COLD: Causal reasOning in cLosed Daily activities

Abhinav Joshi* **Areeb Ahmad*** **Ashutosh Modi**
Department of Computer Science and Engineering
Indian Institute of Technology Kanpur (IIT Kanpur)
Kanpur, India
{ajoshi,areeb,ashutoshm}@cse.iitk.ac.in

Abstract

Large Language Models (LLMs) have shown state-of-the-art performance in a variety of tasks, including arithmetic and reasoning; however, to gauge the intellectual capabilities of LLMs, causal reasoning has become a reliable proxy for validating a general understanding of the mechanics and intricacies of the world similar to humans. Previous works in natural language processing (NLP) have either focused on open-ended causal reasoning via causal commonsense reasoning (CCR) or framed a symbolic representation-based question answering for theoretically backed-up analysis via a causal inference engine. The former adds an advantage of real-world grounding but lacks theoretically backed-up analysis/validation, whereas the latter is far from real-world grounding. In this work, we bridge this gap by proposing the **COLD** (**Causal reasOning in cLosed Daily activities**) framework, which is built upon human understanding of daily real-world activities to reason about the causal nature of events. We show that the proposed framework facilitates the creation of enormous causal queries (~ 9 million) and comes close to the mini-turing test, simulating causal reasoning to evaluate the understanding of a daily real-world task. We evaluate multiple LLMs on the created causal queries and find that causal reasoning is challenging even for activities trivial to humans. We further explore (the causal reasoning abilities of LLMs) using the backdoor criterion to determine the causal strength between events.

1 Introduction

In recent times, Large Language Models (LLMs) have shown remarkable generalization capabilities [Devlin et al., 2019, Radford et al., 2019, Brown et al., 2020]. Consequently, the ability to perform causal reasoning (often considered a core feature of intelligence [Penn and Povinelli, 2007, Pearl and Mackenzie, 2018]) has sparked research interest in the context of LLMs, aiming to answer if causal reasoning is possible with LLMs [Weber et al., 2020, Jin et al., 2023, 2024, Cohrs et al., 2023, Romanou et al., 2023, Yang et al., 2023, Mitchell et al., 2023, Vashishtha et al., 2023, Stolfo et al., 2023]. On a broader level, there are two lines of work; first, that treats the causal reasoning via learning relationships between the events that are grounded in the real world [Gordon et al., 2012, Ho et al., 2022, Zečević et al., 2023, Zhang et al., 2023, Wang et al., 2023]. Second line of work relies on a causal inference engine and establishes relationships between variables via symbolic representation [Jin et al., 2023, 2024]. The former relies on understanding real-world events but lacks formal definitions that adhere to the causal inference theory. The latter solves the issue using a causal inference engine but uses symbolic representations not grounded in the world, making the causal queries more like a test for the understanding of causal theory. Though the first line of work includes real-world events, the causal queries are often limited and could be answered by memorizing the causal relationships between the events. Recent findings that include rigorous analysis using a causal

*Equal Contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0167

inference engine claim LLMs to be *"Causal Parrots"* [Zečević et al., 2023], i.e., the LLMs tend to pick up (memorize) patterns in the training data to perform well on the causal reasoning benchmarks. Moreover, some initial findings by Tang et al. [2023] suggest that LLMs perform significantly better when semantics are consistent with commonsense but struggle to solve symbolic tasks, pointing towards semantic representation to be better for proper validation of LLMs, leading to a conclusion that an in-depth analysis using real-world events is necessary.

In this work, we bridge the gap between the two approaches by proposing the **COLD** (**Causal reasOning in cLosed Daily activities**) framework, based on the human understanding of real-world daily activities capturing commonsense (for example, "making coffee," "boarding an airplane," etc.), that adheres to causal theory literature. It is more natural to frame real-life reasoning-based queries via language; consequently, we follow the literature on Causal Commonsense Reasoning (CCR), which studies the relationships between real-world events (described via natural language). CCR is a non-trivial task of estimating the cause-and-effect relationship between events that are studied under the umbrella of commonsense reasoning [Kuipers, 1984, Gordon et al., 2012, Zhang et al., 2022b, Wang et al., 2023, Chun et al., 2023, Du et al., 2022]. The events in CCR generally refer to actions taking place in an activity in the real world. For example, consider the activity of "`traveling by an airplane`" given in Fig. 1, where the occurrence of all the events is confounded by a universal variable **U** ("intention to perform a task"). Moreover, there are a few events that cause one another. For example, the event "`checking in luggage`" (E_1) caused the occurrence of events like "`waiting at the luggage belt`" (E_2) after the flight, i.e., in an alternate universe where one does not checks in luggage and goes with the cabin bags, will never wait for their luggage after the flight has landed. Moreover, some of the events have no causal impact, like "`find the boarding gate`" (E_3) has no causal relationship with "`checking in luggage`" (E_1). More formally,

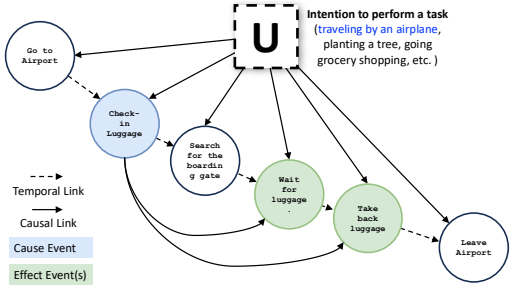

Figure 1: **U** denotes the unobserved variables, confounding all events present in a real-world activity. In an activity, some events cause other events to happen. For example, in *"traveling by an airplane"*, the event of *"check-in luggage"* causes events like "`taking back luggage.`"

$$\begin{aligned}\Delta_{(E_1 \to E_2)} &= \mathbb{P}(E_2|do(E_1)) - \mathbb{P}(E_2|do(\neg E_1)) \\ \Delta_{(E_1 \to E_3)} &= \mathbb{P}(E_3|do(E_1)) - \mathbb{P}(E_3|do(\neg E_1))\end{aligned} \tag{1}$$

where $do(.)$ denotes the **do** operator [Pearl, 2012] showing the intervention on E_1, and Δ is the *causal estimand* capturing the causal strength between two events, i.e., $\Delta(E_1 \to E_2)$ is expected to be higher when compared to $\Delta(E_1 \to E_3)$. Note, CCR excludes the causal events that are beyond the reach of commonsense knowledge, for example, does "`planting trees`" have a direct impact on the "`rainy season`"? or does "`providing free education`" improve the "`economic condition of the country/state`"; does "`carpooling`" directly impact "`air pollution`", etc. A noteworthy point concerning causality is that though the logical temporal (or prototypical) order of these events provides a weak signal about causal relationships, temporal precedence does not always imply causation (§2). For example, one could erroneously argue that "`boarding a plane`" is also the cause of "`waiting at the luggage belt`" since without "`boarding a plane,`" one cannot wait for the luggage belt.

For building a causal reasoning framework (based on CCR) around real-life daily activities, one would require a few primary features readily available: 1) Clear distinction between the events, i.e., the events should encapsulate describing a particular step in an activity, 2) Causal Dependency between the variables/events, i.e., there should be some events causing other events to occur. 3) Causal independence of events with the rest of the world, i.e., the occurrence of events should be independent of events that are not part of the activity (i.e., the covariates are balanced out). We found that *"Scripts"* [Schank, 1975, Schank and Abelson, 1975] provide a concrete medium that satisfies all these requirements. Scripts are defined as a sequence of events describing a prototypical activity, such as going to a restaurant, and hence capture commonsense knowledge about the world. [Schank and Abelson, 1975, Modi et al., 2016, Wanzare et al., 2016, Ostermann et al., 2018, Modi, 2016, 2017, Modi et al., 2017, Modi and Titov, 2014]. Moreover, different people have similar understandings of

https://doi.org/10.52202/079017-0167

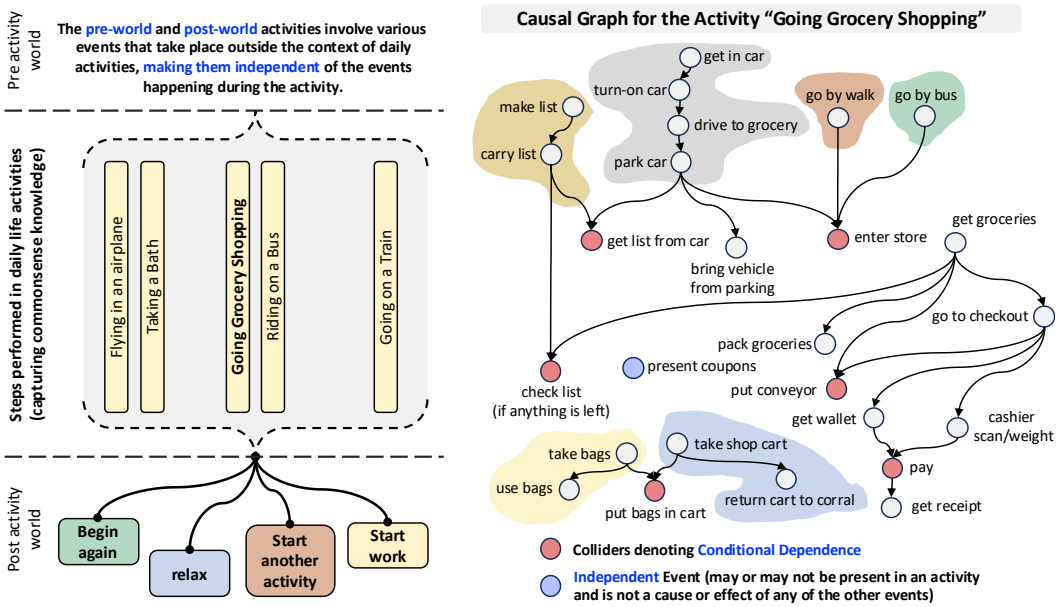

Figure 2: **Left:** the figure represents the closed nature of daily real-world activities (capturing commonsense, commonly understood by humans), start and end given the context of the task, i.e., the pre-activity world and post-activity world activities marginalize out the dependence of event occurring during the activity with the rest of the world. **Right:** Causal Graph for "`going grocery shopping.`" Notice the collider (red nodes) makes the independent set of nodes (highlighted in different colors) unconditionally independent in the causal graph. In contrast, when given a condition on a collider ("`put bags in cart`", the two clusters (yellow and blue) become dependent (if collider is observed, both yellow and blue clusters may have been observed as well).

the activity in the form of scripts that inherently balance out the covariates present in the real world, i.e., all the activities have the same starting and ending point and account for common exogenous and endogenous variables, providing a suitable platform to establish a cause-and-effect relationship between the events. In other words, for an activity like "`flying in an airplane,`" or "`going grocery shopping`" (also see Fig. 2, left) the events that happened before starting the activity and after completing the activity are marginalized out using a common understanding of these activities by different humans and hence will have no causal relations with any of the exogenous events during the activity. Creating a causal graph for script knowledge, i.e., establishing relationships between events taking place during the activity, provides a perfect platform for creating causal queries, thus providing a medium to establish CCR between events. In a nutshell, we make the following contributions:

- We propose **COLD** (**Causal reasOning in cLosed Daily activities**), a CCR framework based on Script knowledge (daily activities involving commonsense) that provides a closed system to test the understanding of causal inference grounded in the real-world. The proposed framework adheres to SUTVA (Stable Unit Treatment Value Assumption) [Cox, 1958, Rubin, 1980] by design (§3). **COLD** consists of activity-specific observational graphs (created via crowd-sourcing) and causal graphs. Further, **COLD** facilitates creating an enormous number of causal queries (e.g., $2,887,950$ per activity) via causal query triplets from the causal graph. This comes close to the mini-Turing test [Pearl and Mackenzie, 2018], where the story becomes the understanding of the daily activity, and the sampled enormous causal queries help in the exhaustive and rigorous evaluation of LMs.
- We devise various design mechanisms for estimating causal strength analytically and show how the representations learned by language models can be validated.
- Via detailed experimentation on the widely used open-weight language models, including encoder-only models (RoBERTa-MNLI) and autoregressive models (gpt-neo-125M, gpt-neo-1.3B, gemma-2b, gpt-neo-2.7B, phi-2, gpt-j-6B, Llama-2-7b-chat-hf, Mistral-7B-v0.1, gemma-7b, and Meta-Llama-3-8B) we estimate the causal reasoning capability of the learned representations. We release the framework, model code, and results via https://github.com/Exploration-Lab/COLD.

https://doi.org/10.52202/079017-0167

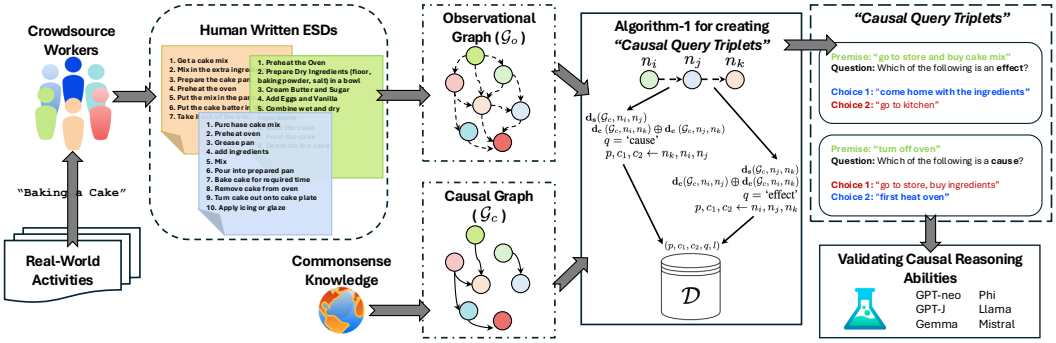

Figure 3: The proposed **COLD** framework for evaluating LLMs for causal reasoning. The human-written Event Sequence Descriptions (ESDs) are obtained from crowdsource workers and include a telegrammic-style sequence of events when performing an activity. The Observational Graph and the Causal Graph for an activity are used to create causal query triplets (details in Algorithm 1), shown towards the right. Using counterfactual reasoning, "going to the kitchen" is possible without going to the market (if the ingredients are already available), making "come home with the ingredients." a more plausible effect among the given choices. Similarly, in the second example, the event "going to market" has no direct relation with the event "heating the oven".

