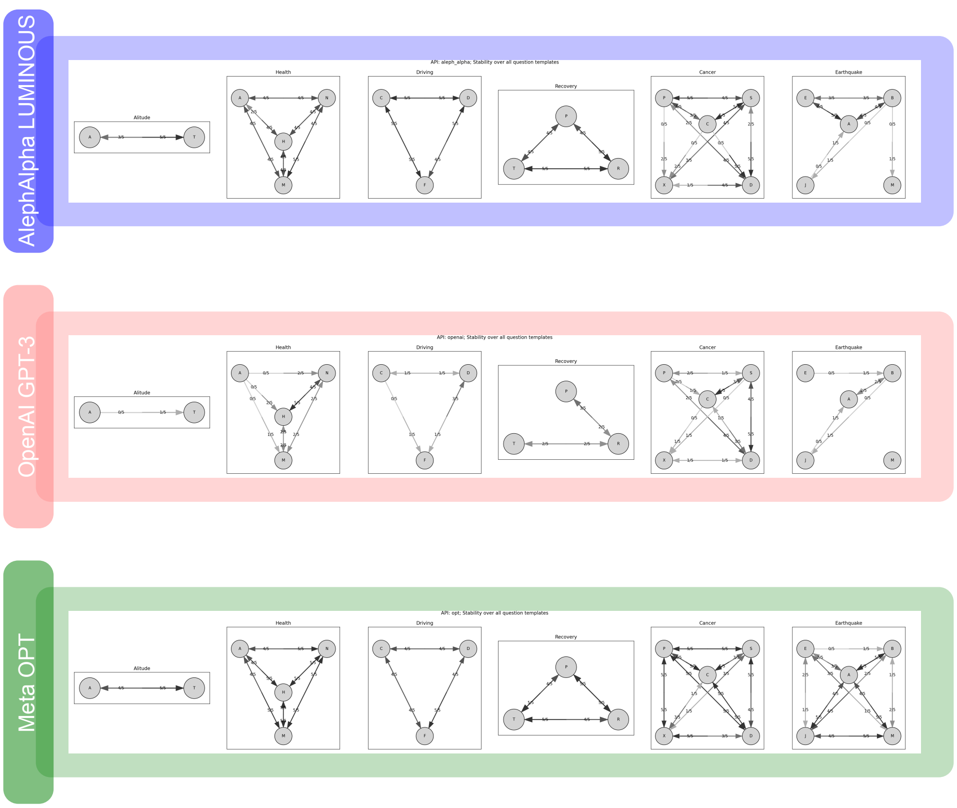



Figure 6: **Stability: Different Query Formulations.** Graph predictions for the three different language LLMs, across six different data sets with known graphs for five different wordings/formulations of a given query. Darker arrows indicate a more frequent prediction of a connection across the different formulations. LLM-L and LLM-O tend to predict dense graphs, whereas LLM-G tends to predict sparse graphs. (Best viewed in color.)

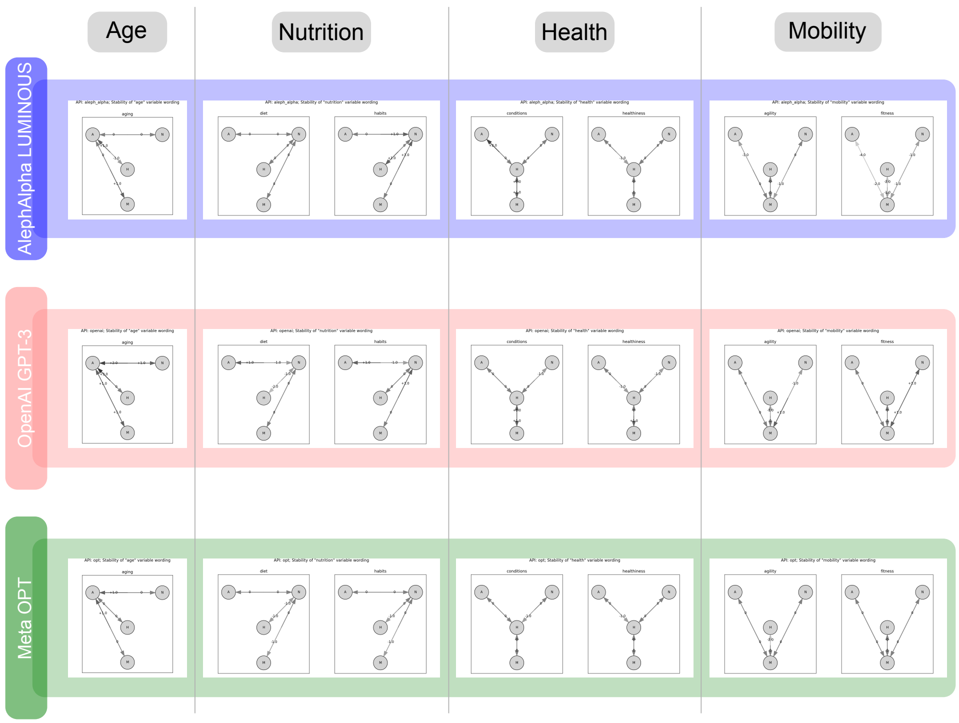



Figure 7: **Stability Experiment: Synonyms and General Variable Name Altercations.** We fix a single data set (here, data set H) and change the formulation of the variable name to either a synonym or a more distant reformulation. Only the edges related to a given node of interest are presented. Some LLMs react sensitive to certain wording changes. For a discussion see **Q2** of the main paper. (Best viewed in color.)

