Uncertainty in Causality: A New Frontier

Shaobo Cui
EPFL
shaobo.cui@epfl.ch

Luca Mouchel
EPFL
luca.mouchel@epfl.ch

Boi Faltings
EPFL
boi.faltings@epfl.ch

Abstract

Understanding uncertainty in causality is vital in various domains, including core NLP tasks like event causality extraction, commonsense reasoning, and counterfactual text generation. However, existing literature lacks a comprehensive examination of this area. This survey aims to fill this gap by thoroughly reviewing the uncertainty in causality. We first introduce a novel trichotomy, categorizing causal uncertainty into aleatoric (inherent randomness in causal data), epistemic (causal model limitations), and ontological (existence of causal links) uncertainty. We then survey methods for quantifying uncertainty in causal analysis and highlight the complementary relationship between causal uncertainty and causal strength. Furthermore, we examine the challenges that large language models (LLMs) face in handling causal uncertainty, such as hallucinations and inconsistencies, and propose key traits for an optimal causal LLM. Our paper reviews current approaches and outlines future research directions, aiming to serve as a practical guide for researchers and practitioners in this emerging field.

1 Introduction

> *The only thing we can count on is uncertainty.* — *Albert Einstein*

Uncertainty is a fundamental aspect of scientific inquiry and practical decision-making, and the study of causality is no exception. Causal uncertainty refers to the ambiguity and unknown factors in identifying, reasoning about, or quantifying causal relationships, commonly arising from random variability, incomplete knowledge, or existential doubts. Causal uncertainty has long been recognized as critical in high-stakes domains such as financial markets (Rigotti and Shannon, 2005), medical diagnosis (Dahm and Crock, 2022), and environmental sciences (López-Gamero et al., 2011), but it also holds significant importance in natural language processing (NLP) tasks. For instance, tasks like event causality extraction (Dasgupta et al., 2018; Liu et al., 2023b), commonsense reasoning (Wang et al., 2023b; Joshi et al., 2024), and counterfactual text generation (Feder et al., 2021; Nguyen et al., 2024) often hinge on recognizing whether causal relationships in text are robust, partially known, or merely correlational (see App. A for details).

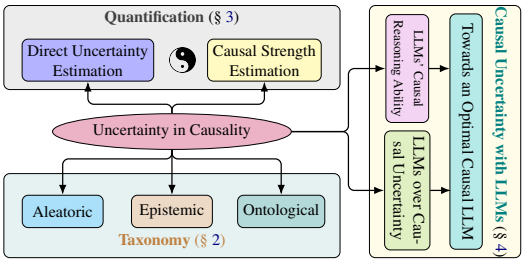


Figure 1: Overview of different aspects of causal uncertainty and their corresponding sections in the survey.

While previous surveys have explored causal reasoning (Yao et al., 2021; Liu et al., 2024b) and uncertainty in machine learning (Abdar et al., 2021; Geng et al., 2024) separately, none have systematically reviewed the intersection of these two critical areas. Though Cui et al. (2024a) use uncertainty levels as a criterion for classifying commonsense causality benchmarks and reasoning methods, they do not present a clear taxonomy or overview of uncertainty quantification methods. Motivated by this gap, we present the first systematic review of uncertainty in causality (see the overview of our survey's structure in Figure 1).

A clear taxonomy of uncertainty in causality forms the essential foundation for advancing research in this area. Unlike the commonly used dichotomy of uncertainty in machine learning (Kendall and Gal, 2017), which primarily distinguishes between aleatoric and epistemic uncertainty, the unique characteristics of causality re-

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8022–8044

quire a more nuanced approach. In § 2, we propose a trichotomy that categorizes uncertainty in causality into three types: (i) Aleatoric uncertainty (), arising from randomness in causal data; (ii) Epistemic uncertainty (), stemming from limitations in modeling causality and the model's causal knowledge; and (iii) Ontological uncertainty (), relating to the conditional existence and validity of causal links. This trichotomy distinguishes inherent randomness, model-dependent uncertainty, and existential ambiguities in causal structures, providing a clearer foundation for studying causal uncertainty.

Additionally, in § 3, we review methods for quantifying uncertainty in causality and highlight how evaluating causal strength provides a complementary perspective. Specifically, we first examine existing quantification methods that could be adapted for causal uncertainty measurement (Lakshminarayanan et al., 2017; Kuhn et al., 2023), and then we discuss established approaches for assessing causal strength measurements (Good, 1961; Suppes, 1973; Pearl, 2009) as a complementary counterpart to causal uncertainty estimation.

Finally, we examine how LLMs contend with causal uncertainty in § 4. Our analysis highlights their successes in causal discovery, inference, and counterfactual reasoning, while also exposing critical vulnerabilities—most notably, hallucinations and self-contradictions when confronted with ambiguous evidence (Gao et al., 2023; Cui et al., 2024b; Mündler et al., 2024). These deficiencies appear to arise predominantly from pattern memorization rather than from authentic causal reasoning (Zečević et al., 2023). Drawing on these insights, we propose key attributes for an optimal causal LLM, emphasizing unwavering consistency, versatility across various causal reasoning tasks, and robust modeling of causal uncertainty.

The full organization of the literature review is summarized in Figure 2. Overall, our contributions are threefold:

- We introduce a novel trichotomy for categorizing uncertainty in causality – aleatoric, epistemic, and ontological uncertainties – that extends beyond the conventional dichotomy to better address the unique complexities inherent in causal reasoning.
- We review methods for quantifying uncertainty in causality from a complementary perspective, encompassing both the direct estimation of uncertainty and the causal strength estimation, highlighting their interrelation.
- We analyze LLMs' performance under causal uncertainty, exposing issues like hallucinations and self-contradictions, and propose key traits for an optimal causal agent: robust consistency, versatile reasoning, and precise uncertainty quantification.

Paper Selection. The papers reviewed in this survey are mainly from renowned conferences and journals in ML and NLP, including but not limited to ACL, EMNLP, NAACL, ICML, NeurIPS, ICLR, AAAI, IJCAI, etc. The primary selection criteria we use are relevance to causal uncertainty, completeness, and influences of the candidate papers. The discussion about related surveys is in App. B.

2 Taxonomy of Uncertainty in Causality

The conventional dichotomy of uncertainty (Kendall and Gal, 2017), which classifies uncertainty into *aleatoric* and *epistemic* types, is not fully tailored to the complexities of causal reasoning[1]. Specifically, causal analysis (i) must determine whether an observed association reflects a genuine causal mechanism rather than mere correlation, (ii) necessarily embodies a directional influence from cause to effect, and (iii) is intrinsically concerned with the outcomes of hypothetical interventions and counterfactual scenarios. These properties motivate our extension of the classical aleatoric–epistemic dichotomy: we introduce *ontological uncertainty* to capture the existential doubt about whether a causal link is real, directional, and robust under intervention. We redefine total causal uncertainty as a combination of aleatoric, epistemic, and ontological uncertainty, providing a more comprehensive framework for understanding uncertainty in causal contexts.

2.1 Aleatoric Uncertainty

Alea, derived from the Latin word "alea", means "dice" (). It pertains to elements of chance, randomness, or unpredictability. Aleatoric uncertainty refers to inherent randomness in the data, such as variability in health outcomes among patients receiving the same treatment, and is generally irreducible even with more data. It can be further categorized into:

[1]The preliminary knowledge about uncertainty sources and uncertainty expressions is detailed in App. C