2 Background

The Mini Turing Test proposed by Pearl and Mackenzie [2018] is designed in a question-answering format to validate the understanding of causal knowledge about a simple story. The primary feature of a mini-Turing test is the enormous number of causal queries that can be framed using the underlying causal graph, which governs the occurrence of events in the story. Due to the enormous number of causal queries, passing the mini-Turing via memorization becomes combinatorially heavy, and hence, the authors argue that it can only be beaten if one has access to the underlying causal graph governing the occurrence of events (i.e., one has the ability to reason causally about the events). In this work, though, we only consider a more straightforward case of choice-based causal triplets; we realize the number of causal queries that can be created is enormous and helps validate the causal reasoning abilities coming close to the mini-Turing test.

***d*-separation:** Establishing the independence of variables becomes non-trivial when dealing with complex interactions among multiple variables. d-separation [Pearl, 1988] facilitates the determination of conditional independence between two sets of nodes $\mathbf{X}$ and $\mathbf{Y}$ in a graphical model $\mathcal{G}$ given another set of nodes $\mathbf{Z}$. d-separation asserts that $\mathbf{X}$ and $\mathbf{Y}$, given the set $\mathbf{Z}$, are d-separated if all paths for every node in $\mathbf{X}$ and every node in $\mathbf{Y}$ are blocked by conditioning on $\mathbf{Z}$, denoted as $\mathbf{X} \perp\!\!\!\perp_{\mathcal{G}} \mathbf{Y} \mid \mathbf{Z}$. A path p is blocked by a set of nodes $\mathbf{Z}$ [Pearl et al., 2016], if and only if: 1) p contains a chain of nodes $\mathbf{A} \rightarrow \mathbf{B} \rightarrow \mathbf{C}$ or a fork $\mathbf{A} \leftarrow \mathbf{B} \rightarrow \mathbf{C}$ such that the middle node $\mathbf{B}$ is in $\mathbf{Z}$ OR 2) p contains a collider $\mathbf{A} \rightarrow \mathbf{B} \leftarrow \mathbf{C}$ such that the collision node $\mathbf{B}$ or its descendant is not in $\mathbf{Z}$.

Backdoor Criterion: A set of variables W satisfies the backdoor criterion relative to T and Y if the following are true:

(A) W blocks all backdoor paths from T to Y i.e. blocking confounding or non-causation association paths
(B) W doesn't contain any descendants of T

Then, W satisfies the backdoor criterion [Pearl et al., 2016, Neal, 2020]. We make use of the backdoor criterion to estimate the causal estimand, capturing the relationship between the causal events. (Refer App. C for more detail)

3 COLD (Causal reasOning in cLosed Daily activities)

We propose **COLD** (**Causal reasOning in cLosed Daily activities**) framework for testing causal reasoning abilities of natural language understanding systems such as LLMs. Fig. 3 gives an overview of the creation process. We use crowd-sourced data of script knowledge to create observational

https://doi.org/10.52202/079017-0167

graphs which is further used along with manual intervention to create causal graphs. Subsequently, an algorithm is used to create an enormous number of causal queries (causal triplets), which are further used to test LLMs for causal reasoning. Next, we explain each of the steps in more detail.

Task Formulation: COLD is motivated by Causal Commonsense Reasoning (CCR), which we define as the task of finding the strength of the cause-and-effect relationship between two events (E_1 and E_2) given in an activity $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of all activities. For example, for an activity like "`going in an airplane`", the central question is to determine the causal relationship between two events that occur during the activity (events like "`checking in luggage`" and "`waiting for luggage`"). Since reasoning about a sequence of events is tedious (and sometimes confusing [Do et al., 2011]), researchers often rely on a more plausible cause rather than defining a definite causal event. For instance, COPA dataset [Gordon et al., 2012] provides a premise event and a corresponding causal query question along with two choices (see Table 1 for an example); a system is required to predict which of the two choices is most plausible cause/effect as required by the question.

Creating a Closed Causal System

Given the nature of Script knowledge (satisfying the criterion of balanced covariates, §1), we use a script corpus called DeScript [Wanzare et al., 2016] for creating the observational graphs. DeScript is a corpus with a telegram-style sequential description of an activity in English (e.g., baking a cake, taking a bath, etc.). DeScript is created via crowd-sourcing. For a given activity, crowd-workers write a point-wise and sequential short description of various events involved in executing the activity (this one complete description is called an ESD (Event Sequence Description)). DeScript collects data for a set of 40 daily activities (100 ESDs each) varying in complexity and background knowledge. Additionally, for a given activity, semantically similar events from different ESDs are manually aligned by human annotators (for more details, refer to Wanzare et al. [2016]). These alignments were later used by Joshi et al. [2023b,a] to create a DAG representing the overall activity. In our work, we use these DAGs as the observational distribution of an activity ($\mathcal{G}_o^{(a)}$, where $a \in \mathcal{A}$, where $\mathcal{A}$ is the set of all activities). These DAGs provide a medium for generating enormous trajectories (scales from $1.6e+16$ to $1.3e+27$, also see Table 2), that are coming directly from human annotations (alignment as well as the ESDs), providing us a proxy to represent the understanding of daily activities.

Observational Distribution ($\mathcal{G}_o$): Note that the graphs $\mathcal{G}_o$, approximately represent (almost) all possible ways in which an ESD can be written for an activity, providing the true observational distribution, i.e., how the combinations of events will look like while performing the activity in the real world (see App. A.3 for examples).

Causal Graphs ($\mathcal{G}_c$): To reason about the causal relationships between the events (nodes of $\mathcal{G}_o$), we would need the underlying causal graph that shows the cause of occurrence of various activities (directly or indirectly). We construct the causal graphs manually by reasoning about the independence of various events in the activity. Fig. 2 shows the pictorial representation of one of the created causal graphs for the activity "`going grocery shopping.`" Notice that various sets of events in the graph create clusters, denoting independence between various events. For example, nodes related to `make list` cause the events that involve the presence of a list and do not cause events like `going via car` (as some of the population will not create a list for shopping). Similarly, the mode of transportation (car/bus/walk) is independent of the events performed inside the store.

Causal Query Triplets: The obtained Causal Graph ($\mathcal{G}_c$), for an activity provides a medium to reason about causal links between the events. Notice in Fig. 2, how the red nodes (colliders) help separate out the independent event clusters. For example, the nodes '`get list from car`' and '`check list (if anything is left)`' being colliders, separate out the making list-related events with the rest of the graph. Similarly, the node '`put bags in cart`' separates out the '`take shop cart`' and '`take bags`'. Another interesting property represented in the obtained causal graph is the conditional dependence between various clusters. For example, the cluster related to '`get in car`' is unconditionally independent of '`make list`'. However, if we condition on the collider ('`get list from car`'), they become dependent, i.e., if '`get list from car`' is observed, it means that the person will have created the list as well as went by car for the grocery shopping (similarly for node '`put bags in cart`'). *d*-separation (§2) provides an easier way to establish conditional/unconditional independence between the set of nodes. For creating the dataset of causal queries (similar to other datasets like COPA [Gordon et al., 2012]), we need a triplet of three events (premise p, Choice-1 c_1, Choice-2 c_2) associated with a question about 'cause' or 'effect' relationship, i.e., given the premise which of the two choices is the cause/effect? (Table 1). We call these triplets *Causal Query*

https://doi.org/10.52202/079017-0167

Table 1: The table shows examples of causal query triplets created using the causal graphs ($\mathcal{G}_c$) and observational graphs ($\mathcal{G}_o$) in Algorithm 1. The top row is taken from the COPA dataset [Gordon et al., 2012] for the purpose of comparison. Note the examples in the table show the samples taken from the instance version.

Dataset	Premise	Choice-1 (1)	Choice-2 (2)	question	answer
COPA	The man turned on the faucet. The girl found a bug in her cereal. The hamburger meat browned.	The toilet filled with water. She poured milk in the bowl. The cook froze it.	Water flowed from the spout. She lost her appetite. The cook grilled it.	effect effect cause	2 2 2
COLD (Cake)	buy proper ingredients. measure ingredients in designated measuring cups. bake until cake is ready. turn off oven. preheat oven to 350 degrees.	go home with ingredients. whisk after each addition.stir to combine. set timer. go to store, buy ingredients. turn off oven.	wait for the timer to go off. clean up the mess. carefully remove cake from pan. first heat oven. prepare the microwave oven and utensils	effect effect cause cause effect	1 1 1 2 1
COLD (Shopping)	pay total. get the bill for groceries. pay for it . go back to the car.put the bags in the car. take the full cart to the checkout lane.	get receipt. pay for the grocery. start at the non-cold side of the store. take your car and drive to grocery shop. watch prices as the checker scans.	place cart into cart corral. return cart to store. go to shelf and get the food items. bring items to checkout. go down aisles.	effect cause cause cause effect	1 1 2 1 1
COLD (Train)	check train schedules. you board your train and find your seat. get off train. go out of the station. arrive at destination.	choose a destination. find your seat or compartment. go out of the station. put carry on luggage in overhead bin. when train reaches destination, exit train.	go to the car. wait for train. take all the luggage out of train. get off at your correct stop. walk to the train platform.	cause effect effect cause effect	1 1 1 2 1
COLD (Tree)	go to garden center. fill hole with dirt and fertilizer. place the tree at the top of the hole. fill in dirt around the tree gently. place tree sapling into hole.	transport it home. get tree. cover the roots with dirt. take it home. pack dirt back in.	choose type of tree. dig hole big enough for tree to grow. get a tree. place the tree in the hole. find place for tree.	effect cause effect cause effect	1 2 1 2 1
COLD (Bus)	pull signal for stop. board the bus when it arrives. step on bus. buy bus ticket. find seat on bus and sit.	bus stops at destination. while boarding, pay the driver the required fee take available seat. when it arrives get on. find out what bus to take.	stand up and go to door. pull signal for stop. wait for the bus to arrive. when your stop approaches, pull cord. the bus arrives at the departure station.	effect effect effect cause cause	1 1 1 1 2

Table 2: The table provides details of the observational graph ($\mathcal{G}_o$) for 5 activities. The Causal Query Triplets represent the total number of triplets generated via Algorithm 1. The instance version shows the number of samples present in the instance version (including different text instances describing the same event) of the created dataset. Table 1 shows a small sample taken for the 5 activities. Overall, the huge number of samples highlights the exhaustive nature of evaluation that can be done for LLMs.

Activity	Nodes	Compact Trajectories	Total Trajectories	Causal Query Triplets	Instance version (Num. samples)
Baking a **Cake**	28	177030	$1.3e+27$	864	2887950
Riding on a **Bus**	20	13945	$1.3e+17$	334	834046
Going Grocery **Shopping**	33	626096	$3.1e+26$	1984	3739184
Going on a **Train**	26	133799	$4.9e+22$	950	1213114
Planting a **Tree**	23	4466	$1.6e+16$	260	846046
Total Dataset Samples	-	-	-	**4392**	**9,520,340**

Triplets that are used to frame a causal query between the events. App. B, Algorithm 1 presents the mechanism to create a dataset of causal query triplets. We start by constructing the set of possible triplets, and sort the nodes in every triplet using the topological order preset in the observational graph (a DAG). Further, using d-separation, we figure out the triplets that have one node d-separated from the other nodes. The d-separated node becomes the wrong choice. The premise and correct choice are determined from the remaining choices, leading to a 'cause' and 'effect' query based on the topological order, i.e., the event occurring before (temporally) in an activity becomes 'cause.' The other event becomes the plausible 'effect.' Note that the temporal precedence is generally assumed essential for defining causation, and it is one of the most important clues that is used to distinguish causal from other types of associations [Mill, 1898, Hill, 1965, Pearl and Verma, 1995]. For our framework, we also consider the topologically sorted order obtained from the observational graphs and use the temporal order to define the causal query triplets, i.e., the cause events will always precede the effect events (see App. B, Algorithm 1, where $\mathcal{G}_o$ helps determine the temporal order of events). Table 1 shows a sample of the created dataset in comparison to the COPA dataset. Note that each node (premise, Choice-1, Choice-2) in $\mathcal{G}_c$ also has multiple texts (instances of text describing the same event) written by different crowd-sourced workers. We can further use these text instances to enrich the created dataset by considering all the available instances to create all possible combinations. This strategy increases the scale of the created dataset by a huge margin. Overall, we found a dataset created from triplets using Algorithm 1 results in **4,392** tuples, which, after using the text instances, increases to **9,520,340**, bringing it close to 'mini turning test" [Pearl and Mackenzie, 2018].

Adherence to SUTVA: In causal literature, a fundamentally acknowledged *Stable Unit Treatment Value Assumption* (SUTVA) [Cox, 1958, Rubin, 1980]) requires that for each unit (e.g., sequence of events), there is only one version of the non-treatment, i.e., for an event in the sequence, there lie only

Table 3: The table provides evaluation results of language models over the created causal triplets.