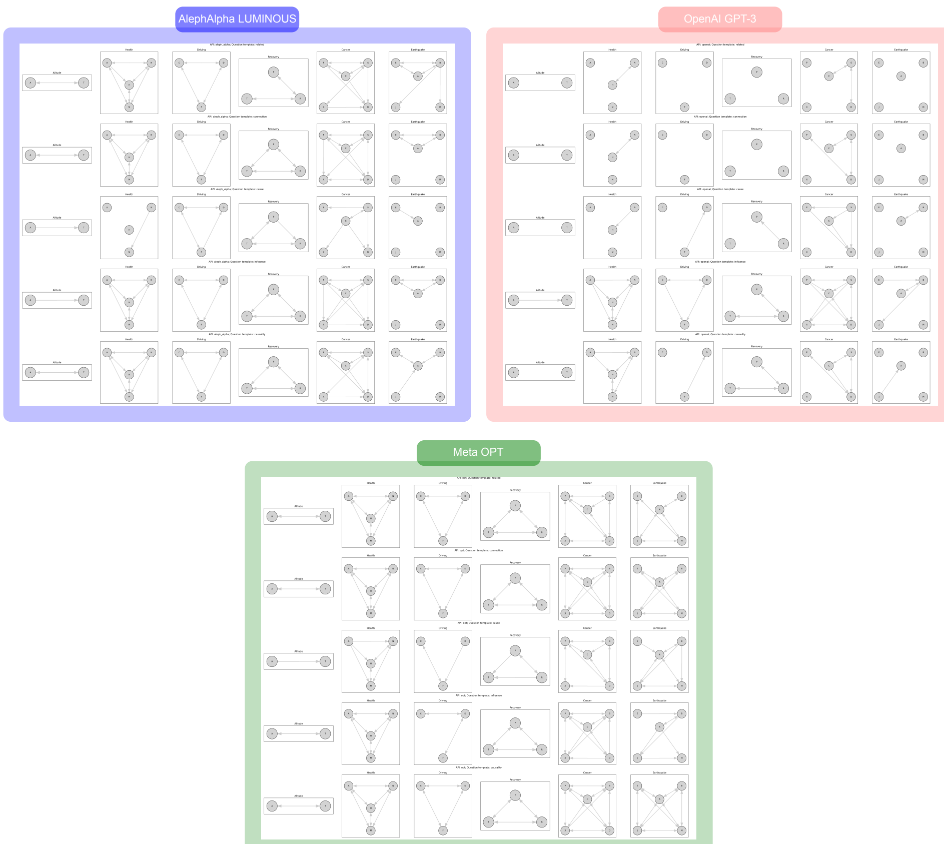



Figure 8: **LLM Graph Prediction.** All separate graph predictions using the naïve LLM structure discovery procedure. Six different data sets, five different formulations, three different LLM. For a discussion reconsider **Q1** in the main paper. (Best viewed in color.)

### A.1 QUESTION-ANSWER TEMPLATES

This section contains the question and answer templates used in the experiments of this paper. The templates are adjusting for plural forms ("Burglaries *cause* alarms.") and articles ("The *the* causes *the* diving style.")

| **Question templates** | |
|---|---|
| **Abbreviation** | **Template text** |
| related | Are A and B causally related? |
| connection | Is there a causal connection between A and B? |
| cause | Does A cause B? |
| influence | Does A influence B? |
| causality | Is there causality between A and B? |

| **Answer templates** | |
|---|---|
| **Abbreviation** | **Template text** |
| related | A and B are causally related. |
| connection | There is a causal connection between A and B. |
| cause | A causes B. |
| influence | A influences B. |
| causality | There is causality between A and B. |