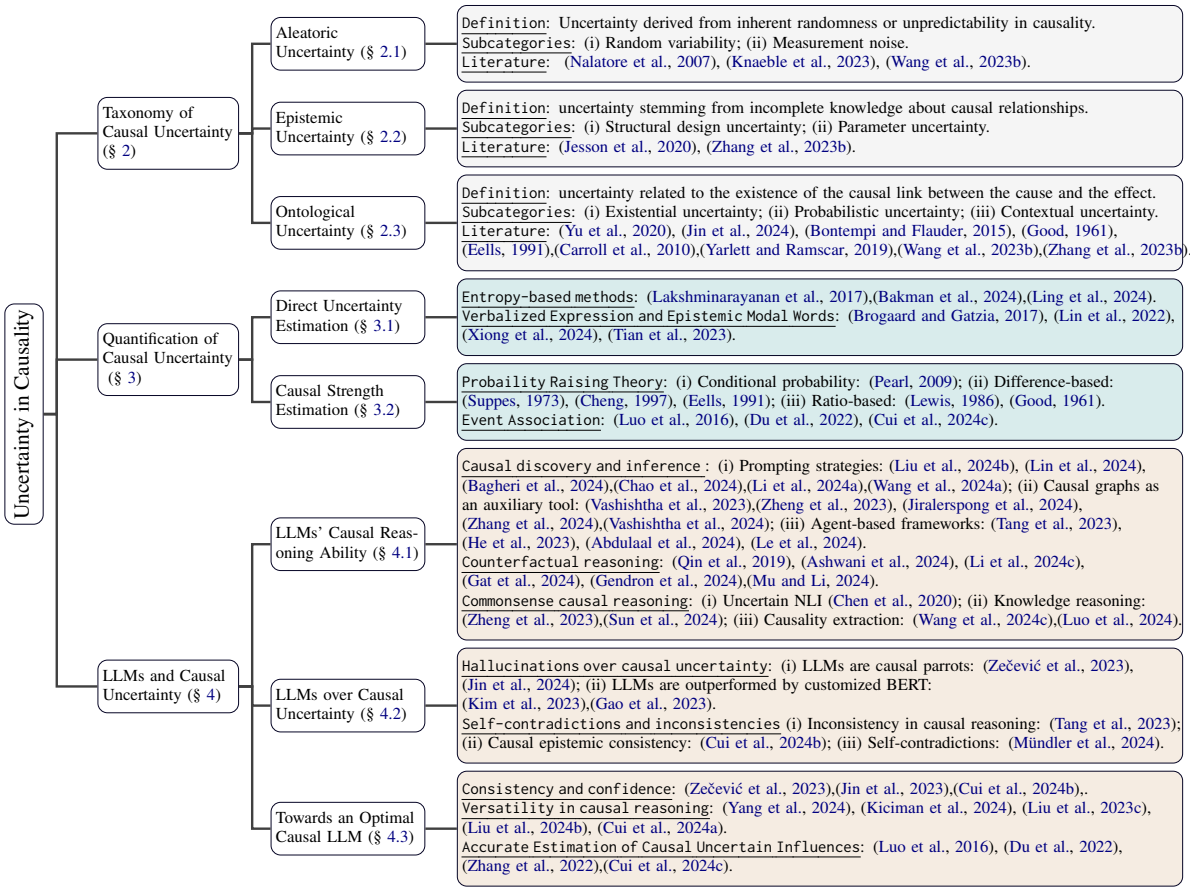


Figure 2: Overview of existing literature on causal uncertainty, including taxonomy and quantification aspects and LLMs' abilities and challenges in managing causal uncertainty.

- *Random variability*: The variability of causality occurs even under identical conditions. A typical example is that "smoking leads to lung cancer." There are situations where two individuals, both with similar smoking habits, may have different health outcomes.
- *Measurement noise*: This uncertainty is due to the limitations of measurement tools. For instance, inaccuracies in data collection or measurement tools can introduce uncertainty in quantifying causal relationships, especially in complex systems.

2.2 Epistemic Uncertainty

Episteme (), as implied by its Greek origin επιστημη means knowledge. Epistemic uncertainty arises from incomplete knowledge about causal mechanisms or limitations inherent to the model structure and parameters. Epistemic uncertainty can be further divided into:

- *Structural design uncertainty*: Uncertainty also arises from the design of the model itself. For instance, different kinds of causal model structures lead to different model prediction results even when trained with the same datasets (Zhang et al., 2023b). Choices such as model architecture, the number of neural network layers, and activation functions contribute to this type of uncertainty.
- *Parameter uncertainty*: In the era of deep learning and following LLMs, there is variability and lack of certainty in the estimates of parameters used in models. Specifically, Jesson et al. (2020) propose an approach for estimating epistemic (outcome) uncertainty in individual-level cause-effect estimates.

2.3 Ontological Uncertainty

The Greek origin of "ontology", i.e., ὀν, means "being". Namely, ontology () deals with questions concerning existence. Ontological uncertainty pertains to whether a causal relationship truly exists. For instance, ice cream sales and shark attacks both increase during the summer months. The hot weather acts as a confounding variable, leading to

Aspect	Aleatoric	Epistemic	Ontological
Definition	The inherent randomness in the causal data.	The lack of knowledge about causal relationships.	The unsureness related to the fundamental existence or non-existence of causal relationships.
Subcategories	(i) Random variability; (ii) Measurement noise.	(i) Model (structure) design uncertainty; (ii) Parameter uncertainty.	(i) Existential uncertainty; (ii) Probabilistic uncertainty; (iii) Contextual uncertainty.
Source	Inherent in the randomness of causal data.	Due to incomplete observation samples, model structure design, and limitations of causal models.	Due to the ambiguity about the existence or definition of concepts, entities, and their relationships.
Reduction method	It is generally considered irreducible even when increasing data or improving model capacity.	Can be mitigated by obtaining more causal data samples or improving model design and capability.	Can be reduced by conceptual clarification, adding contextual information, conducting control experiments, etc.
Illustration with a single example	All these three types of causal uncertainty can be illustrated with the example of studying the effect of a newly developed antihypertensive medicine in lowering blood pressure. Aleatoric uncertainty refers to the situation in which, even under controlled conditions, each patient's body reacts differently to this medicine. Typical epistemic uncertainty, in this case, is the unobserved factors such as participants' dietary habits and fitness routines. One ontological uncertainty example is questioning whether the observed drop in blood pressure is causally related to the medicine intake or just correlated. More examples in the NLP domain are provided in App. A.		

Table 1: Comparison of different kinds of uncertainty in causality: aleatoric, epistemic, and ontological.

higher rates of both activities, but there is no direct causal link between ice cream sales and shark attacks. Ontological causal uncertainty can be further decomposed into three kinds:

- *Existential uncertainty*: As we all know, "correlation does not imply causation." For example, ice cream sales are correlated with shark attacks. However, evidently, there is no causal link between them. People eat more ice cream and swim more frequently in waters where sharks inhabit during hot summers. Namely, the confounder is the hotter weather. However, existing models, even for LLMs like GPT-4, still struggle with distinguishing correlation from causation (Yu et al., 2020; Jin et al., 2024), which highlights the difficulty of identifying existential uncertainty.

- *Probabilistic uncertainty*: The core idea is the link from the cause to the effect is not absolute but probabilistic. Instead of the cause leading to the occurrence of the effect without exception, probabilistic principles state that causes increase the probabilities of their effects' occurrence, but they do not guarantee the occurrence (Good, 1961; Eells, 1991). This principle is also implemented in the probabilistic causal strength metrics (§ 3.2).

- *Contextual uncertainty*: Existing causality with the formulation $C \rightarrow E$ often omit the contextual factors that influence the causal relationship (Carroll et al., 2010; Yarlett and Ramscar, 2019; Wang et al., 2023b; Zhang et al., 2023b). For instance, the causal link between exercise and good health depends on context – such as the exercise being moderate and the absence of underlying health conditions like heart disease.

2.4 Why Trichotomy

While finer subdivisions are possible, these three categories encapsulate the primary facets of causal uncertainty. For instance, *aleatoric* uncertainty focuses on inherent randomness (e.g., two patients responding differently to the same drug), *epistemic* uncertainty stems from knowledge gaps (e.g., not knowing a patient's underlying conditions), and *ontological* uncertainty questions whether a causal link even exists (e.g., distinguishing correlation between ice-cream sales and shark attacks from a true causal relationship).

However, please note that these three categories are intended as a *conceptual lens*, not an exhaustive partition. A real instance of causal uncertainty may bear *multiple* labels (multi-label schemes in (White et al., 2016)). Consider the example: 'hot weather' acts as a hidden confounder, casting *ontological* doubt on whether the observed association *ice-cream sales* $\rightarrow$ *drownings* is causal; the same omission constitutes *epistemic* uncertainty in a model that fails to encode temperature; and individual behavioural variability introduces an *aleatoric* component. However, such overlap does not weaken the trichotomy. Rather, it shows that each category can be treated as a *label* that may co-occur with others. Annotating an instance with multiple labels is therefore analogous to multi-label classification in

NLP: the scheme is not a rigid partition but a set of conceptual anchors that practitioners can combine as the analysis demands. More discussion about the complex interactions among these types of causal uncertainty is given in App. D.

3 Quantification of Uncertainty

Direct estimation of uncertainty and the estimation of causal strength are complementary approaches to understanding causality. A strong causal link often corresponds to lower overall uncertainty, but high uncertainty can also persist if the underlying model remains incomplete. Below, we detail methods for *directly* measuring causal uncertainty, typically focusing on aleatoric and epistemic aspects (§ 3.1), and then turn to causal strength estimation as its complementary counterpart in § 3.2. The interplay between causal uncertainty and causal strength is further elaborated in § 3.3.