Triplets	Model Name	cake	shopping	train	tree	bus
causal triplets	gpt-neo-125M	50.71	50.01	49.99	50.13	50.15
	gpt-neo-1.3B	44.77	45.69	42.52	45.67	42.89
	gemma-2b	53.76	52.19	60.57	60.71	53.64
	gpt-neo-2.7B	50.00	50.01	50.00	50.01	50.00
	phi-2	85.14	83.65	77.29	82.24	71.74
	gpt-j-6B	49.59	50.02	50.29	49.92	49.93
	Llama-2-7b-chat-hf	77.92	72.41	73.48	72.40	68.21
	Mistral-7B-v0.1	77.64	69.38	68.46	72.43	69.37
	gemma-7b	81.47	82.26	77.24	80.78	70.29
	Meta-Llama-3-8B	80.79	76.46	76.08	78.21	67.39

two versions occurring and not occurring. SUTVA plays a vital role in causal inference by ensuring that each unit's treatment assignment has a consistent impact, facilitating the accurate estimation of treatment effects. Our framework closely adheres to SUTVA assumptions (details in A.1).

Comparison with Other Causal Datasets: We briefly compare the created dataset with the existing set of causal reasoning datasets in App. Table 5. The created dataset serves as a middle ground, having both real-world groundings as well as an underlying causal graph to create an exhaustive set of causal queries.

4 Experiments and Results

COLD provides a causal query dataset for evaluating LMs for causal understanding. In particular, we consider the "Causal Query Triplets" (Table 2) coming from compact trajectories as a base and sample the instance version coming from the same skeleton. Since it is not possible to evaluate all the possible causal queries that could be created using our framework, we use $10K$ samples for each activity to report our findings. For a fair comparison between various models and better reproducibility, we freeze the sampled causal query triplets and compare the success rate over the frozen samples. We evaluate via two methods. First, as done in previous work Jin et al. [2024, 2023], Chen et al. [2024], we first experiment with various LLMs using a prompt-based evaluation scheme; second, we propose other mechanisms (based on causal theory, e.g., Average Treatment Effect) that could be used to perform an in-depth analysis of evaluating causal relationships between events.

Causal Reasoning Evaluation of LLMs via Prompts: We start with the prompt-based evaluation of recent open-weight LLMs (gpt-neo-125M, gpt-neo-1.3B, gpt-neo-2.7B [Black et al., 2021], gemma-2b [Team et al., 2024], phi-2 [Javaheripi et al., 2023], gpt-j-6B [Wang and Komatsuzaki, 2021], gemma-7b [Team et al., 2024], Llama-2-7b-chat-hf [Touvron et al., 2023], Mistral-7B-v0.1 [Jiang et al., 2023], and Meta-Llama-3-8B [Dubey et al., 2024]) We frame the prompt as a multi-choice question-answering (MCQA) objective [Robinson and Wingate, 2023]. The prompt is intentionally structured so that the LLM is intended to predict a single choice token (Such as " A", " B", etc.). Robinson and Wingate [2023] highlight the advantages of MCQA-based evaluation over cloze evaluation [Brown et al., 2020](where the LLMs are expected to generate the entire answer in a cloze test), leading to a significant boost in various tasks, including commonsense-based tasks. App. E, Fig. 5 presents various prompt templates for autoregressive experiments, and App. E Fig. 6 shows a few qualitative examples for the framed causal query templates. Table 3 shows the success rate obtained for various LLMs. The success rate corresponds to the percentage of queries where the LLM predicts the desired choice. We observe that reasoning causally about simple daily activities is challenging when a rigorous test is framed, validating the dependencies between the events. Overall, for the more common activities like `baking a cake` and `going grocery shopping`, the LLMs perform better when compared to activities like `boarding a bus` or `planting a tree`. We also experimented with another version of the dataset, where incorrect choice may correspond to temporally plausible but causally implausible events. The results drop significantly in this case; details and results are provided in the App. F.1.

Evaluation using Average Treatment Effect (ATE) (Δ): Computing the Average Treatment Effect (Δ) helps establish the strength of causal links given a context (Eq. 1). In our setup, to estimate

https://doi.org/10.52202/079017-0167

$P(y|do(x))$ (i.e., the causal estimand) from statistical estimands (obtained from observational distribution), we make certain reasonable assumptions about the underlying process that governs the relation among variables/events and then utilize the implications of these assumptions. For any activity taking place, the causal relationships between two events E_1 and E_2 may have a causal link along with a non-causal link through a set of confounders z. We define the confounder $z = \{t_i | t_i \in \mathcal{T}\}$, where $\mathcal{T}$ denotes all the trajectories (sequence of events) from the start of the activity till the event E_1. The temporal nature of events makes this assumption suitable since the occurrence of E_1 and E_2 can be confounded by all the events preceding E_1. Note the possibility of unobserved confounders (events that are not explicitly mentioned but may be affecting the mentioned events) in our case is removed due to two reasons: 1) Keeping a closed system representation with a large number of diverse scripts (written by humans) helps cover the set of most generic and diverse events either implicitly or explicitly as a part of the activity, and 2) The causal reasoning goal is restricted to figuring out the causal effect between the events that are present explicitly. Assuming the unmentioned events have insignificant effects, we can establish that there will not be any unobserved confounders. This assumption makes the observed confounders satisfy the backdoor criterion [Pearl, 1993] that make sufficient adjustment sets. By using the backdoor criterion (App. C), the interventional distributions are estimated as follows:

$$P(E_2|do(E_1)) = \sum_{t_i \in \mathcal{T}} p_*(E_2|E_1, z = t_i) p_*(z = t_i) \tag{2}$$

Note that the true observational distribution, i.e. $p_*(E_2|E_1, z)$ and $p_*(z)$ both are unknown and have to be approximated ($\hat{p}(E_2|E_1, z)$ and $\hat{p}(z)$). Further, we describe ways $\hat{p}$ can be estimated via multiple design mechanisms. Due to space limitation, we only describe the $\hat{p}$ estimation via language models below and move the statistical analysis using the original trajectories and observational graphs to the App. D.2.

ATE using Language Models Since pre-trained LMs capture world knowledge [Devlin et al., 2019, Brown et al., 2020, Li et al., 2023, Nanda et al., 2023, Karvonen, 2024], these provide a suitable proxy for establishing relationships between these events. For our experiments, we consider a simple reasoning capability of language models, i.e., to reason about the temporal order of various events, i.e., given an event, what is the likelihood of the occurrence of another event? We further ask if this can be used to estimate the causal relationship between the events (a similar strategy is used by Zhang et al. [2022b] for zero-shot causal estimation). It is worth noting that for these activities about daily activities, one way to find causes is to establish the temporal likelihood of the events. For each of the multiple language models, we frame the temporal prediction differently.

Encoder-only Models: For BERT-based models trained for mask token prediction, we model the temporal prediction using the probability assigned to the mask tokens "*before*" and "*after*" [Zhang et al., 2022b]. Given two events E_1 and E_2, the temporal link is predicted using a prompt like E_1 <mask> E_2, and the scores corresponding to the before and after tokens are collected. App. D, Fig. 7, the top row highlights the prompt template used for BERT-based models. For encoder-only experiments, we consider RoBERTa MNLI [Liu et al., 2019].

Decoder-only Models: For other language models that are autoregressive in nature, we modify the prompt to predict the temporal order as the last token. We again use the MCQA-based prompting style to frame the temporal order query by providing "*before*" and "*after*" as the options in the prompt. App. D, Fig. 7, the bottom row highlights the prompt template used for Decoder-only Models.

Interventions: We utilize the SUTVA assumption in the proposed framework to devise an intervention over a trajectory in natural language form. App. D, Fig. 8 shows the style of intervention made by an event (E_1) taking place ($do(E_1)$) or not taking place $do(\neg E_1)$.

Given the above strategies, LMs can be used to evaluate $\hat{p}(E_2|\neg E_1, z = t)$, by feeding the prompt that contains E_1, E_2 and the $z = t$) and predict the temporal nature between E_1 and E_2, given a trajectory $z = t$. Further, applying the backdoor criterion for multiple trajectories $\mathcal{T}$, we obtain

$$\begin{aligned} p_{\mathcal{M}}(E_2|do(E_1)) &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \hat{p}(E_2|E_1, z = t) \\ p_{\mathcal{M}}(E_2|do(\neg E_1)) &= \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \hat{p}(E_2|\neg E_1, z = t) \end{aligned} \tag{3}$$

Table 4: Accuracy over the causal triplets for various Δ estimates. The blue text denotes the improvements by backdoor adjustments over the temporal scheme for multiple language models. **Bold text** represents the best-performing method for a particular activity.

$\hat{p}$ **estimation**	**Scheme**	**ATE**	`Cake`	`Shopping`	`Train`	`Tree`	`Bus`
Original Trajectories		Δ_o	28.20	34.30	31.10	30.10	30.40
Observational Graphs	-	Δ_n	30.40	30.10	29.80	28.60	25.4
	-	Δ_t	40.90	47.10	40.30	37.60	40.10
Language Models	Temporal	RoBERTa MNLI	46.80	54.00	45.50	52.70	43.00
		gpt-neo-125M	47.70	55.50	55.50	53.60	48.20
		gpt-neo-1.3B	47.40	45.40	53.30	43.40	52.90
		gemma-2b	43.80	41.70	52.20	49.70	49.80
		gpt-neo-2.7B	50.10	48.90	52.40	47.60	53.70
		phi-2	60.30	59.20	56.90	70.30	49.40
		gpt-j-6B	49.50	46.40	56.00	62.70	56.00
		Llama-2-7b-chat-hf	38.90	42.10	51.00	40.70	47.80
		Mistral-7B-v0.1	50.90	54.40	64.50	60.50	62.30
		gemma-7b	46.8	54.00	45.50	52.70	43.00
		Meta-Llama-3-8B	58.20	54.10	55.6	55.00	64.00
Language Models	Backdoor Adjustments	$\Delta_{\mathcal{M}}$ (RoBERTa MNLI)	59.20	54.40	56.30	57.50	53.30
		$\Delta_{\mathcal{M}}$ (gpt-neo-125M)	59.20	55.10	50.50	52.10	45.50
		$\Delta_{\mathcal{M}}$ (gpt-neo-1.3B)	51.30	50.70	55.00	43.90	49.00
		$\Delta_{\mathcal{M}}$ (gemma-2b)	44.50	45.30	52.60	63.50	43.90
		$\Delta_{\mathcal{M}}$ (gpt-neo-2.7B)	49.10	51.30	51.50	54.00	51.40
		$\Delta_{\mathcal{M}}$ (phi-2)	57.00	66.00	62.10	57.10	45.80
		$\Delta_{\mathcal{M}}$ (gpt-j-6B)	51.30	45.60	50.50	49.10	46.00
		$\Delta_{\mathcal{M}}$ (Llama-2-7b-chat-hf)	62.60	64.60	68.50	70.50	63.80
		$\Delta_{\mathcal{M}}$ (Mistral-7B-v0.1)	63.90	71.40	**73.70**	61.30	**67.00**
		$\Delta_{\mathcal{M}}$ (gemma-7b)	**72.80**	**77.80**	73.60	**71.90**	62.40
		$\Delta_{\mathcal{M}}$ (Meta-Llama-3-8B)	66.00	70.20	68.40	62.00	63.40

which can further be used to estimate the causal strength between the events E_1 and E_2.

$$\Delta_{\mathcal{M}} = p_{\mathcal{M}}(E_2|do(E_1)) - p_{\mathcal{M}}(E_2|do(\neg E_1))$$

Using the multiple Δ estimates defined above, we estimate the causal strength between the events available for an activity. We follow the scheme presented in the App. B, Algorithm 2 to compute the performance in terms of success rates.

Temporal Scheme: In this scheme, we validate if temporal ordering knowledge of LLMs could be directly used to estimate the causal estimand. We make use of templates shown in the App. Fig. 7. The causal estimand is estimated via the difference in logit values when intervening over an event, i.e., does the predicted probability take into account the context of events not happening? Surprisingly, we found that temporal ordering does provide a suitable proxy for estimating causal strength between the events. We further extend this approximation to incorporate the backdoor adjustments in the Δ.

Backdoor Adjustments: For the experiments with Language models, we apply the backdoor adjustment to estimate the causal estimand $\Delta_{\mathcal{M}}$. App. Fig. 8 shows the prompt template used to determine the relationship between the events. The prompt template takes a trajectory, t_i, that contains all the events till the event E_1 in sequential order of occurrence, further, an added prompt determines the intervention ($do(E_1)$ or $do(\neg E_1)$) and the causal estimand is estimated using the logit values associated with the predicted token. App. B, Algorithm 3 provides the designed scheme to compute unbiased causal estimands. We essentially flip the options and generate the scores associated with options 'A' and 'B' for increase and decrease, respectively (more details in the App. C.)

Table 4 shows a comparison between various design choices. We observe that when using LLMs for $\hat{p}$ estimation, the backdoor adjustment increases the performance over the temporal estimation scheme by a significant margin. The understanding of these activities is generic, and LLMs do provide a suitable set of sequences when prompted to generate a list of steps to complete the activity. For example, when prompted with `'Generate the sequential steps in a telegrammic style to perform the activity "going grocery shopping"'`, almost all the models we tested provide

https://doi.org/10.52202/079017-0167

a valid set of steps for the given activity. However, when prompted with causal queries, the lower performance signifies the lack of understanding of the underlying causal mechanism. The constructed dataset helps to rigorously validate the understanding of the activity through an enormous number of causal query triplets. The results show that although the LLMs can explain the activity in detail, including generating correct steps for performing tasks, causally reasoning about the set of events remains challenging.