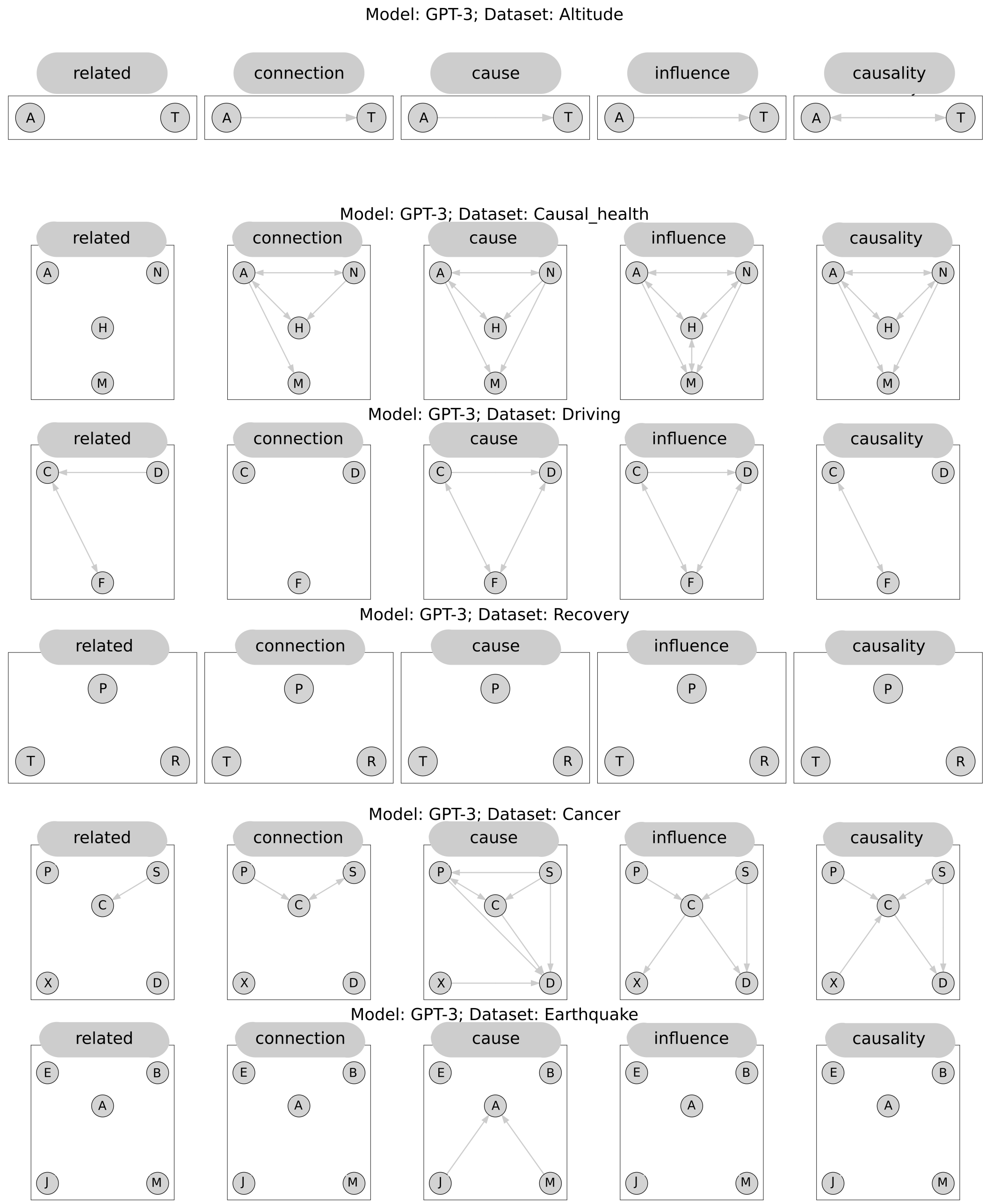
### A.2 CORRECT CAUSAL STATEMENTS

Below follows the list of ten correct causal relations used for computing prototype vectors. The variables are marked in italics, the sentence structure is varied with the answer templates listed in Section A.1. False causal statements are created by querying for the anti-causal direction.

- A monument causes visitors.
- A sunny day causes happiness.
- Heat causes melting ice.
- Source material causes the object‘s weight.
- A doctor's visit causes diagnoses.
- A cold causes coughing.
- Traffic causes air pollution.
- Holidays causes flights.
- Increasing prices causes lowered demand.
- A strong economy causes wealth of a country.