3.1 Direct Quantification: Measuring Uncertainty in Causality

Entropy-Based Methods. Quantifying causal uncertainty can leverage similar approaches to predictive uncertainty (Lakshminarayanan et al., 2017; Bakman et al., 2024; Ling et al., 2024), using methods such as entropy-based estimations, which measure the spread or unpredictability of the potential outcomes of causal relationships. We can quantify causal uncertainty using predictive entropy by modeling causal reasoning between variables C and E as a classification problem (binary or ternary). The label set is either $\mathbb{L} = \{0, 1\}$ or $\mathbb{L} = \{-1, 0, 1\}$, where (i) $l = 0$: there is no causal relationship between C and E; (ii) $l = 1$: C facilitates the occurrence of E; (iii) $l = -1$: C prevents the occurrence of E. In this setting, causal uncertainty ($\Phi(C, E)$) can be quantified using predictive entropy:

$$\Phi(C, E) = -\sum_{l \in \mathbb{L}} p(l|(C, E)) \log p(l|(C, E)) \tag{1}$$

where $\mathbb{L}$ is the set of labels. This approach draws on foundational methods used in uncertainty quantification, which measures uncertainty by calculating the conditional entropy of predicted outputs (Shannon, 1948; Cover and Thomas, 2006).

Verbalized Expressions and Modal Words. Beyond numerical measures, verbalized expressions (e.g., "certain", "most likely", "probably", "likely", "even chance", "possibly", "perhaps", "most unlikely", and "impossible") can reflect varying degrees of causality uncertainty (Brogaard and Gatzia, 2017). Zhou et al. (2023) systematically investigate how verbal and numerical markers influence LLMs' performance over uncertainty. Furthermore, Xiong et al. (2024) and Tian et al. (2023) show that verbalized confidence can often be better calibrated than probabilistic (numerical) values for models like GPT-4 (OpenAI, 2023) and Claude (Anthropic, 2024). However, modal words can express more than epistemic uncertainty, making it vital to discriminate epistemic uses from deontic and other senses to ensure accurate uncertainty measurement.

3.2 Complementary Approach: Causal Strength Estimation

Causal strength measures the intensity of the cause leading to the occurrence of the effect. In causal reasoning, the effects of causes are divided into two parts: facilitative and preventative causal strength. The central difference behind these two concepts is how causes change the probability of their effects. Facilitative causal strength studies how a cause positively contributes to the occurrence of the effect, whereas preventative causal strength focuses on how causes negatively impact effects' occurrence. Despite the contrasting nature of these influences, both operate within the same probabilistic framework, offering a complete picture of causation. Further discussion is presented in App. E. The

Metrics	Formulation
Conditional Probability-Based Metrics	
Pearl (2009)	$P(E\|C)$
Difference-Based Metrics	
Suppes (1973)	$P(E\|C) - P(E)$
Cheng (1997)	$[P(E\|C) - P(E\|\neg C)]/[1 - P(E\|\neg C)]$
Eells (1991)	$P(E\|C) - P(E\|\neg C)$
Ratio-based Metrics	
Lewis (1986)	$P(E\|C)/P(E\|\neg C)$
Good (1961)	$\log[1 - P(E\|\neg C)/(1 - P(E\|C)]$

Table 2: Probabilistic causal strength metrics.

estimation of causal strength can be approached in two primary ways (Cui et al., 2024a):

Probability Raising Theory. This approach is based on the idea that cause C increases the probability of the effect E occurring. Metrics derived from this theory involve probability terms: (i) $P(E)$: the probability of E's occurrence without any knowledge of C; (ii) $P(E|C)$: the probability of E's occurrence given the presence of C; (iii) $P(E|\neg C)$: the probability of E's occurrence

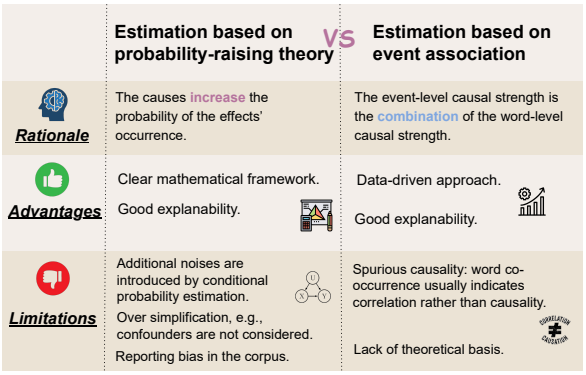


Figure 3: Comparison of metrics based on probability raising theory and event association, highlighting key principles, strengths, and limitations.

given the absence of C. Existing metrics of this line can be roughly classified into three types: (i) The original conditional probability $P(E|C)$; (ii) Difference between probability terms like $P(E|C)$ and $P(E|\neg C)$ or $P(E)$; (iii) Variants based on the ratio of the probability of E's occurrence with the presence of C and without the presence of C. Various metrics are summarized in Table 2.

Events (Words) Association. This approach estimates causal strength based on event co-occurrence frequencies, e.g., "rain" $\rightarrow$ "flood", "heat" $\rightarrow$ "melt", "infect" $\rightarrow$ "sick", etc. Specifically, CEQ (Luo et al., 2016; Du et al., 2022) uses the word co-occurrence frequency as the word-level causal strength, while CESAR (Cui et al., 2024c) uses the association score between events' causal embeddings. Figure 3 highlights the principles, advantages, and limitations of these two methods. However, we should note that association metrics that rely on raw co-occurrence counts inherit the topical and cultural skew of the corpus in which they are measured. Over-reported events (e.g., extreme weather in newswire) artificially inflate apparent causal strength, whereas under-reported or tacit commonsense links (e.g., *drinking water* $\rightarrow$ *quenching thirst*) may appear spuriously weak or even absent. Consequently, reporting bias constitutes an additional source of *epistemic* uncertainty, distinct from sampling variance, because it reflects systematic gaps in what humans choose to write down. Possible mitigations include (i) re-weighting counts with external priors or domain statistics, (ii) triangulating multiple heterogeneous corpora, and (iii) applying causal-feature de-biasing methods that treat medium or genre as a confounder.

3.3 Causal Uncertainty and Causal Strength: Two Sides of the Same Coin

Causal uncertainty and causal strength are inherently complementary but not strictly inversely related. Generally, a high causal strength implies a relatively low causal uncertainty. However, the reverse is not necessarily true. Specifically, high causal uncertainty often indicates ambiguity or incomplete evidence about the causal link, regardless of the intensity and direction (either facilitative or preventative) of causal strength. To understand this, consider that as more contextual information or evidence becomes available, causal uncertainty generally decreases, which can either reinforce or weaken the causal strength. A high facilitative strength typically indicates a solid causal link with minimal uncertainty. For example, a well-established causal relationship like "overeating junk food leads to weight gain" has high causal strength and low uncertainty due to extensive evidence. Nevertheless, high causal uncertainty suggests the evidence supporting causality is unclear without providing any definitive statement about the intensity of the causal link. More comparison between causal strength and causal uncertainty regarding definition, advantages, limitations, and applications is illustrated in Table 3 and App. E.

Aspect	Causal Uncertainty	Causal Strength
Definition	The uncertainty inside the causal relationship (C, E).	The intensity of how likely the cause C leads to the occurrence of the effect E.
Taxonomy	Aleotoric (data randomness), epistemic (knowledge gaps), and ontological (existential questions about causal link).	Facilitative (promoting the effect) and preventative (inhibiting the effect).
Advantages	(i) Beneficial for robust decision; (ii) Helpful for model calibration.	Offering directional insights to indicate either facilitative or preventative.
Limitations	Does not provide information about the direction of causal relationships, i.e., preventative or facilitative.	Sensitive to confounding variables.
Application	(i) Model reliability: enhances trustworthiness in model's causal predictions; (ii) Risk assessment: identifying potential risk by gauging the uncertainty level.	(i) Interventional planning: helping design interventions to strengthen or weaken the desired outcomes; (ii) Causal attributions: identifying the most likely root cause for the investigated events.

Table 3: Comparison of causal uncertainty and causal strength from various aspects.