Human Study: We conducted a small-scale human validation study over the created causal query dataset and asked 5 graduate students to answer 100 randomly sampled causal query triplets (20 per activity). We record an average performance of 92.20%. (More details about the human study are provided in the App. B)

5 Related Work

Causal reasoning has been an active research area in the ML community [Spirtes et al., 2000a, Peters et al., 2017, Schölkopf et al., 2021]. Some of the initial works highlight the causal nature of events present in text [Schank, 1975] as *'causal chains'*. Multiple works have considered creating benchmarks/datasets that capture causal relationships between the events described in the text (see App. Table 5). More recently, with the rapid growth of LLMs on reasoning/understanding tasks, attention has shifted to validating these general-purpose models capturing causal reasoning [Jin et al., 2023, Zečević et al., 2023, Willig et al., 2023a, Liu et al., 2023, Willig et al., 2023b, Zhang et al., 2022a, Jin et al., 2024]. App. A.2 Table 5 shows a broad overview of the existing causal Dataset/Benchmarks presented in the NLP community. In this work, the primary focus is to bridge the gap between various lines of work that consider natural language to learn/validate/reason about causal relationships between events.

6 Limitations and Future Directions

One of the primary limitations of our work is the limited set of activities. Though the frameworks support generating exhaustive/enormous causal queries, finding general commonsense reasoning activities/tasks that are well understood by humans remains challenging. Moreover, creating a causal graph for an activity increases as we move toward more long-term tasks. However, as a general test of causal intelligence, our framework provides a suitable platform to validate the reasoning capabilities more rigorously. In the future, it would be interesting to sample trajectories from the observational distribution $\mathcal{G}_o^a$ to create a training dataset and check if causal reasoning ability can be acquired by language modeling objectives (including other variants like presented in Lampinen et al. [2023]). We leave this detailed analysis for future endeavors. The proposed algorithm for causal triplet generation generates the simplest variant of causal queries in the form of causal triplets (also referred to as Pairwise Causal Discovery (PCD) task by [Chen et al., 2024]). More complicated causal queries can be generated, such as considering cases with common confounders, long/short causal chain dependency, etc. Moreover, taking formal definitions. (i.e., using the formal causal inference language) causal queries inspired from Jin et al. [2023, 2024] can be framed for a more rigorous analysis. Being at the initial state, we stick to the simple causal queries that provide two choices, and the task is to choose the more plausible cause. The creation of underlying causal graphs provides endless possibilities for creating varied versions of causal queries. In this work, we only consider an unconditional version of d-separation. In the future, the same causal graphs could be used to define more datasets for covering other rungs of the *'causal ladder'* [Pearl and Mackenzie, 2018].

7 Conclusion

In this paper, we proposed the **COLD** (**Causal reasOning in cLosed Daily activities**) framework for generating causal queries that can be used to rigorously evaluate LLMs. We performed extensive experimentation with LLMs for the task of Causal Commonsense Reasoning. Results indicate that LLMs are still far from a complete understanding of daily commonsensical activities and fail to answer causal queries when analyzed in an exhaustive manner. We believe this framework will provide a good platform for future research in understanding the causal reasoning abilities of LLMs.

Acknowledgments

We would like to thank the anonymous reviewers and the meta-reviewer for their insightful comments and suggestions. We would like to thank Google Deepmind India for helping us with the conference travel support.

References

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen tau Yih, and Yejin Choi. Abductive commonsense reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Byg1v1HKDB. [Cited on page 24.]

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, March 2021. URL https://doi.org/10.5281/zenodo.5297715. [Cited on page 7.]

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. CausalQA: A benchmark for causal question answering. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.291. [Cited on page 24.]

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165. [Cited on pages 1, 7, and 8.]

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. Causal evaluation of language models, 2024. [Cited on pages 7 and 10.]

Changwoo Chun, SongEun Lee, Jaehyung Seo, and Heuiseok Lim. CReTIHC: Designing causal reasoning tasks about temporal interventions and hallucinated confoundings. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10334–10343, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.693. URL https://aclanthology.org/2023.findings-emnlp.693. [Cited on page 2.]

Kai-Hendrik Cohrs, Emiliano Diaz, Vasileios Sitokonstantinou, Gherardo Varando, and Gustau Camps-Valls. Large language models for constrained-based causal discovery. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2023. URL https://openreview.net/forum?id=NEAoZRWHPN. [Cited on page 1.]

D. R. Cox. *Planning of Experiments*. Wiley, New York,, 1958. [Cited on pages 3, 6, and 23.]

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423. [Cited on pages 1 and 8.]

 https://doi.org/10.52202/079017-0167

Quang Do, Yee Seng Chan, and Dan Roth. Minimally supervised event causality identification. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://aclanthology.org/D11-1027. [Cited on pages 5, 24, 25, and 30.]

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.33. URL https://aclanthology.org/2022.acl-long.33. [Cited on pages 2, 23, and 24.]

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon

Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. [Cited on page 7.]

Jesse Dunietz, Lori Levin, and Jaime Carbonell. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In Nathan Schneider and Nianwen Xue, editors, *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-0812. URL https://aclanthology.org/W17-0812. [Cited on page 24.]

Jörg Frohberg and Frank Binder. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.229. [Cited on page 24.]

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Eneko Agirre, Johan Bos, Mona Diab, Suresh Manandhar, Yuval Marton, and Deniz Yuret, editors, **SEM 2012: The*

 https://doi.org/10.52202/079017-0167

First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052. [Cited on pages 1, 2, 5, 6, 23, and 24.]

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In Katrin Erk and Carlo Strapparava, editors, *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL https://aclanthology.org/S10-1006. [Cited on page 24.]

Austin Bradford Hill. The environment and disease: association or causation?, 1965. [Cited on pages 6 and 24.]

Matthew Ho, Aditya Sharma, Justin Chang, Michael Saxon, Sharon Levy, Yujie Lu, and William Yang Wang. Wikiwhy: Answering and explaining cause-and-effect questions, 2022. [Cited on page 1.]

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 2023. [Cited on page 7.]

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825. [Cited on page 7.]

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2023. [Cited on pages 1, 7, 10, and 24.]

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. Cladder: Assessing causal reasoning in language models, 2024. [Cited on pages 1, 7, 10, and 24.]

Abhinav Joshi, Areeb Ahmad, Umang Pandey, and Ashutosh Modi. From scripts to rl environments: Towards imparting commonsense knowledge to rl agents. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pages 2801–2803, 2023a. [Cited on page 5.]

Abhinav Joshi, Areeb Ahmad, Umang Pandey, and Ashutosh Modi. Scriptworld: Text based environment for learning procedural knowledge. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5095–5103. International Joint Conferences on Artificial Intelligence Organization, 8 2023b. doi: 10.24963/ijcai.2023/566. URL https://doi.org/10.24963/ijcai.2023/566. Main Track. [Cited on page 5.]

Adam Karvonen. Emergent world models and latent variable estimation in chess-playing language models, 2024. [Cited on page 8.]

Benjamin Kuipers. Commonsense reasoning about causality: Deriving behavior from structure. *Artificial Intelligence*, 24(1):169–203, 1984. ISSN 0004-3702. doi: https://doi.org/10.1016/0004-3702(84)90039-0. URL https://www.sciencedirect.com/science/article/pii/0004370284900390. [Cited on page 2.]

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. TellMeWhy: A dataset for answering why-questions in narratives. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.53. URL https://aclanthology.org/2021.findings-acl.53. [Cited on page 24.]

Andrew Kyle Lampinen, Stephanie C.Y. Chan, Ishita Dasgupta, Andrew Joo Hun Nam, and Jane X Wang. Passive learning of active causal strategies in agents and language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BRpi8YAfac. [Cited on page 10.]

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=DeG07_TcZvT. [Cited on page 8.]

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. The magic of IF: Investigating causal reasoning abilities in large language models of code. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9009–9022, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.574. URL https://aclanthology.org/2023.findings-acl.574. [Cited on page 10.]

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. [Cited on page 8.]

John Stuart Mill. *A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Princilples of Evidence and the Methods of Scientific Investigation*. Longmans, Green, 1898. [Cited on pages 6 and 24.]

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. Annotating causality in the TempEval-3 corpus. In Oleksandr Kolomiyets, Marie-Francine Moens, Martha Palmer, James Pustejovsky, and Steven Bethard, editors, *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-0702. URL https://aclanthology.org/W14-0702. [Cited on page 24.]

Melanie Mitchell, Alessandro B. Palmarini, and Arseny Moskvichev. Comparing humans, gpt-4, and gpt-4v on abstraction and reasoning tasks, 2023. [Cited on page 1.]

Ashutosh Modi. Event Embeddings for Semantic Script Modeling. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, 2016. doi: 10.18653/v1/K16-1008. URL https://aclanthology.org/K16-1008. [Cited on page 2.]

Ashutosh Modi. *Modeling Common Sense Knowledge via Scripts*. PhD thesis, Saarland University, 2017. [Cited on page 2.]

Ashutosh Modi and Ivan Titov. Inducing neural models of script knowledge. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 49–57, 2014. [Cited on page 2.]

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016. URL https://aclanthology.org/L16-1555. [Cited on page 2.]

Ashutosh Modi, Ivan Titov, Vera Demberg, Asad Sayeed, and Manfred Pinkal. Modeling Semantic Expectation: Using Script Knowledge for Referent Prediction. *Transactions of the Association for Computational Linguistics*, 2017. doi: 10.1162/tacl_a_00044. URL https://aclanthology.org/Q17-1003. [Cited on page 2.]

 https://doi.org/10.52202/079017-0167

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In Martha Palmer, Ed Hovy, Teruko Mitamura, and Tim O'Gorman, editors, *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-1007. URL https://aclanthology.org/W16-1007. [Cited on page 24.]

Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent linear representations in world models of self-supervised sequence models. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 16–30, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.2. URL https://aclanthology.org/2023.blackboxnlp-1.2. [Cited on page 8.]

Brady Neal. Introduction to causal inference. *Course Lecture Notes (draft)*, 2020. [Cited on pages 4 and 27.]

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. In *LREC*, 2018. [Cited on page 2.]

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988. [Cited on page 4.]

Judea Pearl. Comment: Graphical models, causality and intervention. 8(3):266–269, August 1993. doi: 10.1214/ss/1177010894. [Cited on page 8.]

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, page 3–11, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989. [Cited on page 2.]

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition, 2018. ISBN 046509760X. [Cited on pages 1, 3, 4, 6, 10, and 25.]

Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995. [Cited on pages 6 and 24.]

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. [Cited on pages 4, 25, and 27.]

Derek Penn and Daniel Povinelli. Causal cognition in human and nonhuman animals: A comparative, critical review. *Annual Review of Psychology*, 58:97–118, 02 2007. doi: 10.1146/annurev.psych.58.110405.085555. [Cited on page 1.]

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2017. ISBN 978-0-262-03731-0. URL https://mitpress.mit.edu/books/elements-causal-inference. [Cited on page 10.]

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. Counterfactual story reasoning and generation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5043–5053, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1509. URL https://aclanthology.org/D19-1509. [Cited on page 24.]

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. URL https://api.semanticscholar.org/CorpusID:160025533. [Cited on page 1.]

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. Event2Mind: Commonsense inference on events, intents, and reactions. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1043. URL https://aclanthology.org/P18-1043. [Cited on page 24.]

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B. [Cited on pages 7 and 29.]

Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. CRAB: Assessing the strength of causal relationships between real-world events. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.940. URL https://aclanthology.org/2023.emnlp-main.940. [Cited on pages 1 and 24.]

Donald B. Rubin. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593, 1980. [Cited on pages 3, 6, and 23.]

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. Atomic: An atlas of machine commonsense for if-then reasoning, 2019a. [Cited on page 24.]

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. Social IQa: Commonsense reasoning about social interactions. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1454. URL https://aclanthology.org/D19-1454. [Cited on page 24.]

Roger C. Schank. The structure of episodes in memory. In DANIEL G. BOBROW and ALLAN COLLINS, editors, *Representation and Understanding*, pages 237–272. Morgan Kaufmann, San Diego, 1975. ISBN 978-0-12-108550-6. doi: https://doi.org/10.1016/B978-0-12-108550-6.50014-8. URL https://www.sciencedirect.com/science/article/pii/B9780121085506500148. [Cited on pages 2 and 10.]

Roger C. Schank and Robert P. Abelson. Scripts, Plans, and Knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence*, IJCAI, 1975. [Cited on page 2.]

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. [Cited on page 10.]

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 883–898, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.78. URL https://aclanthology.org/2021.findings-acl.78. [Cited on page 24.]

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000a. [Cited on page 10.]

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000b. [Cited on page 28.]

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W.

 https://doi.org/10.52202/079017-0167

Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz,

Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. [Cited on page 24.]

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. A causal framework to quantify the robustness of mathematical reasoning with language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–561, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.32. URL https://aclanthology.org/2023.acl-long.32. [Cited on page 1.]

Xiaojuan Tang, Zilong Zheng, Jiaqi Li, Fanxu Meng, Song-Chun Zhu, Yitao Liang, and Muhan Zhang. Large language models are in-context semantic reasoners rather than symbolic reasoners, 2023. [Cited on page 2.]

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open models based on gemini research and technology, 2024. URL https://arxiv.org/abs/2403.08295. [Cited on page 7.]

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra,

 https://doi.org/10.52202/079017-0167

Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288. [Cited on page 7.]

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N Balasubramanian, and Amit Sharma. Causal inference using llm-guided discovery, 2023. [Cited on page 1.]

Ben Wang and Aran Komatsuzaki. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax, May 2021. [Cited on page 7.]

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. COLA: Contextualized commonsense causal reasoning from the causal inference perspective. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5253–5271, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.288. URL https://aclanthology.org/2023.acl-long.288. [Cited on pages 1, 2, 24, and 28.]