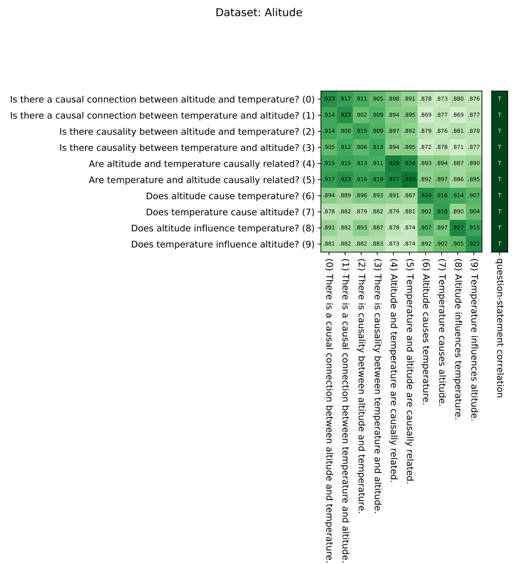
### A.3 PROTOTYPE GRAPHS

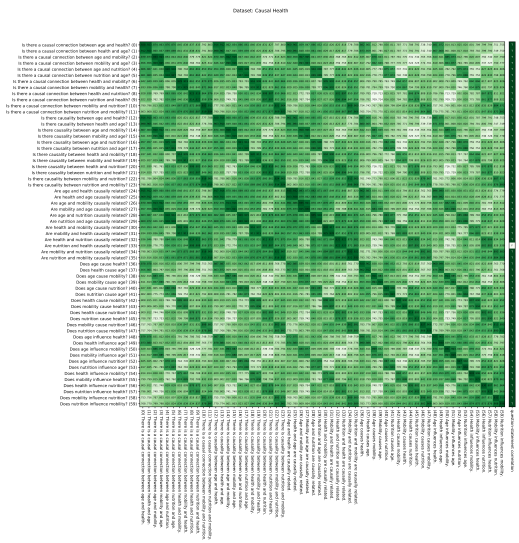
Graphs generated from comparisons to prototype embeddings. Causal 'truth' prototype vectors yield features that are correlated with the high level concept of *causal truth*. By averaging embeddings of known causally correct facts, we hope to zero out vector features that encode e.g. the domain of a sentence (which will vary over the set of sentence embeddings) and keep positive values the locations that encode causal truth (which is present in all sentences). A description of the graph inference procedure is described in the main paper (Section 4).

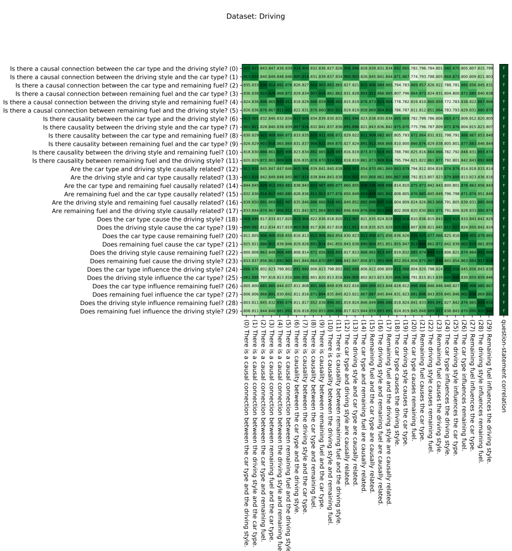


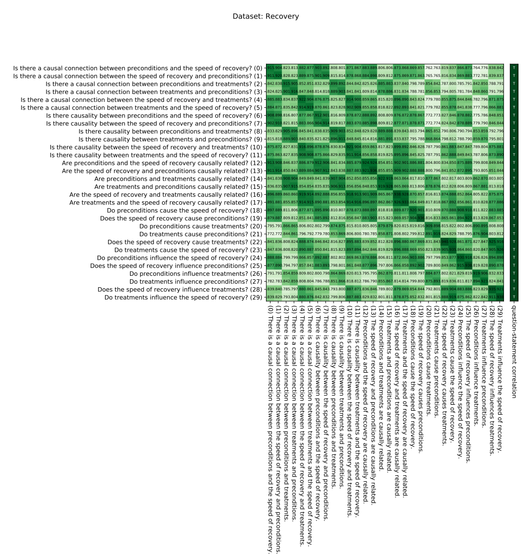
### A.4 Question-Answer Statement Correlation

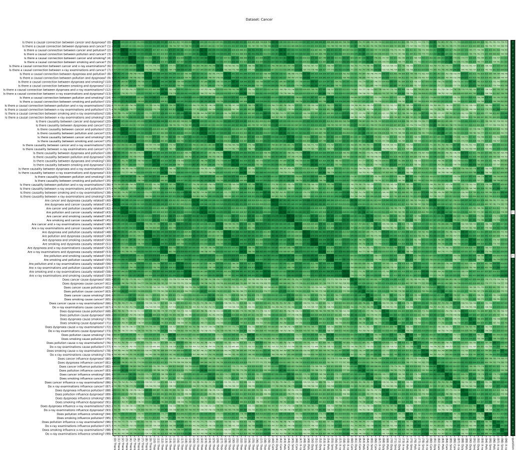
This section contains similarity matrices of question and answer statement embeddings using cosine similarity. Darker color indicates a stronger correlation. The right column indicates whether a question, e.g. "Does A cause B?", correlates stronger to its equivalent statement, "A causes B". Correlations of these pairings are placed on the main diagonal. Pairings of a question and its anti-causal answer "Does A cause B?"-"B causes A" are placed alternating left and right to the main diagonal. The column to the right of the matrix indicates whether a question correlates strong with its causally equivalent answer (colored green) or its anti-causal answer (indicated by white color).