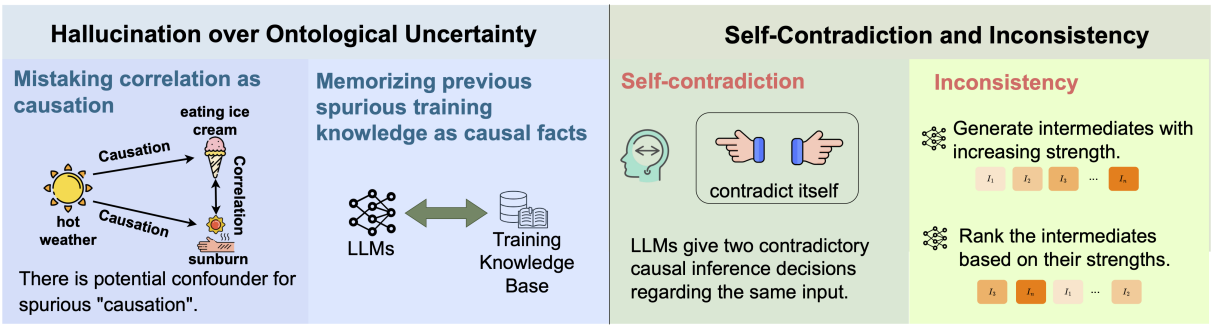


Figure 4: LLMs over causal uncertainty, illustrating challenges like hallucinations and inconsistencies.

4 Causal Uncertainty with LLMs

4.1 LLMs' Causal Reasoning Abilities

LLMs have demonstrated strong capabilities in causal reasoning tasks, which can be categorized into three major areas:

Causal Discovery and Inference. Causal discovery identifies relationships between variables, while causal inference determines their effects. Recent causality-focused methods improve LLMs' performance in these areas, which can be broadly classified into three types: (i) *Prompting strategies*: Prompt engineering enhances LLMs' causal reasoning using Chain-of-Thought (Wei et al., 2022), Few-Shot prompting (Brown et al., 2020), and extensions like multi-turn reasoning (Liu et al., 2024b,a; Bagheri et al., 2024), in-context contrastive learning (Chao et al., 2024), heuristic semantic dependency (Li et al., 2024a), synthetic control (Wang et al., 2024a), and expert reasoning prompts (Lin et al., 2024); (ii) *Causal graphs as auxiliary tools*: LLMs can benefit from causal graphs in reasoning tasks (Jiralerspong et al., 2024; Zheng et al., 2023; Zhang et al., 2024). Vashishtha et al. (2023, 2024) show that topological ordering of graph variables suffices for causal inference, while axiomatic training further enhances reasoning; and (iii) *Agent-based frameworks*: Multi-agent systems help improve LLM-based causal inference (Tang et al., 2023; He et al., 2023; Abdulaal et al., 2024; Le et al., 2024). Tang et al. (2023) propose a multi-agent system where a reasoner LLM generates solutions, and evaluators challenge them with counterfactuals. Abdulaal et al. (2024) introduce an agent unifying metadata and data-based modeling for reasoning, while other frameworks focus on causal explanation (He et al., 2023) and discovery (Le et al., 2024).

Counterfactual Causal Reasoning. Counterfactual causal reasoning, the ability to explore "what if" scenarios, is crucial for robust causal reasoning. Empirically, Li et al. (2024c) improve smaller language models' performance in natural language inference through counterfactual generation. Ashwani et al. (2024) propose a novel architecture to enhance LLMs' causal reasoning and explainability. Mu and Li (2024) introduce a VAE-based method leveraging event commonsense in narratives. Gendron et al. (2024) develop an end-to-end framework that extracts causal graphs from text and performs counterfactual inference.

Commonsense Causal Reasoning. Commonsense causal reasoning enables AI systems to infer cause-and-effect relationships in everyday scenarios using intuitive world knowledge (Cui et al., 2024a). LLMs can predict action outcomes, infer causes, and identify event causality (Hobbhahn et al., 2022; Ko et al., 2023; Sun et al., 2024; Zhang et al., 2023c; Nie et al., 2023). Sun et al. (2024) find that prompt engineering techniques, such as Chain-of-Thought and Tree-of-Thought, enhance LLMs' causal reasoning by structuring reasoning steps and capturing hierarchical dependencies. Zheng et al. (2023) mitigate catastrophic forgetting by using causal inference to retain commonsense knowledge during fine-tuning. Wang et al. (2024c) propose a document-level approach for extracting commonsense causal relations with LLMs.

4.2 LLMs over Causal Uncertainty

While advancements such as improved prompting techniques and multi-agent frameworks have enhanced LLMs' causal reasoning performance, LLMs often generate hallucinations (i.e., plausible but incorrect causal links) and self-contradictions when reasoning under causal uncertainty.

LLMs' Hallucinations over Causal Uncertainty. LLMs often struggle with uncertainty in causality, leading to hallucinations - plausible but incorrect causal links - in their reasoning. Specifically, Zeče-

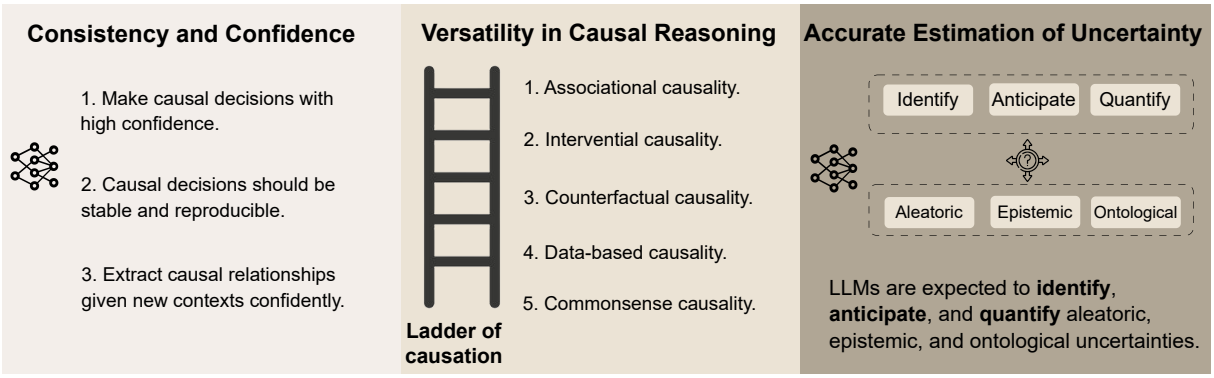


Figure 5: Desired characteristics of an optimal causal LLM: (i) maintaining consistency and confidence; (ii) versatility in different levels of causal reasoning tasks; and (iii) accurate estimation (quantification) of uncertainty.

vić et al. (2023) argue that LLMs are not inherently causal and tend to rely on reciting knowledge learned from their training data rather than performing genuine causal inference. Similarly, Jin et al. (2024) conclude LLMs do not arrive at their answers via genuine reasoning, but rather through memorizing corresponding question and answer pairs. As summarized in Figure 4, these shortcomings highlight the difficulty LLMs have in distinguishing correlation from causation, a crucial aspect of ontological uncertainty (see § 2.3).

Self-Contradictions and Inconsistencies. LLMs frequently struggle to maintain consistency under uncertainty, leading to self-contradictions and inconsistencies (Mündler et al., 2024; Liu et al., 2024d; Li et al., 2024b). Self-contradictions in causal analysis occur when LLMs provide conflicting responses to logically equivalent or minimally altered prompts, usually due to their struggle to resolve ambiguous causal links. Inconsistencies occur when models produce conflicting responses across tasks, for instance, generating intermediates for cause-effect pairs with increasing strengths and ranking them based on their strengths produces a different output, as shown in Figure 4. These issues commonly stem from the partial knowledge of correct causal mechanisms (epistemic) and spurious correlations (ontological). Empirically, Cui et al. (2024b) show that small LLMs ($\leq$7B) barely outperform random baselines in maintaining causal consistency when generating and ranking intermediates in cause-effect pairs. To improve causal consistency, Tang et al. (2023) propose CaCo-CoT, which combines reasoning and evaluating agents to mitigate inconsistencies seen in CoT (Wei et al., 2022) and Self-Consistent CoT (Wang et al., 2023a). Additionally, emerging techniques like causal abstraction (Tan, 2023) and incorporating external knowledge bases (Liu et al., 2023a) show promise in enhancing LLMs' reasoning abilities without producing contradictory statements. More details about causal uncertainty's impact on LLMs are presented in App. F.

4.3 Towards an Optimal Causal LLM

Zhang et al. (2023a) identify three types of causal questions to assess the causal ability of LLMs: (i) identifying causality using domain knowledge; (ii) delivering new causal knowledge from data; and (iii) quantitatively estimating the consequences of actions. However, these categories do not address the inherent uncertainty in causal reasoning. Our survey isolates this topic, clarifying future research on LLMs and causal uncertainty. From the uncertainty perspective, and the shortcomings presented in § 4.2, an optimal causal LLMs should demonstrate the abilities depicted in Figure 5, which describes the following key characteristics:

Consistency and Confidence. LLMs must reliably make confident causal decisions, distinguishing between strong and weak causal links without ambiguity. Decisions should be stable, reproducible, and consistent, minimizing errors from conflicting information (Cui et al., 2024b). Additionally, LLMs should be able to extract causal relationships given new contexts confidently, without relying on the training data they are fed to make conclusions (Zečević et al., 2023; Jin et al., 2023).