Lilian D. A. Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. A Crowdsourced Database of Event Sequence Descriptions for the Acquisition of High-quality Script Knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, 2016. URL https://aclanthology.org/L16-1556. [Cited on pages 2, 5, and 28.]

Noah Weber, Rachel Rudinger, and Benjamin Van Durme. Causal inference of script knowledge. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7583–7596, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.612. URL https://aclanthology.org/2020.emnlp-main.612. [Cited on page 1.]

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Probing for correlations of causal facts: Large language models and causality, 2023a. URL https://openreview.net/forum?id=UPwzqPOs4-. [Cited on page 10.]

Moritz Willig, Matej Zečević, Jonas Seng, and Florian Peter Busch. Causal concept identification in open world environments. In Martin Mundt, Keiland W. Cooper, Devendra Singh Dhami, Adéle Ribeiro, James Seale Smith, Alexis Bellot, and Tyler Hayes, editors, *Proceedings of The First AAAI Bridge Program on Continual Causality*, volume 208 of *Proceedings of Machine Learning Research*, pages 52–58. PMLR, 07–08 Feb 2023b. URL https://proceedings.mlr.press/v208/willig23a.html. [Cited on page 10.]

Linying Yang, Oscar Clivio, Vik Shirvaikar, and Fabian Falck. A critical review of causal inference benchmarks for large language models. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*, 2023. URL https://openreview.net/forum?id=mRwgczYZFJ. [Cited on page 1.]

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=tv46tCzs83. [Cited on pages 1, 2, 10, and 24.]

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. Understanding causality with large language models: Feasibility and opportunities, 2023. [Cited on page 1.]

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. In *International Joint Conference on Artificial Intelligence*, 2022a. URL https://api.semanticscholar.org/CorpusID:248986434. [Cited on page 10.]

Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. ROCK: Causal inference principles for reasoning about commonsense causality. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26750–26771. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/zhang22am.html. [Cited on pages 2, 8, 23, and 29.]

https://doi.org/10.52202/079017-0167

Appendix

Table of Contents

A COLD Framework Details **23**

A.1 Adherence to SUTVA 23

A.2 Comparison with previous Causal Reasoning Datasets/Benchmarks 23

A.3 Observational Graphs 24

B Algorithms in the COLD Framework **24**

C Backdoor Adjustments **26**

D Experiments and Results **28**

D.1 Compute Resources 28

D.2 Evaluation using Average Treatment Effect (ATE) 28

E Prompt Templates for Language Model based Experiments **29**

F Additional Results **30**

F.1 Temporally Plausible Choices in Causal Triplets 30

List of Tables

5 Comparison with existing Causal Benchmarks/Datasets in NLP 24

6 Human performance 25

7 Results for Causal and Causal Temporal Triplets 32

List of Figures

4 Causal Graphical Model of Events. E_1 temporally precedes E_2, and z is trajectory variable, which assumes a values t where $t \in$ All trajectories from start to E_1 . . . 26

5 Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models 29

6 Qualitative examples for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). 30

7 Input prompt formats for the Δ estimation (temporal) via Language models. 30

8 Input prompt formats for computing the causal estimand using the autoregressive open-weight models (using backdoor criterion) 31

9 The *"observational graph"* for the activity `Baking a Cake`. 33

10 The *"observational graph"* for the activity `Going Grocery Shopping`. 34

11 The *"observational graph"* for the activity `Going on a Train`. 35

12 The *"observational graph"* for the activity `Planting a Tree`. 36

13 The *"observational graph"* for the activity `Riding on a Bus`. 37

https://doi.org/10.52202/079017-0167

A COLD Framework Details

The **COLD** framework consists of Observational Distributions represented in the form of DAGs ($\mathcal{G}_o$) along with the corresponding causal graphs ($\mathcal{G}_c$) governing the dependency of occurrence between the events. Table 2 highlights the total number of causal queries that can be created using the framework. Table 1 shows a qualitative comparison with the COPA dataset Gordon et al. [2012] and the triplet samples coming from the **COLD** framework.

A.1 Adherence to SUTVA

In causal literature, a fundamentally acknowledged *Stable Unit Treatment Value Assumption* (SUTVA) [Cox, 1958, Rubin, 1980]) requires that for each unit (e.g., sequence of events), there is only one version of the non-treatment, i.e., for an event in the sequence, there lie only two versions occurring and not occurring. SUTVA plays a vital role in causal inference by ensuring that each unit's treatment assignment has a consistent impact, facilitating the accurate estimation of treatment effects. Although, in the past, researchers have created various datasets that capture the causal relationship between real-world events [Gordon et al., 2012, Du et al., 2022], the problem of achieving the SUTVA assumption has remained challenging. For example, given events (taken from the COPA dataset [Gordon et al., 2012]) E_1: "`The teacher assigned homework to students`" and E_2: "`The students groaned,`" it becomes challenging to define $\neg E_1$ since there are enormous possibilities that may have occurred at the same time (in place of E_1) that negates E_1, making it difficult to define an event of not having done something. Recent work by Zhang et al. [2022b], proposes to use multiple alterations of events for capturing $\neg E_1$, violating the SUTVA assumption. In this work, we highlight that if we define a closed system, capturing a commonsense activity, it facilitates adherence to SUTVA assumptions as closely as possible. For example, in the activity of "`going via an airplane,`" one would have either "`checked-in the luggage`" (E_1) or "`skipped checking-in luggage`" due to smaller bags ($\neg E_1$). Moreover, developing a causal setup with observations has always been a challenging problem in the wild and often requires few assumptions, as the strong causal link can only be established in an ideal world where randomized controlled trials (RCTs) are feasible. In our framework, adhering to SUTVA comes naturally where, in a trajectory, an occurrence of an event can be intervened to obtain an alternate trajectory, reaching an ideal setup facilitating causal reasoning in daily commonsensical activities.

A.2 Comparison with previous Causal Reasoning Datasets/Benchmarks

Table 5 shows a broad overview of the existing causal Dataset/Benchmarks presented in the NLP community. We find that most of the existing set of work relies on real-world events to reason about causality in NLP, where human annotators are asked to reason causally between the nature of events. However, most of these datasets/benchmarks try to establish a connection using a simple question prompt, which may not be enough to construct the underlying causal graph. Moreover, most of the real-world grounding-based methods remain open-ended due to the events taking place in the wild, making it difficult to consider constructing a causal graph where multiple variables play a role. More recently, with increased research attention on the causal reasoning abilities of LLMs, researchers have tried framing causal queries based on a causal inference engine, requiring the underlying causal graphs. However, when constructing causal queries from prompting LLMs, natural language is used to verbalize the causal concepts in the form of symbolic variables that may not have a real-world grounding.[2] Moreover, the created causal queries are difficult for a human with little or no knowledge of causal inference concepts.[3] Table 5 shows a comparison of all these features in detail, where **COLD** satisfies all the features.

We realize this is a first-of-its-kind framework built over real-world events and contains the underlying causal graph. Having both the Observational Distribution (representing the enormous event sequences present in daily activity) and the manually created underlying causal graph helps facilitate an in-depth analysis of the causal reasoning abilities of LLMs. Moreover, the same framework can further be extended in various ways: 1) Extending the number of activities: In the current version of the framework, we only consider 5 daily activities to provide an in-depth analysis. In the future, those can be extended to incorporate more such activities. 2) Extending the scope of activities: The tasks

[2]https://huggingface.co/datasets/causalnlp/corr2cause

[3]https://huggingface.co/datasets/causalnlp/CLadder

https://doi.org/10.52202/079017-0167

Table 5: **Comparison of causal experimental settings used in prior LLM evaluation benchmarks.** The real-world grounding plays a crucial role in evaluating LLMs, which is not present in the symbolic benchmarks.

Datasets/Benchmarks	Real-World	Causal Graph	Symbolic	Exhaustive	# Samples
SemEval2021 Task8 [Hendrickx et al., 2010]	✓	✗	✗	✗	1331
EventCausality [Do et al., 2011]	✓	✗	✗	✗	414
COPA [Gordon et al., 2012]	✓	✗	✗	✗	1000
Causal-TimeBank [Mirza et al., 2014]	✓	✗	✗	✗	-
CaTeRS [Mostafazadeh et al., 2016]	✓	✗	✗	✗	320 stories (1.6K sent)
BECauSE [Dunietz et al., 2017]	✓	✗	✗	✗	-
Event2Mind [Rashkin et al., 2018]	✓	✗	✗	✗	25K event phrases
ATOMIC [Sap et al., 2019a]	✓	✗	✗	✗	877K
SocialIQA [Sap et al., 2019b]	✓	✗	✗	✗	37K
TimeTravel [Qin et al., 2019]	✓	✗	✗	✗	81.4K
Abductive (ART) [Bhagavatula et al., 2020]	✓	✗	✗	✗	20K narr, 200K expl
Com2Sense [Singh et al., 2021]	✓	✗	✗	✗	4k sentence pairs.
TellMeWhy [Lal et al., 2021]	✓	✗	✗	✗	30K questions
CRASS [Frohberg and Binder, 2022]	✓	✗	✗	✗	274 PCT
e-CARE [Du et al., 2022]	✓	✗	✗	✗	20K CR questions
CausalQA [Bondarenko et al., 2022]	✓	✗	✗	✗	1.1 Million
COLA [Wang et al., 2023]	✓	✗	✗	✗	1,360 event pairs
CRAB [Romanou et al., 2023]	✓	✗	✗	✗	2.7K pairs
CausalJudgement [Srivastava et al., 2023]	✓	✓	✗	✗	-
Corr2Cause [Jin et al., 2023]	✗	✓	✓	✓	200K
CausalParrots [Zečević et al., 2023]	✓	✓	✓	✗	-
CLadder [Jin et al., 2024]	✗	✓	✓	✗	10K
COLD (ours)	✓	✓	✓	✓	∼9.52 Million

used in the activities are generic and capture commonsense; for validating domain-specific causal reasoning abilities, the framework could be extended to domain activities, for example, cooking a specific recipe where adding different ingredients causes a variation in taste. 3) Extending the type of causal queries: while constructing the causal queries, we considered the simplest task of finding the more plausible cause/effect given two options as events, keeping only unconditional $d-$separation as the primary condition. The framework can directly be extended, keeping causal queries inspired from Jin et al. [2023, 2024].

On the analysis front, we realize that the possibilities of an in-depth analysis increase by a significant margin. In this work, we shed light on a few mechanisms for validating causal reasoning abilities via zero-shot CCR (compared to previous works that rely on training and further testing on similar datasets). We specifically focused on open-weight models for better applicability in the future and proposed a few mechanisms for estimating the causal relationships between the events. This opens up several possible avenues for an in-depth analysis of LLMs.

A.3 Observational Graphs

Fig. 9, Fig. 10, Fig. 11, Fig. 12, and Fig. 13 shows the *"observational graphs"* for the activity `Baking a Cake`, `Going Grocery Shopping`, `Going on a Train`, `Planting a Tree`, and `Riding on a Bus` respectively.

B Algorithms in the COLD Framework

In this section, we provide insights into the Algorithms used in the **COLD** framework. We start with Algorithm 1, which creates causal query triplets given the observational graphs $\mathcal{G}_o$ and along with the Causal Graphs $\mathcal{G}_c$.

Remark: Temporal precedence is generally assumed essential for defining causation, and it is one of the most important clues that is used to distinguish causal from other types of associations [Mill, 1898, Hill, 1965, Pearl and Verma, 1995]. For our framework, we also consider the topologically sorted order obtained from the observational graphs and use the temporal order to define the causal query triplets, i.e., the cause events will always precede the effect events.

Creating Causal Query Triplets: The Algorithm 1 is designed to sample all the possible causal query triplets to construct a dataset for validating causal reasoning ability over an activity. Provided the observational graphs $\mathcal{G}_o$ and the Causal Graphs $\mathcal{G}_c$ for an activity, we first sample all the possible node triplets in the graph. Later, we iterate over the set of triplets and check if one of the nodes in

Table 6: **Human validation done for a small sample of 100 causal query triplets.** Overall we find that humans do perform well in causal reasoning about these daily activities.

Human Annotators	`cake`	`shopping`	`train`	`tree`	`bus`	Average
Subject 1	95	95	90	100	90	94
Subject 2	100	100	90	95	90	95
Subject 3	100	100	85	85	70	88
Subject 4	100	100	95	90	85	94
Subject 5	100	85	95	90	80	90
Average	99.00	96.00	91.00	92.00	83.00	**92.20**

the triplet (n_i, n_j, n_k) is d-separated. Further, the d-separated node becomes the wrong choice. The remaining two events become premise and correct choices, depending on their temporal order. For example, if n_i is the node that is d-separated from n_j and n_k, we check if n_j and n_k have a causal link between them in $\mathcal{G}_o$. If n_j and n_k are found to have a causal link, we create two triplets using the temporal ordering between n_j and n_k. The temporal link $n_j \rightarrow n_k$ leads to an 'effect' query where n_j becomes 'premise' and n_k becomes 'correct choice,' and a 'cause' query where n_j becomes 'correct choice' and n_k becomes the 'premise.' Note in Algorithm 1 [Store tuple], we only show one such instance for brevity; the implementation will consist of another mirror instance (i.e., for every [Store tuple] both 'cause' and 'effect' question triplets are stored to the dataset).

The understanding of these activities is generic, and LLMs do provide a suitable set of sequences when prompted to generate a list of steps to complete the activity. The constructed dataset helps to rigorously validate the understanding of the activity through an enormous number of causal query triplets. The results show that although the LLMs can explain the activity in detail, including generating correct steps for performing tasks, causally reasoning about the set of events remains challenging.