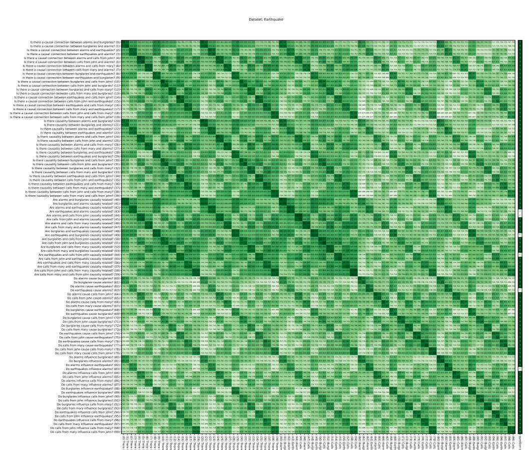

Dataset: Altitude
Is there a causal connection between altitude and temperature? (0)
Is there a causal connection between temperature and altitude? (1)
Is there causality between altitude and temperature? (2)
Is there causality between temperature and altitude? (3)
Are altitude and temperature causally related? (4)
Are temperature and altitude causally related? (5)
Does altitude cause temperature? (6)
Does temperature cause altitude? (7)
Does altitude influence temperature? (8)
Does temperature influence altitude? (9)
.923 .917 .911 .905 .898 .891 .878 .873 .880 .876
.914 .923 .902 .909 .894 .895 .869 .877 .869 .877
.914 .908 .915 .909 .897 .892 .879 .876 .881 .878
.905 .912 .906 .913 .894 .895 .872 .878 .871 .877
.915 .915 .913 .911 .929 .924 .893 .894 .887 .890
.917 .923 .916 .919 .927 .933 .892 .897 .886 .895
.894 .889 .896 .893 .891 .887 .924 .916 .914 .907
.878 .882 .879 .882 .879 .881 .902 .918 .890 .904
.891 .882 .893 .887 .878 .874 .907 .897 .927 .915
.881 .882 .882 .883 .873 .874 .892 .902 .905 .922
T
(0) There is a causal connection between altitude and temperature.
(1) There is a causal connection between temperature and altitude.
(2) There is causality between altitude and temperature.
(3) There is causality between temperature and altitude.
(4) Altitude and temperature are causally related.
(5) Temperature and altitude are causally related.
(6) Altitude causes temperature.
(7) Temperature causes altitude.
(8) Altitude influences temperature.
(9) Temperature influences altitude.
question-statement correlation