Versatility in Causal Reasoning. An optimal causal LLM should be capable of handling various levels of causal reasoning tasks (Yang et al., 2024; Kiciman et al., 2024), including (i) *Associational Causality*: Identifying patterns and correlations within data; (ii) *Interventional Causality*:

Predicting outcomes of potential interventions (Liu et al., 2023c); (iii) *Counterfactual Causality*: Assessing hypothetical scenarios to determine alternative outcomes; (iv) *Data-based Causality*: assessing data-based statistical causal reasoning (Liu et al., 2024b); and (v) *Commonsense Causality*: conducting causal reasoning based on commonsense knowledge, i.e., general public's intuition how the occurrence of one event contributes to another happening (Cui et al., 2024a). This versatility ensures the LLM can adapt to different contexts and provide accurate causal reasoning ability.

Accurate Estimation of Causal Uncertain Influences. An optimal causal LLM should identify, anticipate, and quantify the impact of uncertain factors, encompassing aleatoric (data randomness), epistemic (knowledge gaps), and ontological (existence of causal links) uncertainty (Jin et al., 2023; Stolfo et al., 2023). LLMs should also effectively model these uncertainties to provide probabilistic estimations that reflect potential variability in causal predictions (Kiciman et al., 2024). This includes generating confidence intervals or probability distributions over possible outcomes, thereby enabling more informed decision-making in the presence of uncertainty.

Together, these traits define an optimal causal agent that is robust, reliable, and capable of navigating complex causal relationships while managing the uncertainty in causality. More details on future directions are in App. G.

5 Concluding Remarks

This survey provides a comprehensive review of the uncertainty in causality, categorizing it into aleatoric, epistemic, and ontological types. We synthesize existing methods for quantifying uncertainty and explore their complementary relationship with causal strength. We further highlight LLMs' limitations over causal uncertainty and underscore future research directions. This survey aims to provide valuable guidance for researchers and practitioners regarding causal uncertainty.

Limitations

We acknowledge the following limitations in our work. Firstly, the taxonomy of causal uncertainty is relatively subjective. Our trichotomy of aleatoric, epistemic, and ontological causal uncertainties strives to capture the most distinct facets of causal uncertainty: data randomness, model limitations, and existential doubt. However, it still inevitably retains a degree of subjectivity. In particular, alternative classifications may slice these categories differently or introduce finer-grained subtypes. For instance, splitting epistemic uncertainty into structural and confounder-focused dimensions. Although we believe our trichotomy strikes a balance between conceptual clarity and practical utility, the field may benefit from further theoretical or domain-specific refinements. Second, though we try our best to cover influential and recent studies across both machine learning and NLP venues, the field's rapid growth means that some relevant work may not be fully captured. Due to the page limit, the description of the techniques is generally sketchy. In particular, area-specific research such as event causality extraction and causal reasoning may present further nuances that we have not fully covered. We suggest readers consult area-specific publications for deeper insights.

Ethical Considerations

Our paper summarizes existing literature and comprehensively reviews uncertainty in causal reasoning. As a survey paper, we do not foresee significant ethical considerations or risks in our paper. We use ai assistants to check the grammar issues in our writing. While this survey does not introduce direct ethical concerns, the propagation of biased or inaccurate causal models, particularly in high-stake fields like healthcare or criminal justice, could have severe implications if causal uncertainty is mismanaged. To mitigate such risks, we advocate for the transparent reporting of model uncertainty. Moreover, it is essential to acknowledge that the studies we cite may contain unrecognized biases or unfairness. These biases can inadvertently be perpetuated through our review. We encourage readers to critically assess these potential biases when understanding the findings and conclusions presented in this survey.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul W. Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion*, 76:243–297.

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-

Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. 2024. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*.

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. In *Anthropic website*.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? In *Proceedings of the AAAI Symposium Series*, volume 4, pages 2–9.

Abdolmahdi Bagheri, Matin Alinejad, Kevin Bello, and Alireza Akhondi Asl. 2024. C2P: featuring large language models with causal reasoning. *CoRR*, abs/2407.18069.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7752–7767, Bangkok, Thailand. Association for Computational Linguistics.

Txus Blasco, J. Salvador Sánchez, and Vicente García. 2024. A survey on uncertainty quantification in deep learning for financial time series prediction. *Neurocomputing*, 576:127339.

Gianluca Bontempi and Maxime Flauder. 2015. From dependency to causality: A machine learning approach. *Journal of Machine Learning Research*, 16(74):2437–2457.

Brit Brogaard and Dimitria Electra Gatzia. 2017. Introduction: epistemic modals. *Topoi*, 36(1):127–130.

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

Margarida Campos, António Farinhas, Chrysoula Zerva, Mário A. T. Figueiredo, and André F. T. Martins. 2024. Conformal prediction for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 12:1497–1516.

Christopher Carroll, Patricia Cheng, and Hongjing Lu. 2010. Uncertainty in causal inference: The case of retrospective revaluation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Liang Chao, Wei Xiang, and Bang Wang. 2024. In-context contrastive learning for event causality identification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 868–881, Miami, Florida, USA. Association for Computational Linguistics.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200, Bangkok, Thailand. Association for Computational Linguistics.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8772–8779, Online. Association for Computational Linguistics.

Ziwei Chen, Linmei Hu, Weixin Li, Yingxia Shao, and Liqiang Nie. 2023. Causal intervention and counterfactual reasoning for multi-modal fake news detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 627–638, Toronto, Canada. Association for Computational Linguistics.

Patricia W Cheng. 1997. From covariation to causation: A causal power theory. *Psychological review*, 104(2):367.

Changwoo Chun, SongEun Lee, Jaehyung Seo, and Heuiseok Lim. 2023. CReTIHC: Designing causal reasoning tasks about temporal interventions and hallucinated confoundings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10334–10343, Singapore. Association for Computational Linguistics.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory (2. ed.)*. Wiley.

Shaobo Cui, Zhijing Jin, Bernhard Schölkopf, and Boi Faltings. 2024a. The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16722–16763, Miami, Florida, USA. Association for Computational Linguistics.

Shaobo Cui, Junyou Li, Luca Mouchel, Yiyang Feng, and Boi Faltings. 2024b. Nuance matters: Probing epistemic consistency in causal reasoning. *CoRR*, abs/2409.00103.

Shaobo Cui, Lazar Milikic, Yiyang Feng, Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Boi Faltings. 2024c. Exploring defeasibility in causal reasoning. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6433–6452, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shiyao Cui, Jiawei Sheng, Xin Cong, Quangang Li, Tingwen Liu, and Jinqiao Shi. 2022. Event causality extraction with event argument correlations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2300–2312, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Maria R. Dahm and Carmel Crock. 2022. Understanding and Communicating Uncertainty in Achieving Diagnostic Excellence. *JAMA*, 327(12):1127–1128.

Tirthankar Dasgupta, Rupsa Saha, Lipika Dey, and Abir Naskar. 2018. Automatic extraction of causal relations from text using linguistically informed deep neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 306–316, Melbourne, Australia. Association for Computational Linguistics.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Ellery Eells. 1991. *Probabilistic Causality*, volume 1. Cambridge University Press.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Yair Ori Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amit Sharma, and Roi Reichart. 2024. Faithful explanations of black-box NLP models using LLM-generated counterfactuals. In *The Twelfth International Conference on Learning Representations*.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, et al. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(Suppl 1):1513–1589.

Gael Gendron, Joze M Rozanec, Michael Witbrock, and Gillian Dobbie. 2024. Counterfactual causal inference in natural language with large language models. In *Causality and Large Models@ NeurIPS 2024*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Irving J Good. 1961. A causal calculus (i). *The British journal for the philosophy of science*, 11(44):305–318.

Mingyue Han and Yinglin Wang. 2021. Doing good or doing right? exploring the weakness of commonsense causal reasoning models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 151–157, Online. Association for Computational Linguistics.

Jianfeng He, Linlin Yu, Shuo Lei, Chang-Tien Lu, and Feng Chen. 2024. Uncertainty estimation on sequential labeling via uncertainty transmission. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2823–2835, Mexico City, Mexico. Association for Computational Linguistics.