Human validation: To get a rough estimate of the human performance on the created causal reasoning queries, we also perform a small-scale human study, where the annotators are given a set of randomly chosen 100 causal queries. The human subjects were graduate students of computer science who were given a brief tutorial about counterfactual reasoning. Table 6 shows the obtained results. We would like to mention that validating human performance is challenging due to the nature of the causal reasoning task. The nature of counterfactual reasoning requires the human/algorithm to assume a similar alternate world/universe with only a particular happening or not happening to approximate the causal strength between the events. These imaginations can be expressed in statements as highlighted by Pearl and Mackenzie [2018], Pearl et al. [2016], containing an "if" statement in which the "if" portion is untrue or unrealized (aka counterfactual). The "if" portion of a counterfactual is called the hypothetical condition, or more often, the antecedent, making it challenging (cognitively heavy) to conduct a human evaluation. Please note that the study is performed using only a small sample of 100 causal query triplets out of thousands of queries, and the presented results only provide a rough estimate that may not generalize for a larger number of queries. Hence, a comparison of human study results and LLMs is not fair, and the presented human performance estimates may not be true representative of the entire population. Interactions with human subjects also revealed that they tend to confuse temporality and causality (similar findings were reported by Do et al. [2011]).

Evaluating over Causal Query Triplets: Algorithm 2 makes use of the causal estimands Δ to compare the causal strength between the premise event and the choice events. We consider the causal estimand computed between the premise and the available set of choices and predict the label corresponding to the high Δ values. For a given causal query from the created causal query triplet dataset $\mathcal{D}$, where each data entry $\mathcal{D}_i$ corresponds to (p, c_1, c_2, q, l), i.e., premise event, choice 1, choice 2, question and the label respectively. As the task is to predict the more plausible cause/effect of the given premise event, we create two event pairs, (p, c_1) and (p, c_2), and compute the causal estimand Δ for both the pairs using the temporal or the backdoor scheme (described below in Algorithm 3). Note that the order of events given to $\Delta_{\mathcal{M}}$ is in E_1 and E_2 format, i.e. $\Delta_{\mathcal{M}}(E_1, E_2)$. Using the temporal precedence (highlighted as remark above), the cause event will always precede the effect event temporally. Hence, for a causal query with the question as 'cause', the causal estimand is estimated as $\Delta_{\mathcal{M}}(c_1^i, p^i)$, $\Delta_{\mathcal{M}}(c_2^i, p^i)$ and $\Delta_{\mathcal{M}}(p^i, c_1^i)$, $\Delta_{\mathcal{M}}(p^i, c_2^i)$ when the question is 'effect.' Further, based on the estimated $\Delta_{\mathcal{M}}$ scores, the more plausible cause/effect is predicted.

https://doi.org/10.52202/079017-0167

Algorithm 1 Creating Causal Query Triplets

$\mathcal{G}_c$: Causal Graphical Model; $\mathcal{G}_o$: Observational Graph;
$\mathbf{d_s}(\mathbf{G}, \mathbf{x}, \mathbf{y})$: True iff $(\mathbf{x}, \mathbf{y})$ are *d-separated* unconditionally in any DAG $\mathbf{G}$;
$\mathbf{d_c}(\mathbf{G}, \mathbf{x}, \mathbf{y})$: True iff $(\mathbf{x}, \mathbf{y})$ are *d-connected* unconditionally in any DAG $\mathbf{G}$;
$\mathbf{genSamples}(\mathbf{G})$: Generates all possible node triplets (n_i, n_j, n_k) from DAG $\mathbf{G}$ such that $i \leq j \leq k$, where (i, j, k) are the respective indices of nodes in a topologically sorted list of nodes;
$\mathbf{A}(\mathbf{G}, \mathbf{x}, \mathbf{y})$: True iff $\mathbf{x}$ is an ancestor of $\mathbf{y}$ in DAG $\mathbf{G}$.
$\oplus$: Exclusive OR operator;
p: premise; c_1: choice 1; c_2: choice 2; q: question; l: answer (label)
$(p, c_1, c_2, q, l) \in \mathcal{D}$
$\mathcal{D} = [\ \ ]$ [Empty Dataset]
$\mathcal{S} \sim \mathbf{genSamples}(\mathcal{G_o})$ [Generate Samples]
for (n_i, n_j, n_k) in $\mathcal{S}$
  if $\mathbf{d_s}(\mathcal{G}_c, n_i, n_j)$ **then**
    if $\mathbf{d_c}\ (\mathcal{G}_c, n_i, n_k) \oplus \mathbf{d_c}\ (\mathcal{G}_c, n_j, n_k)$
      $q =$ 'cause'
      $l = \arg\max[\mathbf{d_c}(\mathcal{G}_c, n_i, n_k), \mathbf{d_c}(\mathcal{G}_c, n_j, n_k)]$
      $p, c_1, c_2 \leftarrow n_k, n_i, n_j$
      if $\mathbf{A}(\mathcal{G}_c, c_l, p)$ **then** APPEND$(\mathcal{D}, [(p, c_1, c_2, q, l))$ [Store tuple]
  else if $\mathbf{d_s}(\mathcal{G}_c, n_j, n_k)$ **then**
    if $\mathbf{d_c}(\mathcal{G}_c, n_i, n_j) \oplus \mathbf{d_c}(\mathcal{G}_c, n_i, n_k)$
      $q =$ 'effect'
      $l = \arg\max[\mathbf{d_c}(\mathcal{G}_c, n_i, n_k), \mathbf{d_c}(\mathcal{G}_c, n_i, n_j)]$
      $p, c_1, c_2 \leftarrow n_i, n_j, n_k$
      if $\mathbf{A}(\mathcal{G}_c, p, c_l)$ **then** APPEND$(\mathcal{D}, (p, c_1, c_2, q, l))$ [Store tuple]
return $\mathcal{D}$

Computing $\Delta_{\mathcal{M}}$: Algorithm 3 depicts the process of computing an unbiased estimate for the causal estimand. The causal strength is computed between two events E_1 and E_2 where E_1 is assumed to be preceding E_2 temporally. To make an unbiased estimate based on the provided options, we consider normalizing the obtained probability scores by flipping the options and providing the same query prompt to the Language Model.

$$f_{\mathcal{M}}(E_1, E_2, \phi) \leftarrow \frac{s_{\mathcal{M}}(E_1, E_2, \phi) + s_{\mathcal{M}}(E_1, E_2, \phi_f)}{s_{\mathcal{M}}(E_1, E_2, \phi) + s_{\mathcal{M}}(E_1, E_2, \phi_f) + \tilde{s}_{\mathcal{M}}(E_1, E_2, \phi) + \tilde{s}_{\mathcal{M}}(E_1, E_2, \phi_f)},$$

where ϕ denotes the prompt template as shown in Figure 8 (top) and ϕ_f denotes the same prompt with flipped options, Figure 8 (bottom). The overall equation helps normalize the prediction probabilities of the 'Increase' option by using the probabilities of the 'Decrease' option. Finally, these normalized scores are computed for multiple trajectories t_i in the backdoor adjustment scheme to compute the causal estimands $p_{\mathcal{M}}(E2 \mid do(E_1))$ and $p_{\mathcal{M}}(E2 \mid do(\neg E_1))$ that help estimate the causal strength $\Delta_{\mathcal{M}}$ between the events E_1 and E_2.

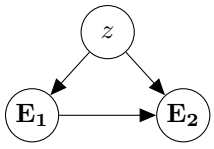

Figure 4: Causal Graphical Model of Events. E_1 temporally precedes E_2, and z is trajectory variable, which assumes a values t where $t \in$ All trajectories from start to E_1

C Backdoor Adjustments

A set of variables W satisfies the backdoor criterion relative to T and Y if the following are true

(A) W blocks all backdoor paths from T to Y i.e. blocking confounding or non-causation association paths

(B) W doesn't contain any descendants of T

https://doi.org/10.52202/079017-0167

Algorithm 2 Evaluating Causal Query Triplets

$\mathcal{T}_n$: n unique trajectories from Start node (start of the activity) to node E_1

$p_{\mathcal{M}}(E_2|do(E_1)) = \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} \hat{p}(E_2|E_1, z = t)$

$p_{\mathcal{M}}(E_2|do(\neg E_1)) = \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} \hat{p}(E_2|\neg E_1, z = t)$

$\Delta_{\mathcal{M}}$: Returns **Average treatment effect** $(p_{\mathcal{M}}(E_2|do(E_1)) - p_{\mathcal{M}}(E_2|do(\neg E_1)))$ to determine the causal effect of event E_1 on event E_2

p: premise; c_1: choice 1; c_2: choice 2; q: question; l: label

$\mathcal{D}$: Set of all causal queries

$\mathcal{D}_i$: Causal query (p, c_1, c_2, q, l), where $(p, c_1, c_2, q, l) \in \mathcal{D}$

for $\mathcal{D}_i$ in $\mathcal{D}$ **do**

  if q^i == 'cause' **then**

    prediction $\leftarrow \arg\max \left[\Delta_{\mathcal{M}}(c_1^i, p^i), \Delta_{\mathcal{M}}(c_2^i, p^i)\right]$

  else if q^i == 'effect' **then**

    prediction $\leftarrow \arg\max \left[\Delta_{\mathcal{M}}(p^i, c_1^i), \Delta_{\mathcal{M}}(p^i, c_2^i)\right]$

  end if

  $\eta \leftarrow \eta + \mathbb{1}(prediction == l_i)$

end for

return $\eta/|\mathcal{D}|$

Algorithm 3 Computing Causal Estimand

E_1, E_2 : Events in an given activity

$\mathcal{T}_n$: Set of n trajectories (temporally ordered sequence of events)

$s_{\mathcal{M}}(E_1, E_2, \phi) \leftarrow$ score of token **A** (associated with option '$increase$') using prompt ϕ of model $\mathcal{M}$.

$s_{\mathcal{M}}(E_1, E_2, \phi_f) \leftarrow$ score of token **B** (associated with option '$increase$') in prompt ϕ_f (flipped options in prompt ϕ) of model $\mathcal{M}$.

$\tilde{s}_{\mathcal{M}}(E_1, E_2, \phi) \leftarrow$ score of token **B** (associated with option '$decrease$') using prompt ϕ of model $\mathcal{M}$.

$\tilde{s}_{\mathcal{M}}(E_1, E_2, \phi_f) \leftarrow$ score of token **A** (associated with option '$decrease$') in prompt ϕ_f (flipped options in prompt ϕ)of model $\mathcal{M}$

Norm. Score $f_{\mathcal{M}}(E_1, E_2, \phi) \leftarrow \frac{s_{\mathcal{M}}(E_1, E_2, \phi) + s_{\mathcal{M}}(E_1, E_2, \phi_f)}{s_{\mathcal{M}}(E_1, E_2, \phi) + s_{\mathcal{M}}(E_1, E_2, \phi_f) + \tilde{s}_{\mathcal{M}}(E_1, E_2, \phi) + \tilde{s}_{\mathcal{M}}(E_1, E_2, \phi_f)}$

$\hat{p}(E_2|E_1, z = t) \leftarrow f_{\mathcal{M}}((t, E_1), E_2, \phi)$

$\hat{p}(E_2|\neg E_1, z = t) \leftarrow f_{\mathcal{M}}((t, \neg E_1), E_2, \phi)$

$p_{\mathcal{M}}(E_2|do(E_1)) \leftarrow \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} \hat{p}(E_2|E_1, z = t)$

$p_{\mathcal{M}}(E_2|do(\neg E_1)) \leftarrow \frac{1}{|\mathcal{T}_n|} \sum_{t \in \mathcal{T}_n} \hat{p}(E_2|\neg E_1, z = t)$

$\Delta_{\mathcal{M}} \leftarrow p_{\mathcal{M}}(E_2|do(E_1)) - p_{\mathcal{M}}(E_2|do(\neg E_1))$

return $\Delta_{\mathcal{M}}$

Then, W satisfies the backdoor criterion [Pearl et al., 2016, Neal, 2020].

Adhering to the above conditions of the backdoor criterion, it is reasonable to assume that the trajectory t (temporally ordered sequence of events) till E_1 will contain the events that confound the event E_1 and event E_2 (condition A). Every event trajectory till E_1 will temporally precede E_1(condition B). Hence, the trajectory variable will satisfy backdoor criteria in the proposed closed system. The domain of the trajectory variable is a set of all trajectories till E_1. Therefore, conditioning on t closes all paths that induce non-causal associations. The generic representation of an approximate causal graphical model involving E_1, E_2, and t is shown in Figure 4, and can be formulated as:

$$p_{\mathcal{M}}(E_2|do(E_1)) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \hat{p}(E_2|E_1, z = t)$$

Where $\mathcal{T}$ is a set of all trajectories from the start of the activity till the event E_1(excluding E_1).

https://doi.org/10.52202/079017-0167

D Experiments and Results

D.1 Compute Resources

We perform all the experiments using a machine with 5 NVIDIA A100 GPUs. We use only the open-weight models with frozen parameters to present the results for better reproducibility in the future.

D.2 Evaluation using Average Treatment Effect (ATE)

Establishing Causal Relationships: To validate the causal reasoning ability, the MCQA-based approach can be further extended to estimate the causal estimation and denote the causal strengths between the events. Establishing cause-and-effect relationships can be achieved through various statistical analyses. The strength of cause-and-effect relationships is approximated by statistically analyzing events' behavior using observational data (PC-Algorithm, Spirtes et al. [2000b]). Moreover, some of the recent works [Wang et al., 2023] highlight the role of context in determining the causal relationships between the events. To extend our analysis of causal reasoning abilities in the proposed framework, we use the backdoor adjustments in LLMs as explained in the main paper. Moreover, we also perform an interesting analysis of the observational graphs for estimating Δ statistically.