Dataset: Causal Health


Dataset: Driving

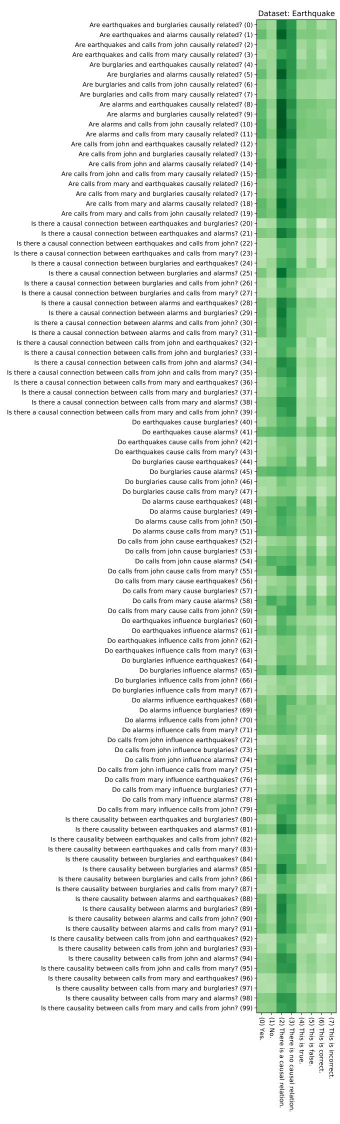

Dataset: Recovery

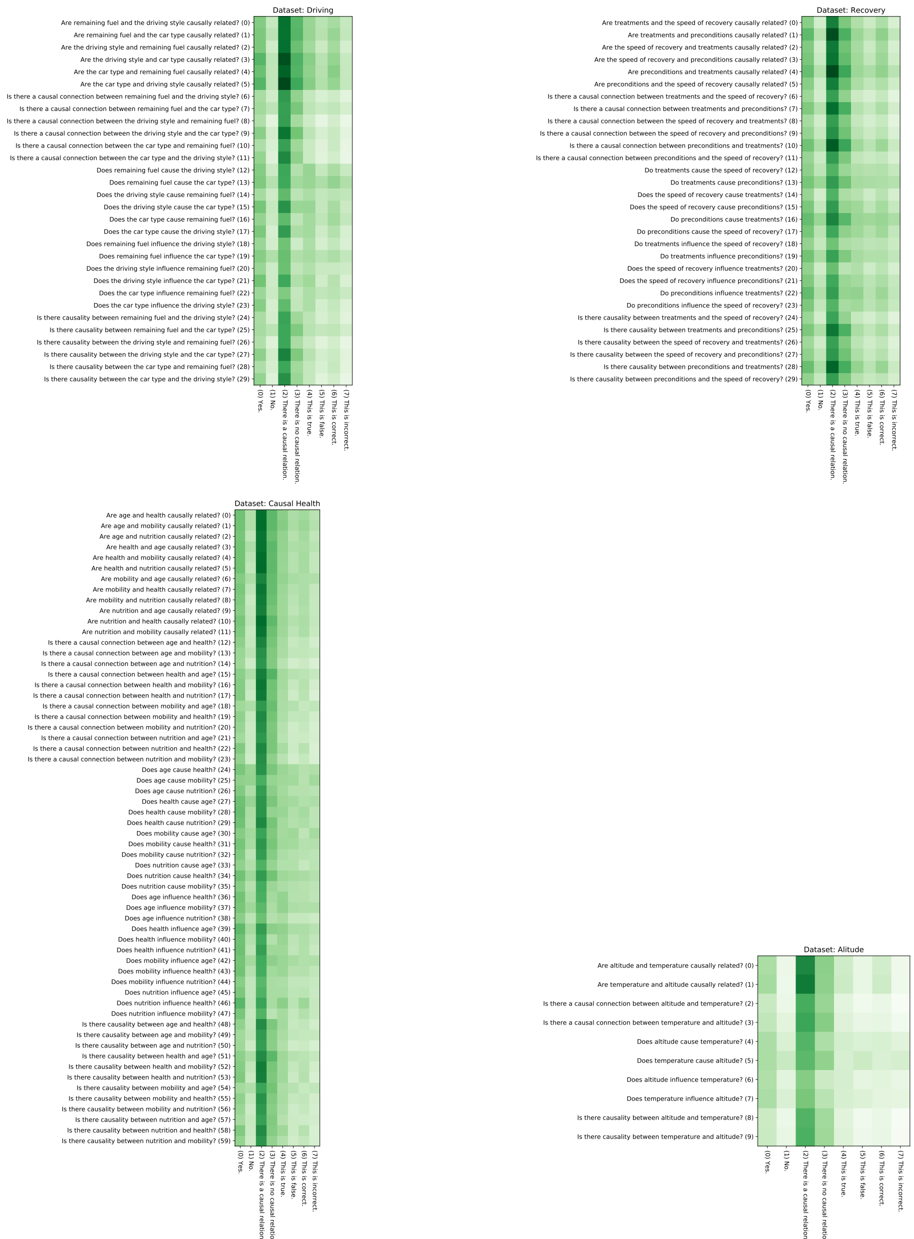

Dataset: Cancer


Dataset: Earthquake

### A.5 Question-Answer Correlation

This section contains similarity matrices of question and answer embeddings using cosine similarity. Darker color indicates a stronger correlation. Every cell in the matrix indicates the embedding correlation between a question, e.g. “Does Altitude cause Temperature?” and one of the following answer embeddings: “Yes”, “No”, “There is a causal relation”, “There is no causal relation”, “This is true”, “This is false”, “This is correct”, “This is incorrect”.