Zhitao He, Pengfei Cao, Yubo Chen, Kang Liu, Ruopeng Li, Mengshu Sun, and Jun Zhao. 2023. LEGO: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9142–9163, Singapore. Association for Computational Linguistics.

Marius Hobbhahn, Tom Lieberum, and David Seiler. 2022. Investigating causal understanding in LLMs. In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Pedram Hosseini, David A. Broniatowski, and Mona Diab. 2022. Knowledge-augmented language models for cause-effect relation classification. In *Proceedings of the First Workshop on Commonsense Representation and Reasoning (CSRR 2022)*, pages 43–48, Dublin, Ireland. Association for Computational Linguistics.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in natural language processing: Sources, quantification, and applications. *CoRR*, abs/2306.04459.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. 2020. Identifying causal-effect inference failure with uncertainty-aware models. In *Advances in Neural Information Processing Systems*, volume 33, pages 11637–11649. Curran Associates, Inc.

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. 2023. Cladder: A benchmark to assess causal reasoning capabilities of language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. 2024. Efficient causal graph discovery using large language models. In *ICLR 2024 Workshop: How Far Are We From AGI*.

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. 2024. COLD: causal reasoning in closed daily activities. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Amita Kamath, Robin Jia, and Percy Liang. 2020. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5684–5696, Online. Association for Computational Linguistics.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5574–5584.

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*. Featured Certification.

Yuheun Kim, Lu Guo, Bei Yu, and Yingya Li. 2023. Can ChatGPT understand causal language in science claims? In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 379–389, Toronto, Canada. Association for Computational Linguistics.

Brian Knaeble, Braxton Osting, and Placede Tshiaba. 2023. An asymptotic threshold of sufficient randomness for causal inference. *Stat*, 12(1):e609.

Dohwan Ko, Ji Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo Kim. 2023. Large language models are temporal and causal reasoners for video question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4300–4316, Singapore. Association for Computational Linguistics.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6402–6413.

Hao Duong Le, Xin Xia, and Zhang Chen. 2024. Multi-agent causal discovery using large language models. *CoRR*, abs/2407.15073.

David Lewis. 1986. Causal explanation. In David K. Lewis, editor, *Philosophical Papers Vol. II*, pages 214–240. Oxford University Press.

Haoran Li, Qiang Gao, Hongmei Wu, and Li Huang. 2024a. Advancing event causality identification via heuristic semantic dependency inquiry network. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Miami, Florida, USA. Association for Computational Linguistics.

Jiaxuan Li, Lang Yu, and Allyson Ettinger. 2023. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 804–815, Toronto, Canada. Association for Computational Linguistics.

Jierui Li, Vipul Raheja, and Dhruv Kumar. 2024b. ContraDoc: Understanding self-contradictions in documents with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6509–6523, Mexico City, Mexico. Association for Computational Linguistics.

Yongqi Li, Mayi Xu, Xin Miao, Shen Zhou, and Tieyun Qian. 2024c. Prompting large language models for counterfactual generation: An empirical study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13201–13221, Torino, Italia. ELRA and ICCL.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Tianjun Lin, Vishvak Khattar, Yichen Huang, Junxian Hong, Ruoxi Jia, Chuangang Liu, Alberto Sangiovanni-Vincentelli, and Ming Jin. 2024. Causalprompt: Enhancing llms with weakly supervised causal reasoning for robust performance in non-language tasks. In *ICLR Workshop: Tackling Climate Change with Machine Learning*.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Cheng Liu, Wei Xiang, and Bang Wang. 2024a. Identifying while learning for document event causality identification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3815–3827, Bangkok, Thailand. Association for Computational Linguistics.

Jintao Liu, Zequn Zhang, Zhi Guo, Li Jin, Xiaoyu Li, Kaiwen Wei, and Xian Sun. 2023a. Kept: Knowledge enhanced prompt tuning for event causality identification. *Knowledge-Based Systems*, 259:110064.

Jintao Liu, Zequn Zhang, Kaiwen Wei, Zhi Guo, Xian Sun, Li Jin, and Xiaoyu Li. 2023b. Event causality extraction via implicit cause-effect interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6792–6804, Singapore. Association for Computational Linguistics.

Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. 2024b. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9215–9235, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023c. The magic of IF: Investigating causal reasoning abilities in large language models of code. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9009–9022, Toronto, Canada. Association for Computational Linguistics.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian J. McAuley, Wei Ai, and Furong Huang. 2024c. Large language models and causal inference in collaboration: A comprehensive survey. *CoRR*, abs/2403.09606.

Ziyi Liu, Soumya Sanyal, Isabelle Lee, Yongkang Du, Rahul Gupta, Yang Liu, and Jieyu Zhao. 2024d. Self-contradictory reasoning evaluation and detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3725–3742, Miami, Florida, USA. Association for Computational Linguistics.

Kun Luo, Tong Zhou, Yubo Chen, Jun Zhao, and Kang Liu. 2024. Open event causality extraction by the assistance of LLM in task annotation, dataset, and method. In *Proceedings of the Workshop: Bridging Neurons and Symbols for Natural Language Processing and Knowledge Graphs Reasoning (NeusymBridge) @ LREC-COLING-2024*, pages 33–44, Torino, Italia. ELRA and ICCL.

Zhiyi Luo, Yuchen Sha, Kenny Q. Zhu, Seung-won Hwang, and Zhongyuan Wang. 2016. Commonsense causal reasoning between short texts. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Fifteenth International Conference, KR 2016, Cape Town, South Africa, April 25-29, 2016*, pages 421–431. AAAI Press.

María D. López-Gamero, José F. Molina-Azorín, and Enrique Claver-Cortés. 2011. Environmental uncertainty and environmental management perception: A multiple case study. *Journal of Business Research*, 64(4):427–435.

Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems*, 31.

José Mena, Oriol Pujol, and Jordi Vitrià. 2021. A survey on uncertainty estimation in deep learning classification systems from a bayesian perspective. *ACM Computing Surveys (CSUR)*, 54(9):1–35.

Feiteng Mu and Wenjie Li. 2024. A causal approach for counterfactual reasoning in narratives. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6556–6569, Bangkok, Thailand. Association for Computational Linguistics.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Hariharan Nalatore, Mingzhou Ding, and Govindan Rangarajan. 2007. Mitigating the effects of measurement noise on granger causality. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 75(3):031123.

Van Bach Nguyen, Christin Seifert, and Jörg Schlötterer. 2024. CEval: A benchmark for evaluating counterfactual text generation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 55–69, Tokyo, Japan. Association for Computational Linguistics.

Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori B. Hashimoto, and Tobias Gerstenberg. 2023. Moca: Measuring human-language model alignment on causal and moral judgment tasks. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

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

Judea Pearl and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic books.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. In *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. Association for Computational Linguistics.

Luca Rigotti and Chris Shannon. 2005. Uncertainty and risk in financial markets. *Econometrica*, 73(1):203–243.

Rachel Rudinger, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi. 2020. Thinking like a skeptic: Defeasible inference in natural language. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4661–4675, Online. Association for Computational Linguistics.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Shirong Shen, Heng Zhou, Tongtong Wu, and Guilin Qi. 2022. Event causality identification via derivative prompt joint learning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2288–2299, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alessandro Stolfo, Zhijing Jin, Kumar Shridhar, Bernhard Schoelkopf, and Mrinmaya Sachan. 2023. A causal framework to quantify the robustness of mathematical reasoning with language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 545–561, Toronto, Canada. Association for Computational Linguistics.

Yaru Sun, Ying Yang, and Wenhao Fu. 2024. Exploring synergies between causal models and largelanguage models for enhanced understanding and inference. In *Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition*, CVIPPR '24, New York, NY, USA. Association for Computing Machinery.

Patrick Suppes. 1973. A probabilistic theory of causality. *British Journal for the Philosophy of Science*, 24(4).

Juanhe (TJ) Tan. 2023. Causal abstraction for chain-of-thought reasoning in arithmetic word problems. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 155–168, Singapore. Association for Computational Linguistics.

Ziyi Tang, Ruilin Wang, Weixing Chen, Keze Wang, Yang Liu, Tianshui Chen, and Liang Lin. 2023. Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms. *CoRR*, abs/2308.11914.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15109–15126, Toronto, Canada. Association for Computational Linguistics.

Dennis Ulmer, Jes Frellsen, and Christian Hardmeier. 2022. Exploring predictive uncertainty and calibration in NLP: A study on the impact of method & data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2707–2735, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dennis Thomas Ulmer, Christian Hardmeier, and Jes Frellsen. 2023. Prior and posterior networks: A survey on evidential deep learning methods for uncertainty estimation. *Transactions on Machine Learning Research*.