1) Through Original Trajectories: DeScript Wanzare et al. [2016] collects data by considering ~ 100 ESDs written by different crowd-sourced workers. We use the original Trajectories (ESDs) $\mathcal{T}_o$ written by humans present in the DeScript dataset. These ESDs provide the original flow in the graph directly coming from crowdsourced workers. We consider these as the original trajectories $\mathcal{T}_o$. Applying the backdoor criterion (Eq. 2) over these trajectories $\mathcal{T}_o$. An interventional distribution similar to the previous section is computed considering the likelihood of occurrence of E_2 under each treatment (E_1 and $\neg E_1$) for only these trajectories $\mathcal{T}_o$. These estimations are further used to compute the treatment effect using the Eq. 1. We denote the causal risk difference (Δ) computed with $\mathcal{T}_o$ as Δ_o.

2) Through Observational Graphs: The observational graphs provide a proxy for the underlying knowledge about the activity, covering all possible sets of events, i.e., starting from the start node, one can trace multiple trajectories that will essentially define the way of performing the activity. For every pair of connected events (e_i, e_j), the edge between them represents the probable transition from e_i to e_j with some non-zero probability. However, a noteworthy point is that the transition probability between two connected events (e_i, e_j) can vary depending on the design choices/transition function $T(e_i, e_j) \rightarrow (0, 1]$. We define this transition function in two ways: 1) **Uniform Node Transition** (T_n): The transition probability from current node e_i to next probable events $e_j \in E_{ij}$ would be uniform that is $T(e_i, e_j) = \frac{1}{|o_i|}$, where $|o_i|$ represents number of outgoing edges from event e_i. (i.e., assuming after an event, the choice of the next event is uniform from the possible events). 2) **Uniform Trajectory Transition (T_t):** Another way to take the set of events in an activity (trajectory) is by considering all the possible paths being equally probable, i.e., across the entire population the same activity will be represented with one of the possible trajectories. Hence we can define the transition function with each trajectory $t_i = (e_{start}, e_2, ..., e_{end})$ (sequence of events from starting to ending) having the same probability, i.e.:

$$p(t_i) = p(t_j) \forall t_i, t_j \in \mathcal{T}$$
$$p(t_i) = \prod_{l,m \in t_i} T_t(e_l \rightarrow e_m)$$

Further, given a transition function T, computation of $\hat{p}(E_2|E_1, z = t_i)$ becomes straightforward as $p(E_2|E_1)$, since the next course of trajectories after E_1 will be decided given E_1 has occurred. Analytically, it can be computed by counting every trajectory (i.e., $\mathcal{T}_{ij}$) from E_1 that leads to E_2

$$p(E_2|E_1) = \sum_{t \in \mathcal{T}_{ij}} \prod_{l,m \in t} T(e_l \rightarrow e_m) \tag{4}$$

$$p(E_2|E_1) = \left[\sum_{k=1}^{k=M} T^k \right] (E_1 \rightarrow E_2) \tag{5}$$

For estimating the probability $\hat{p}(E_2|\neg E_1, z = t_i)$, we make use of the observational graph by considering all the parent nodes of E_1 and compute the probability of reaching E_2 from the parent (i.e., last event of trajectory t_i) avoiding the occurrence of E_1.

$$\hat{p}(E_2|\neg E_1, z = t_i) = \sum_{\substack{t \in \mathcal{T}_{ij} \\ E_1 \notin t}} \prod_{l,m \in t} T(e_l \rightarrow e_m) \tag{6}$$

also, $p(z = t_i)$ can be computed as product of each transition in t_i, i.e

$$p(z = t_i) = \prod_{ij} T(e_i \rightarrow e_j) \tag{7}$$

Computations from Equations 5, 6, and 7 are used in the backdoor adjustment defined in Equation 2 for estimating the interventional likelihood of occurrence of E_2 under each treatment (E_1 and $\neg E_1$) and causal risk difference (Δ). Note depending on the choice of transition function (T_n or T_t), we obtain two deltas Δ_n and Δ_t, respectively.

E Prompt Templates for Language Model based Experiments

We present the various prompt templates used to estimate the temporal link between the events in Figure 7. For BERT-based models, we use the MLM-trained models for predicting the masked token given a sentence (Previously, a similar approach was adopted by Zhang et al. [2022b]). In contrast, for autoregressive models, we frame the prompt as a question-answer objective, taking inspiration from [Robinson and Wingate, 2023], where a multiple-choice-based question is framed to predict the answer in the form of the option IDs. The prompt is intentionally structured so that the LLM is intended to predict a single token (Such as "A", "B", etc.). Robinson and Wingate [2023] highlights the advantages of MCQA-based evaluation over cloze evaluation (where the LLMs are expected to generate the entire answer in a cloze test), leading to a significant boost in various tasks, including commonsense-based tasks.

For our prompt-based evaluation experiments over the generated causal triplets, we follow the same MCQA-based strategy and frame the prompts accordingly for a fair evaluation. Figure 5 presents various prompt templates for autoregressive experiments, and Figure 6 shows a few qualitative examples for the framed causal query templates.

```
Consider the activity of activity name.
[ in-context examples (if few-shot/in-context learning experiment) ]
Which of the following events (given as options A or B) is a plausible question
(cause/effect) of the event premise?
A. choice1
B. choice2
Answer: A
```

```
The following are multiple choice questions about activity name. You should directly
answer the question by choosing the correct option.
[ in-context examples (if few-shot/in-context learning experiment) ]
Which of the following events (given as options A or B) is a plausible question
(cause/effect) of the event premise?
A. choice1
B. choice2
Answer: A
```

Figure 5: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input. The `orange text` is the input from the created causal query triplets, where the `activity name` denotes the description of the activity like `baking a cake`. The next-token prediction probabilities of the option IDs at the `red text` is used as the observed prediction distribution.

https://doi.org/10.52202/079017-0167

```
Consider the activity of baking a cake.
Which of the following events (given as options A or B) is a plausible effect of the event
preheat oven to 350 degrees.?
A. turn off oven.
B. prepare the microwave oven and required utensils
Answer: A
```

```
The following are multiple choice questions about going on a train. You should directly
answer the question by choosing the correct option.
Which of the following events (given as options A or B) is a plausible cause of the event
get the bill for groceries. ?
A. pay the cashier for your items.
B. place cart into cart corral.
Answer: A
```

Figure 6: Qualitative examples for the MCQA-based evaluation of autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.).

```
In terms of 'before' and 'after', the event: "first event text" would have happened
<mask_token> the event: "second event text"
```

```
Consider the activity of activity name.
Question: Determine the temporal order.
The following events took place: 1. first event text, 2. second event text
Did the first event occur 'before' or 'after' the second event? (choose from the given
options)
A: before
B: after
Answer: A
```

Figure 7: Input prompt formats for the Δ estimation via Language models. The first row shows the prompt template used for BERT-based language models, where the mask token is predicted. The second row shows the template for autoregressive open-weight models (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input. The `orange text` is the input from the created causal query triplets, where the `first event text` and `second event text` comes from the premise and available set of choices. The mask-token prediction probabilities of 'before' and 'after' and next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution for BERT-based and GPT-based open-weight models.

F Additional Results

F.1 Temporally Plausible Choices in Causal Triplets

Some of the initial studies [Do et al., 2011] highlight the difficulty in choosing between the causal-effect events and temporal events (that occur in close proximity to the premise event), i.e., temporal relationships are sometimes considered as a causal relationship by human annotators. We also create another version of created causal triplets where the wrong choices are replaced by temporally near nodes (nodes that are at a one-hop distance from the premise node). We call these '*causally hard triplets*.' Note the temporal nodes are obtained from the observational graphs $\mathcal{G}_o$. Table 7 shows the performance comparison with causal triplets and causal-temporal triplets versions of the same queries. We observe a significant performance drop on the causal-temporal triplets version for most models, highlighting the increased confusion.

CAUSAL REASONING ANALYSIS:
Context: For the activity activity name. During the activity, the following set of sequences occurred in order:
[ordered list of events present in a Trajectory till E_1] # Trajectory ($z = \mathcal{T}_i$)
Further, the event 'event text for E_1' took place. # Intervention ($do(E_1)$)
Question: Given the above information, will the chances of the occurrence of the event 'event text for E_2' increase or decrease?
A. Increase
B. Decrease
Answer: A # $(p(E_2 \mid do(E_1), z = \mathcal{T}_i)$

CAUSAL REASONING ANALYSIS:
Context: For the activity activity name. During the activity, the following set of sequences occurred in order:
[ordered list of events present in a Trajectory till E_1] # Trajectory ($\mathcal{T}_i$)
Further, the event 'event text for E_1' did NOT take place. # Intervention ($do(\neg E_1)$)
Given the above information, will the chances of the occurrence of the event 'event text for E_2' increase or decrease?
A. Increase
B. Decrease
Answer: B # $(p(E_2 \mid do(\neg E_1), z = \mathcal{T}_i)$

Flipped options variant of the above Prompt Template

CAUSAL REASONING ANALYSIS:
Context: For the activity activity name. During the activity, the following set of sequences occurred in order:
[ordered list of events present in a Trajectory till E_1] # Trajectory ($z = \mathcal{T}_i$)
Further, the event 'event text for E_1' took place. # Intervention ($do(E_1)$)
Given the above information, will the chances of the occurrence of the event 'event text for E_2' increase or decrease?
A. Decrease
B. Increase
Answer: B # $(p(E_2 \mid do(E_1), z = \mathcal{T}_i)$

CAUSAL REASONING ANALYSIS:
Context: For the activity activity name. During the activity, the following set of sequences occurred in order:
[ordered list of events present in a Trajectory till E_1] # Trajectory ($\mathcal{T}_i$)
Further, the event 'event text for E_1' did NOT take place. # Intervention ($do(\neg E_1)$)
Given the above information, will the chances of the occurrence of the event 'event text for E_2' increase or decrease?
A. Decrease
B. Increase
Answer: A # $(p(E_2 \mid do(\neg E_1), z = \mathcal{T}_i)$

Figure 8: Input prompt formats for computing the causal estimand using the autoregressive open-weight models (using backdoor criterion) (e.g., `llama(-2)`, `GPT-J`, etc.). The `black text` is the templated input. The `orange text` is the input from the created causal query triplets, where the `activity name` denotes the description of the activity like `baking a cake`. The trajectory $\mathcal{T}_i$ is obtained using the observational graph $\mathcal{G}_o$ and contains the sequence of events before the event E_1. The next-token prediction probabilities of the option IDs at the `red text` is used as the observed prediction distribution. The flipped options variants of the prompts contain the same query with flipped options (i.e., the option 'Increase' becomes 'Decrease' and vice versa). This is done to make the causal estimand unbiased towards the predicted option token as highlighted in Algorithm 3.

https://doi.org/10.52202/079017-0167

Table 7: The table provides evaluation results of Language models over the causal and causal temporal triplets.

Triplets	Model Name	cake	shopping	train	tree	bus
causal triplets	gpt-neo-125M	50.71	50.01	49.99	50.13	50.15
	gpt-neo-1.3B	44.77	45.69	42.52	45.67	42.89
	gemma-2b	53.76	52.19	60.57	60.71	53.64
	gpt-neo-2.7B	50.00	50.01	50.00	50.01	50.00
	phi-2	85.14	83.65	77.29	82.24	71.74
	gpt-j-6B	49.59	50.02	50.29	49.92	49.93
	Llama-2-7b-chat-hf	77.92	72.41	73.48	72.40	68.21
	Mistral-7B-v0.1	77.64	69.38	68.46	72.43	69.37
	gemma-7b	81.47	82.26	77.24	80.78	70.29
	Meta-Llama-3-8B	80.79	76.46	76.08	78.21	67.39
causally hard triplets	gpt-neo-125M	50.60	49.80	49.90	50.00	50.20
	gpt-neo-1.3B	49.50	51.20	48.80	47.50	48.00
	gemma-2b	52.30	51.00	56.10	52.20	50.00
	gpt-neo-2.7B	50.00	50.00	50.00	50.00	50.00
	phi-2	80.00	74.70	67.90	87.50	66.50
	gpt-j-6B	50.20	50.00	50.30	50.00	49.80
	Llama-2-7b-chat-hf	71.60	66.60	68.30	77.00	65.40
	Mistral-7B-v0.1	69.20	63.10	64.20	67.90	62.00
	gemma-7b	76.30	76.40	69.70	89.70	63.70
	Meta-Llama-3-8B	77.30	72.20	69.30	83.20	64.30