Dataset: Driving
Are remaining fuel and the driving style causally related? (0)
Are remaining fuel and the car type causally related? (1)
Are the driving style and remaining fuel causally related? (2)
Are the driving style and car type causally related? (3)
Are the car type and remaining fuel causally related? (4)
Are the car type and driving style causally related? (5)
Is there a causal connection between remaining fuel and the driving style? (6)
Is there a causal connection between remaining fuel and the car type? (7)
Is there a causal connection between the driving style and remaining fuel? (8)
Is there a causal connection between the driving style and the car type? (9)
Is there a causal connection between the car type and remaining fuel? (10)
Is there a causal connection between the car type and the driving style? (11)
Does remaining fuel cause the driving style? (12)
Does remaining fuel cause the car type? (13)
Does the driving style cause remaining fuel? (14)
Does the driving style cause the car type? (15)
Does the car type cause remaining fuel? (16)
Does the car type cause the driving style? (17)
Does remaining fuel influence the driving style? (18)
Does remaining fuel influence the car type? (19)
Does the driving style influence remaining fuel? (20)
Does the driving style influence the car type? (21)
Does the car type influence remaining fuel? (22)
Does the car type influence the driving style? (23)
Is there causality between remaining fuel and the driving style? (24)
Is there causality between remaining fuel and the car type? (25)
Is there causality between the driving style and remaining fuel? (26)
Is there causality between the driving style and the car type? (27)
Is there causality between the car type and remaining fuel? (28)
Is there causality between the car type and the driving style? (29)
(0) Yes.
(1) No.
(2) There is a causal relation.
(3) There is no causal relation.
(4) This is true.
(5) This is false.
(6) This is correct.
(7) This is incorrect.
Dataset: Recovery
Are treatments and the speed of recovery causally related? (0)
Are treatments and preconditions causally related? (1)
Are the speed of recovery and treatments causally related? (2)
Are the speed of recovery and preconditions causally related? (3)
Are preconditions and treatments causally related? (4)
Are preconditions and the speed of recovery causally related? (5)
Is there a causal connection between treatments and the speed of recovery? (6)
Is there a causal connection between treatments and preconditions? (7)
Is there a causal connection between the speed of recovery and treatments? (8)
Is there a causal connection between the speed of recovery and preconditions? (9)
Is there a causal connection between preconditions and treatments? (10)
Is there a causal connection between preconditions and the speed of recovery? (11)
Do treatments cause the speed of recovery? (12)
Do treatments cause preconditions? (13)
Does the speed of recovery cause treatments? (14)
Does the speed of recovery cause preconditions? (15)
Do preconditions cause treatments? (16)
Do preconditions cause the speed of recovery? (17)
Do treatments influence the speed of recovery? (18)
Do treatments influence preconditions? (19)
Does the speed of recovery influence treatments? (20)
Does the speed of recovery influence preconditions? (21)
Do preconditions influence treatments? (22)
Do preconditions influence the speed of recovery? (23)
Is there causality between treatments and the speed of recovery? (24)
Is there causality between treatments and preconditions? (25)
Is there causality between the speed of recovery and treatments? (26)
Is there causality between the speed of recovery and preconditions? (27)
Is there causality between preconditions and treatments? (28)
Is there causality between preconditions and the speed of recovery? (29)
(0) Yes.
(1) No.
(2) There is a causal relation.
(3) There is no causal relation.
(4) This is true.
(5) This is false.
(6) This is correct.
(7) This is incorrect.
Dataset: Causal Health
Are age and health causally related? (0)
Are age and mobility causally related? (1)
Are age and nutrition causally related? (2)
Are health and age causally related? (3)
Are health and mobility causally related? (4)
Are health and nutrition causally related? (5)
Are mobility and age causally related? (6)
Are mobility and health causally related? (7)
Are mobility and nutrition causally related? (8)
Are nutrition and age causally related? (9)
Are nutrition and health causally related? (10)
Are nutrition and mobility causally related? (11)
Is there a causal connection between age and health? (12)
Is there a causal connection between age and mobility? (13)
Is there a causal connection between age and nutrition? (14)
Is there a causal connection between health and age? (15)
Is there a causal connection between health and mobility? (16)
Is there a causal connection between health and nutrition? (17)
Is there a causal connection between mobility and age? (18)
Is there a causal connection between mobility and health? (19)
Is there a causal connection between mobility and nutrition? (20)
Is there a causal connection between nutrition and age? (21)
Is there a causal connection between nutrition and health? (22)
Is there a causal connection between nutrition and mobility? (23)
Does age cause health? (24)
Does age cause mobility? (25)
Does age cause nutrition? (26)
Does health cause age? (27)
Does health cause mobility? (28)
Does health cause nutrition? (29)
Does mobility cause age? (30)
Does mobility cause health? (31)
Does mobility cause nutrition? (32)
Does nutrition cause age? (33)
Does nutrition cause health? (34)
Does nutrition cause mobility? (35)
Does age influence health? (36)
Does age influence mobility? (37)
Does age influence nutrition? (38)
Does health influence age? (39)
Does health influence mobility? (40)
Does health influence nutrition? (41)
Does mobility influence age? (42)
Does mobility influence health? (43)
Does mobility influence nutrition? (44)
Does nutrition influence age? (45)
Does nutrition influence health? (46)
Does nutrition influence mobility? (47)
Is there causality between age and health? (48)
Is there causality between age and mobility? (49)
Is there causality between age and nutrition? (50)
Is there causality between health and age? (51)
Is there causality between health and mobility? (52)
Is there causality between health and nutrition? (53)
Is there causality between mobility and age? (54)
Is there causality between mobility and health? (55)
Is there causality between mobility and nutrition? (56)
Is there causality between nutrition and age? (57)
Is there causality between nutrition and health? (58)
Is there causality between nutrition and mobility? (59)
(0) Yes.
(1) No.
(2) There is a causal relation.
(3) There is no causal relation.
(4) This is true.
(5) This is false.
(6) This is correct.
(7) This is incorrect.
Dataset: Altitude
Are altitude and temperature causally related? (0)
Are temperature and altitude causally related? (1)
Is there a causal connection between altitude and temperature? (2)
Is there a causal connection between temperature and altitude? (3)
Does altitude cause temperature? (4)
Does temperature cause altitude? (5)
Does altitude influence temperature? (6)
Does temperature influence altitude? (7)
Is there causality between altitude and temperature? (8)
Is there causality between temperature and altitude? (9)
(0) Yes.
(1) No.
(2) There is a causal relation.
(3) There is no causal relation.
(4) This is true.
(5) This is false.
(6) This is correct.
(7) This is incorrect.

### A.6 Intuitive Physics Questions

| There is a tilted board above a bucket. Where does a ball end up if it is placed on the board? | |
|---|---|
| GPT-3 | The ball will end up in the bucket. |
| Luminous | A: The ball will roll down the board. |
| OPT | A. It will fall into the bucket.<br>B. It will fall into the bucket.<br>C. It will fall into the bucket. |

| There is a tilted board above a bucket. Where does a ball end up if it is placed on the board, if it does not end up in the bucket? | |
|---|---|
| GPT-3 | The ball will end up on the ground. |
| Luminous | A ball is placed on a tilted board. Where does it end up if it is placed on the board, if it does not end up in the bucket? |
| OPT | A. It will fall into the bucket.<br>B. It will fall into the bucket.<br>C. It will fall into the bucket. |

| There is a board above a bucket. Where does a ball end up if it is placed on the board? | |
|---|---|
| GPT-3 | The ball will end up in the bucket. |
| Luminous | A ball is placed on a table. Where does it end up if it is rolled under a table? |
| OPT | A. It falls into the bucket.<br>B. It falls into the bucket and bounces off the board.<br>C. It falls into the bucket and bounces off the board and into the bucket below. |