Jordy Van Landeghem, Matthew Blaschko, Bertrand Anckaert, and Marie-Francine Moens. 2022. Benchmarking scalable predictive uncertainty in text classification. *IEEE Access*, 10:43703–43737.

Aniket Vashishtha, Abhinav Kumar, Abbavaram Gowtham Reddy, Vineeth N. Balasubramanian, and Amit Sharma. 2024. Teaching transformers causal reasoning through axiomatic training. *CoRR*, abs/2407.07612.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. 2023. Causal inference using LLM-guided discovery. In *AAAI 2024 Workshop on "Are Large Language Models Simply Causal Parrots?"*.

Haoyu Wang, Fengze Liu, Jiayao Zhang, Dan Roth, and Kyle Richardson. 2024a. Event causality identification with synthetic control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1725–1737, Miami, Florida, USA. Association for Computational Linguistics.

Siyin Wang, Jie Zhou, Changzhi Sun, Junjie Ye, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022. Causal intervention improves implicit sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6966–6977, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024b. A survey on natural language counterfactual generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4798–4818, Miami, Florida, USA. Association for Computational Linguistics.

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2023b. COLA: Contextualized commonsense causal reasoning from the causal inference perspective. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5253–5271, Toronto, Canada. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024c. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, Texas. Association for Computational Linguistics.

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations*.

Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck. 2024. A critical review of causal reasoning benchmarks for large language models. *CoRR*, abs/2407.08029.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2021. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5).

Daniel Yarlett and Michael Ramscar. 2019. Uncertainty in causal and counterfactual inference. In *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society*, pages 956–961. Routledge.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*.

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023a. Understanding causality with large language models: Feasibility and opportunities. *CoRR*, abs/2304.05524.

Chi Zhang, Karthika Mohan, and Judea Pearl. 2023b. Causal inference under interference and model uncertainty. In *2nd Conference on Causal Learning and Reasoning*.

Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. 2020. Causal intervention for weakly-supervised semantic segmentation. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Jiayao Zhang, Hongming Zhang, Weijie Su, and Dan Roth. 2022. Rock: Causal inference principles for reasoning about commonsense causality. In *International Conference on Machine Learning*, pages 26750–26771. PMLR.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023c. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. 2024. Causal graph discovery with retrieval-augmented generation based large language models. *CoRR*, abs/2402.15301.

Junhao Zheng, Qianli Ma, Shengjie Qiu, Yue Wu, Peitian Ma, Junlong Liu, Huawen Feng, Xichen Shang, and Haibin Chen. 2023. Preserving commonsense knowledge from pre-trained language models via causal inference. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9155–9173, Toronto, Canada. Association for Computational Linguistics.

Kaitlyn Zhou, Jena Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3623–3643, Bangkok, Thailand. Association for Computational Linguistics.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524, Singapore. Association for Computational Linguistics.

Xinlei Zhou, Han Liu, Farhad Pourpanah, Tieyong Zeng, and Xizhao Wang. 2022. A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing*, 489:449–465.

A Examples and Impact of Causal Uncertainty in NLP Tasks

While causal uncertainty is a general concept applicable to numerous domains, it holds particular significance for many NLP tasks. In what follows, we demonstrate how causal uncertainty impacts textual cause-effect analysis across diverse NLP tasks.

A.1 Event Causality Extraction

Causal uncertainty in NLP tasks is often reduced to event causality extraction (Dasgupta et al., 2018; Hosseini et al., 2022; Shen et al., 2022; Cui et al., 2022; Liu et al., 2023b), where models identify cause-effect links in textual content. The following are examples of these three types of causal uncertainty in the event causality extraction task:

- *Aleatoric uncertainty:* For instance, coverage of the same geopolitical event and its sequence in news articles from different sources can diverge in tone or detail, yielding inherently noisy data. A model might inconsistently assign causal labels because certain news agencies exaggerate or underreport pivotal triggers.

- *Epistemic uncertainty:* If a causal event extraction model has no background knowledge about the cultural or historical context in its training corpus, it may fail to distinguish a direct cause from tangential events. The system's incomplete knowledge of these contexts in its training corpus further raises doubts about the extracted cause-effect pair.

- *Ontological uncertainty:* The existence of a cause-effect link itself can be questionable if news reports only highlight correlation. For example, when a model detects correlation patterns (e.g., "Ice cream sales increase during the hot summer months. At the same time, police report a rise in drowning incidents along crowded beaches and pools."), *ontological uncertainty* may arise if the text only presents correlational cues without direct evidence of causal relationships. In this example, we could see the true cause for "drowning incidents" is the "hot weather and more swimming" rather than the correlation factor of "increased ice cream sales".

In summary, understanding these uncertainties is vital for ensuring accurate extraction of cause-effect relationships from textual inputs. By distinguishing spurious causal links from genuine causal links, models can extract more reliable causal event graphs, which is key to downstream applications.

A.2 Commonsense Reasoning

Commonsense reasoning tasks often require a nuanced understanding of everyday cause-effect relationships in text (Ponti et al., 2020; Han and Wang, 2021; Chun et al., 2023; Zhang et al., 2023c; Chen et al., 2023). For example, consider a logistics planning question: "Why might ignoring an upcoming severe weather front lead to shipping delays?" Answering this question involves applying a general causal reasoning rule in general cases ("extreme weather" $\rightarrow$ "flight or route disruptions") alongside domain-specific knowledge about supply chains (e.g., rerouting costs, limited carrier options). *Epistemic causal uncertainty* emerges if the training corpus lacks explicit coverage of the interplay between inclement weather and shipping routes, leading the model to produce incomplete or unconvincing answers. Moreover, *aleatoric uncertainty* can lead to hallucinations over noisy or ambiguous textual mentions about the severity of the bad weather (e.g., "strong winds" vs. "hurricane conditions"), making it difficult to gauge the precise level of logistic disruption.

To sum up, commonsense reasoning in text often relies on incomplete clues and ambiguous descriptions, making it easy for models to confuse inherent randomness or partial knowledge with genuine causal rules. By explicitly and accurately handling these uncertainties, models can deliver more robust and reliable reasoning conclusions, especially in high-stake domains like policy-making and healthcare.

A.3 Counterfactual Reasoning and Counterfactual Text Generation

Counterfactual reasoning (Qin et al., 2019; Liu et al., 2023c; Li et al., 2023; Mu and Li, 2024) and counterfactual text generation (Yang et al., 2020; Treviso et al., 2023; Nguyen et al., 2024; Wang et al., 2024b) involve assessing "what if" scenarios, such as rewriting a story by removing a causal trigger or altering a key causal condition. Suppose a generative model aims to produce an alternate storyline where a triggering causal event does not happen (e.g., removing a character's decision to

light a fire). In that case, it must manage *epistemic uncertainty* regarding unknown downstream ramifications in the narrative. Additionally, *aleatoric uncertainty* may arise if multiple endings are equally plausible based on existing textual clues. Without full knowledge of *ontological uncertainty*, the system might hallucinate cause-effect chains that were never grounded in the original story.

To sum up, explicitly classifying, identifying, and quantifying causal uncertainty in NLP tasks is not merely a theoretical exercise. It tangibly bolsters downstream NLP tasks and helps to build robust intelligent causal agents. This makes understanding causal uncertainty a necessary component for complicated language understanding and reasoning tasks.

B Related Surveys

In Table 4, we present various surveys that are associated with the concept of causal uncertainty. These surveys are organized into four distinct categories:

- *Surveys of uncertainty in machine learning*: These surveys (Gawlikowski et al., 2023) cover the classification of uncertainty and the quantification methods.
- *Surveys of causal inference*: Yao et al. (2021) and (Liu et al., 2024b) review causal inference methods and the integration with machine learning and LLMs.
- *Surveys of Uncertainty in language modeling*: there are multiple surveys (Hu et al., 2023; Geng et al., 2024) that particularly focus on uncertainty in language modeling.
- *Surveys of commonsense causality*: Although Cui et al. (2024a) use uncertainty level as a criterion for classifying existing commonsense causality benchmarks and reasoning methods, they do not explicitly illustrate different kinds of uncertainty sources. Our survey not only introduces a novel trichotomy but also illustrates the complementary relationship between causal strength and causal uncertainty. Additionally, we elaborate on existing literature regarding LLMs' performance in addressing causal uncertainty.