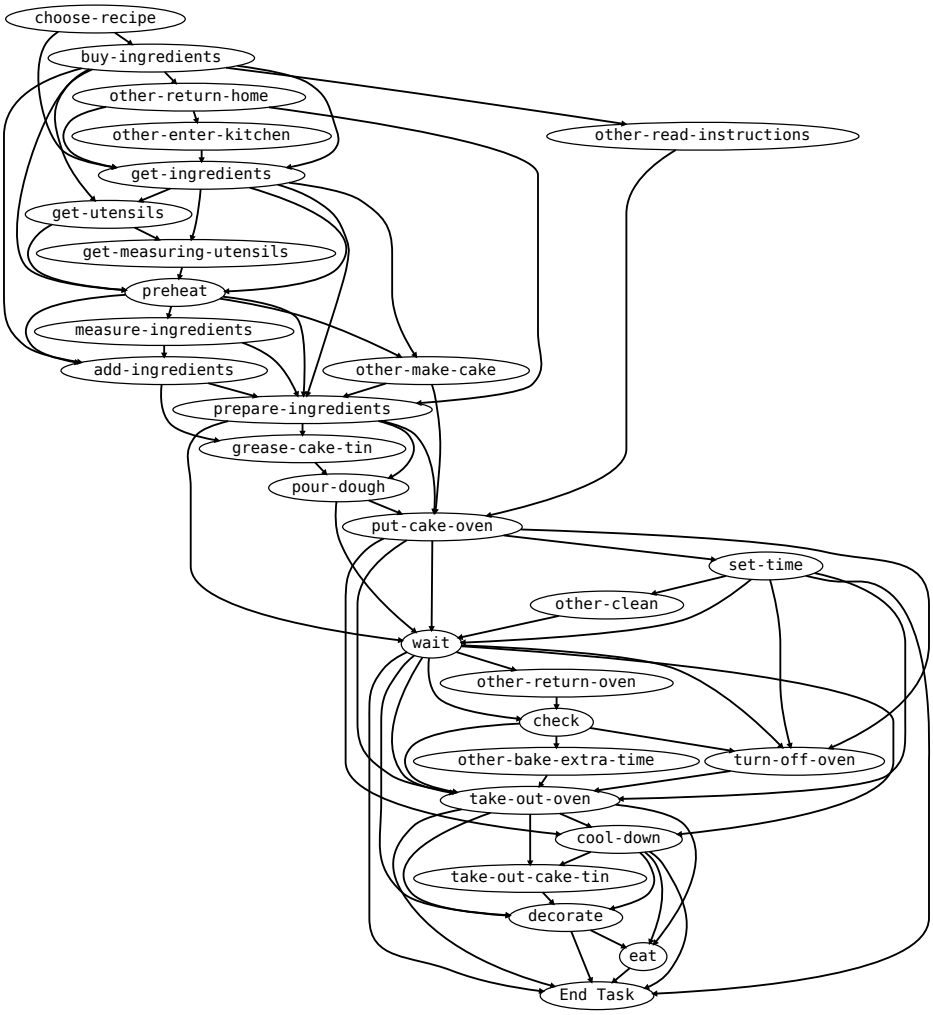

Figure 9: The figure shows the *"observational graph"* for the activity `Baking a Cake`.

https://doi.org/10.52202/079017-0167

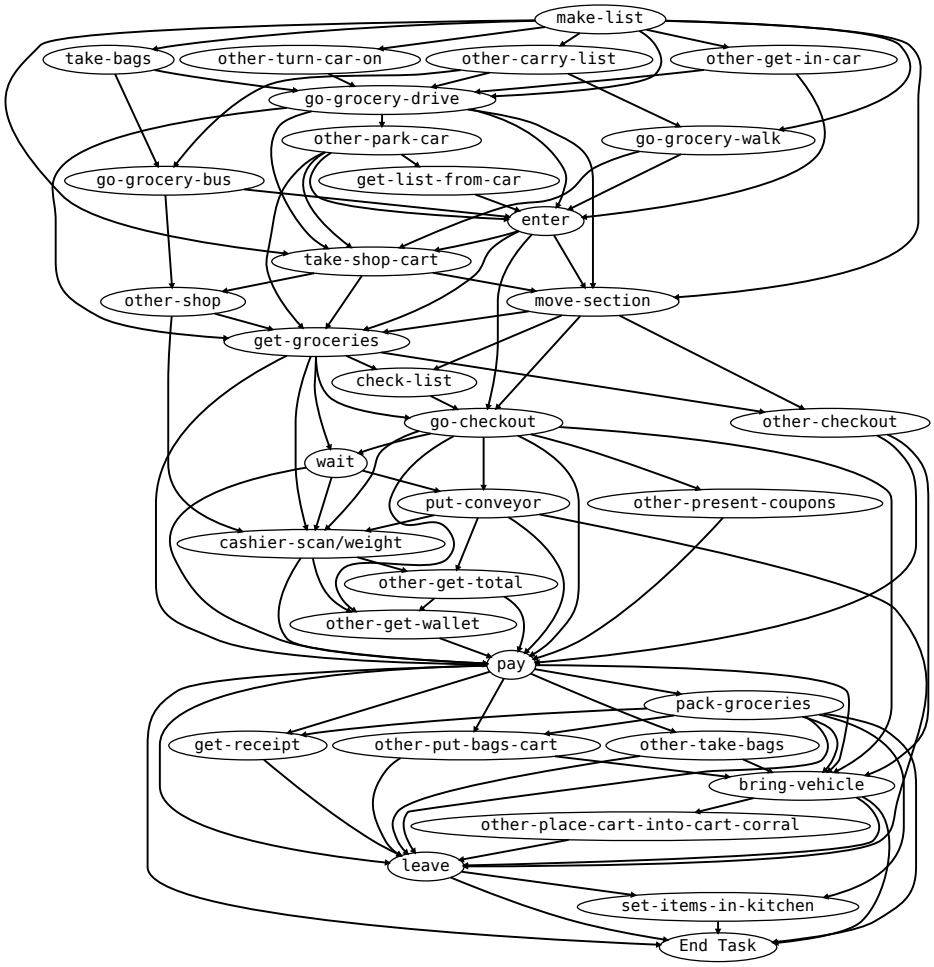

Figure 10: The figure shows the *"observational graph"* for the activity `Going Grocery Shopping`.

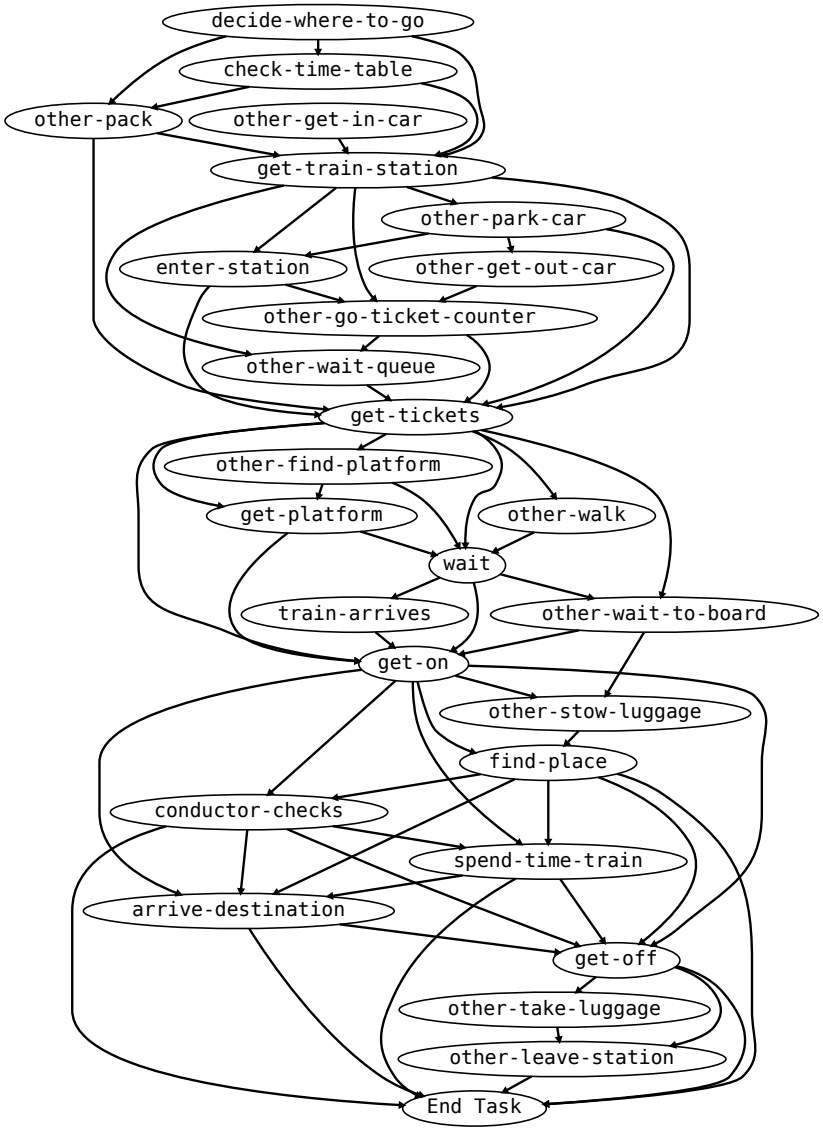

Figure 11: The figure shows the *"observational graph"* for the activity `Going on a Train`.

https://doi.org/10.52202/079017-0167

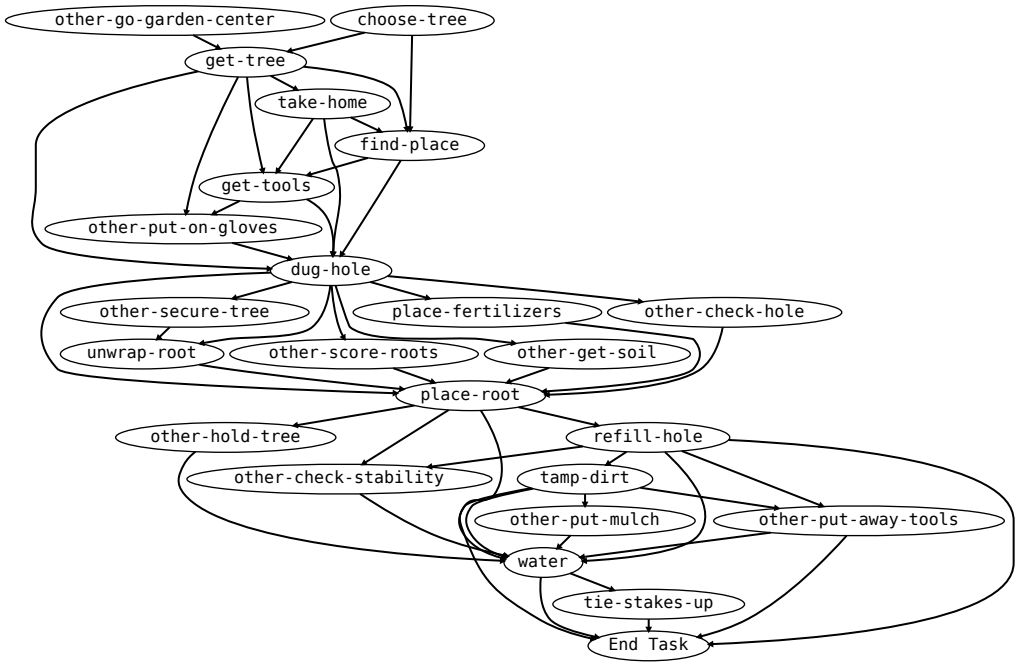

Figure 12: The figure shows the *"observational graph"* for the activity `Planting a Tree`.

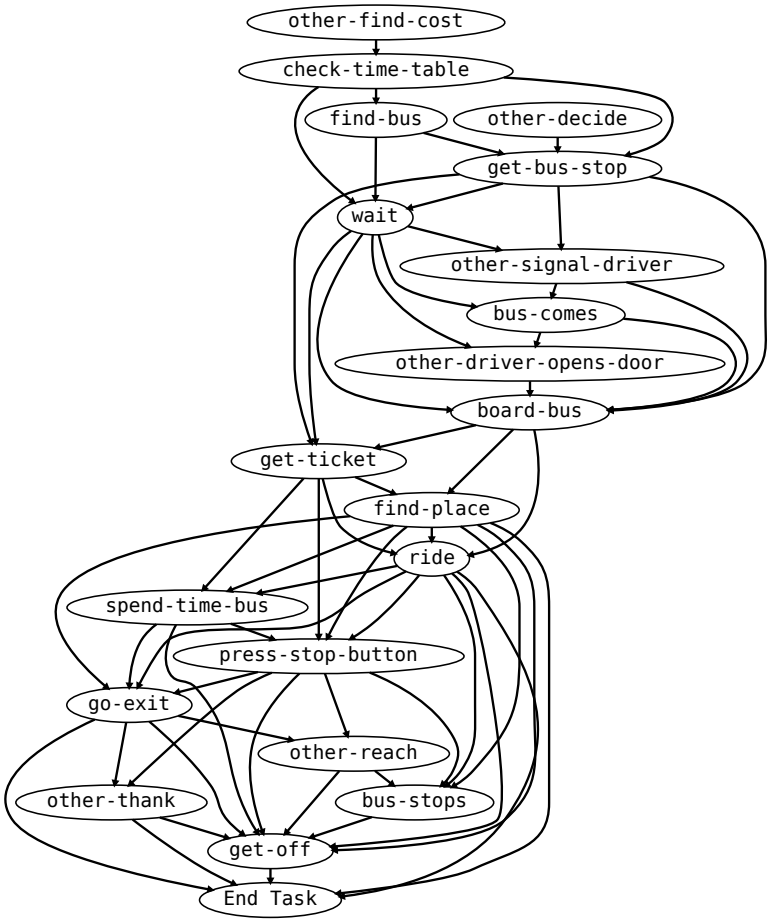

Figure 13: The figure shows the *"observational graph"* for the activity `Riding on a Bus`.

https://doi.org/10.52202/079017-0167

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide details about the main claims in the Abstract and Introduction (Section 1).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a separate section on Limitations (Sections 6)

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not have any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide details in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-0167

Answer: [Yes]

Justification: We provide details in the Introduction.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: As discussed in Section 4, we only perform evaluation on pre-trained models and do not train/fine-tune any new model.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: As discussed in Section 4, we only perform evaluation on pre-trained models and do not train/fine-tune any new model.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix Section D.1 provides details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we read the code of ethics and follow these.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: To the best of our knowledge the research proposed in the paper does not have any negative social impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-0167

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Not applicable for our paper.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have used only open-source resources and cited relevant owners of the various resources, tools and models.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not create any new asset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We do not perform any human experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable in our case.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0167