| There is a leveled board above a bucket. Where does a ball end up if it is placed on the board? | |
|---|---|
| GPT-3 | The ball will end up in the bucket. |
| Luminous | A: The ball will roll down the board and into the bucket. |
| OPT | A. It falls into the bucket.<br>B. It falls into the bucket and bounces off the board.<br>C. It falls into the bucket and bounces off the board and into the bucket below. |

| A ball is placed on a table and rolls off. What does this tell us about the table? | |
|---|---|
| GPT-3 | The table is not level. |
| Luminous | A. The table is flat.<br>B. The table is round.<br>C. The table is square.<br>D. The table is rectangular.<br>Answer: C<br>Explanation: The table is rectangular |
| OPT | A ball is placed on a table and rolls off. What does this tell us about the table? |

| A ball is placed on a surface and rolls off. What does this tell us about the surface? | |
|---|---|
| GPT-3 | The surface is not level. |
| Luminous | A ball is placed on a surface and rolls off. What does this tell us about the surface? |
| OPT | A ball is placed on a surface and rolls off. What does this tell us about the surface? |

| A ball is placed on a table and rolls to the center. What does this tell us about the table? | |
|---|---|
| GPT-3 | The table is level. |
| Luminous | A ball is placed on a table and rolls to the center. What does this tell us about the table? |
| OPT | A ball is placed on a table and rolls to the center. What does this tell us about the table? |

| A ball is placed on a surface and rolls to the center. What does this tell us about the surface? | |
|---|---|
| GPT-3 | The surface is curved. |
| Luminous | A ball is placed on a surface and rolls to the center. What does this tell us about the surface? |
| OPT | A ball is placed on a surface and rolls to the center. What does this tell us about the surface? |

**Remaining Intuitive Physics Results.** Can be found here: https://anonymous.4open.science/r/LLMsAndCausality-E220/media/Intuitive-Physics.pdf

### A.7 Causal Chains Questions

| If A causes B and B causes C. Does A cause C? | |
|---|---|
| GPT-3 | It is possible that A causes C, but it is also possible that A and C are unrelated. |
| Luminous | A: The answer is no.<br>The cause of a thing is the thing that makes it happen. |
| OPT | If A causes B and B causes C, then A causes C.<br>If A causes B and B causes C, then A does not cause C.<br>If A causes B and B does not cause C, then A does not cause C. |

| If A causes B and B causes C. Does A cause B? | |
|---|---|
| GPT-3 | There is not enough information to answer this question. |
| Luminous | A: The answer is no.<br>The cause of a thing is the thing that makes it happen. |
| OPT | If A causes B and B causes C, then A causes C.<br>If A causes B and B does not cause C, then A does not cause C. |

| If A causes B and B causes C. Does B cause C? | |
|---|---|
| GPT-3 | Yes, B causes C. |
| Luminous | A: The answer is no.<br>The cause of a thing is the thing that makes it happen. |
| OPT | If A causes B and B causes C, then A causes C.<br>If A causes B and B causes C, then A does not cause C.<br>If A causes B and B does not cause C, then A does not cause C. |

| If A causes B and B causes C. Does A cause A? | |
|---|---|
| GPT-3 | A does not cause A. |
| Luminous | A: The answer is no.<br>The cause of a thing is the thing that caused it. |
| OPT | If A causes B and B causes C, then A causes C.<br>If A causes B and B causes C, then A does not cause C.<br>If A causes B and B does not cause C, then A does not cause C. |

| If A causes B and B causes C. Does B cause A? | |
|---|---|
| GPT-3 | There is no definitive answer to this question as it depends on the specific situation. In some cases, B may cause A, while in others A may cause B. |
| Luminous | A: The answer is no.<br>The cause of A is A itself.<br>The cause of B is A and B.<br>The cause of C is A, B and C. |
| OPT | If A causes B and B causes C, then A causes C.<br>If A causes B and B causes C, then A does not cause C.<br>If A causes B and B does not cause C, then A does not cause C. |

| If A causes B and B causes C. Does C cause A? | |
|---|---|
| GPT-3 | There is no definitive answer to this question as it depends on the specific situation. In some cases, it may be true that C causes A, while in other cases it may not be true. |
| Luminous | A: The answer is no.<br>The cause-effect relationship is a logical relationship. It is not a physical relationship. |
| OPT | If A causes B and B causes C, then A causes C.<br>If A causes B and B causes C, then B causes C.<br>If A causes B and B causes C, then A does not cause C. |

| If A causes B, B causes C and C causes D. Does A cause D? | |
| --- | --- |
| GPT-3 | A causes D. |
| Luminous | A: The answer is no.<br>The cause of A is A itself.<br>The cause of B is A and B.<br>The cause of C is A, B and C. |
| OPT | If A causes B, B causes C and C causes D. Does A cause D?<br>If A causes B, B causes C and C causes D. Does A cause D? |

**Remaining Abstract Reasoning Results.** Can be found here: https://anonymous.4open.science/r/LLMsAndCausality-E220/media/Abstract-Reasoning.pdf