Our survey distinguishes itself from existing surveys by bridging the gap between uncertainty and causality. Furthermore, we uniquely distinguish between aleatoric, epistemic, and ontological uncertainty in causality. This trichotomy is tailored to uncertainty in causality, diverging from traditional uncertainty in machine learning (Gawlikowski et al.,

Citation	Summary
Uncertainty in Machine Learning	
Zhou et al. (2022)	They provide a comprehensive review of epistemic uncertainty in supervised learning and decompose the epistemic uncertainty into bias and variance types.
Gawlikowski et al. (2023)	They review types of uncertainty: model and data uncertainty. They focus on approaches for estimating uncertainty, such as Bayesian inference, ensemble methods, and test-time augmentation.
Mena et al. (2021)	They review the definition and quantification of uncertainty when applied to classification systems.
Ulmer et al. (2023)	This survey reviews evidential deep learning (Sensoy et al., 2018) for uncertainty estimation, which is different from predictive uncertainty that involves distributions over parameters.
Blasco et al. (2024)	They review existing works on uncertainty quantification methods to predict the behavior of financial assets. These works span the years from 2001 to 2022.
Causal Inference	
Yao et al. (2021)	They review causal inference methods within the potential outcome framework, focusing on the integration with machine learning. The review covers methods from both statistical and machine-learning perspectives.
Liu et al. (2024c)	They review the intersection of causal inference and LLMs, focusing on how causal models can enhance reasoning, fairness, and explainability. They also discuss how LLMs can discover causal relationships.
Uncertainty in Language Modeling	
Hu et al. (2023)	They first classified the sources of uncertainty in NLP systems into three kinds: input, system, and output. And then, they review the works focusing on uncertainty quantification in NLP systems.
Geng et al. (2024)	They review confidence estimation and calibration techniques for LLMs, highlighting factual errors and instability.
Campos et al. (2024)	This paper reviews the application of conformal prediction in various language modeling tasks.
Commonsense Causality	
Cui et al. (2024a)	They focus on taxonomies, benchmarks, acquisition methods, qualitative reasoning, and quantitative measurements in commonsense causality.

Table 4: Related surveys categorized by their research areas.

2023; Mena et al., 2021; Blasco et al., 2024). Additionally, we underscore the complementary relationship between causal uncertainty quantification and causal strength (more details in § 3 and App. E).

C Preliminary Knowledge

C.1 Uncertainty Sources

There are two major sources of uncertainty involved in machine learning models: data-induced (aleatoric) and modeling-induced (epistemic). Data-induced uncertainty (aleatoric uncertainty) includes labeling noise, measurement errors, or inherent variability in the data and is generally irreducible. Modeling-induced uncertainty (epistemic uncertainty) includes ambiguities related to the machine learning models, tools, or methods, and can often be reduced with more data or improved modeling techniques.

However, this traditional dichotomy of aleatoric and epistemic uncertainty is insufficient when applied to causality, as it does not account for the complexities introduced by the *existence* of the causal relationship and its *directionality*. This requires a more nuanced treatment, as causal relationships inherently involve questions of existence and directionality that are not present in conventional predictive modeling.

C.2 Uncertainty Expression

In machine learning, there are three major ways of expressing uncertainty:

- *Numerical Uncertainty Quantification*: Modeling the level of uncertainty with a numerical score. Calibration scores, conditional probabilities, predictive entropy, and confidence regions (Xiao et al., 2022; Kuhn et al., 2023; Fadeeva et al., 2024; He et al., 2024; Bakman et al., 2024; Duan et al., 2024; Chen and Mueller, 2024) can be categorized into this type.

- *Set-Valued Prediction*: Instead of providing only one answer, provide a set of candidate answers. Methods like conformal predictions (Angelopoulos and Bates, 2021; Ulmer et al., 2022) fall into this category.

- *Abstain Answering (Selective Prediction)*: Instead of giving a definite answer or giving a set of candidates, avoiding answering is another way of expressing uncertainty when the model's confidence is not high enough (Kamath et al., 2020; Xin et al., 2021; Zhou et al., 2024).

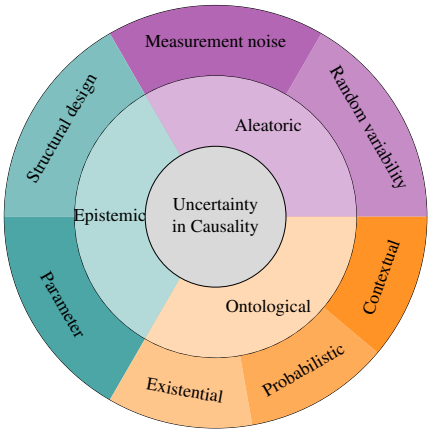


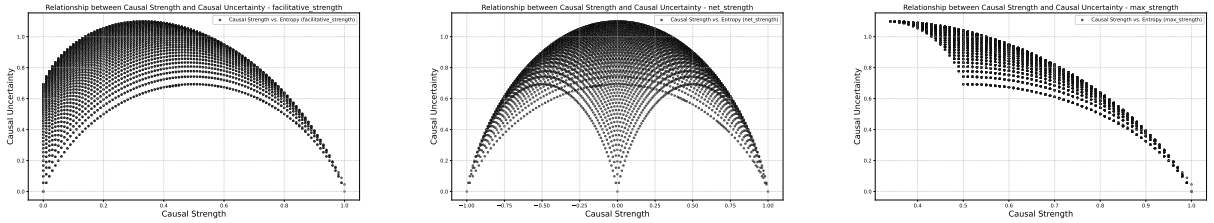
Figure 6: A concentric view of causal uncertainty types—aleatoric, epistemic, and ontological. Although depicted as conceptually distinct, real-world cases frequently blur these boundaries, highlighting their interwoven nature in practice.

These existing methods primarily handle uncertainty from a predictive modeling perspective and are not well adapted to the specific demands of causal inference. Causality involves understanding not just correlations but the underlying mechanisms and effects of interventions, which requires a more refined framework for uncertainty expression.

D Entangled Uncertainties: How Aleatoric, Epistemic, and Ontological Uncertain Factors Intertwine?

Although we distinguish aleatoric, epistemic, and ontological uncertainties as three conceptually separate types(as viewed in Figure 6), real-world causal problems often exhibit intricate interactions among them. Below are two illustrative scenarios:

- Ontological uncertainty leading to model limitations (epistemic uncertainty): For instance, a hidden or unaccounted causal factor may introduce ontological uncertainty about the underlying structure of a medical study. If a genetic trait is unobserved, researchers face existential doubts about whether a purported cause actually influences a health outcome. This gap in causal structure then manifests as epistemic uncertainty: the model remains limited in predicting outcomes because it is missing a critical piece of information.



(a) Facilitative causal strength. (b) Net causal strength. (c) Maximum causal strength.

Figure 7: Relationship between causal strength and causal uncertainty using different formulations of causal strength.

- From random variability (aleatoric) to undecided existence (ontological): For instance, in agricultural experiments, the crop yields usually fluctuate due to various factors, including weather, soil health, and seed genetic differences. Such inherent randomness (aleatoric) may obscure the causal inference conclusion of whether a new fertilizer truly increases crop yield or not. In this case, the systems encounter ontological uncertainty as they are not sure about whether a causal relationship holds or not.

These examples underscore that real-world causal analysis rarely isolates just one type of uncertainty. Instead, apparent shortfalls in data, models, or even the existence of a causal link often blur the lines between aleatoric, epistemic, and ontological considerations.

E More Discussion about Causal Strength and Causal Uncertainty

There can be various definitions of causal strength and causal uncertainty. Suppose there is a ternary classification problem with labels of (i) $l = 1$: event C facilitates the occurrence of event E; (ii) $l = -1$: event C prevents the occurrence of event E; (iii) $l = 0$: event C has no influence on the occurrence of event E. In this setting, there is a relationship between causal uncertainty and causal strength.

Causal Uncertainty. In this case, the predictive entropy about causal uncertainty is defined as

$$\Phi(C, E) = -\sum_{l \in \mathbb{L}} p(l|(C, E)) \log p(l|(C, E)) \tag{2}$$

where $\mathbb{L} = \{+1, -1, 0\}$ is the set of labels. This entropy measures the degree of uncertainty in predicting the causal relationship between C and E. A higher entropy indicates greater uncertainty, while a lower entropy suggests more confidence in the prediction.

Causal Strength. The causal strength, $\text{CS}(C, E)$ and defined in the following three ways:

- *Facilitative Causal Strength:* Focuses on the facilitative effect of C on E.

$$\text{CS}(C, E) = p(l = +1 \mid (C, E)) \tag{3}$$

This definition captures how likely C is to positively impact E. It is useful in contexts where the main interest is in enhancing positive outcomes, such as determining the effectiveness of a treatment or intervention.

- *Net Causal Strength:* Balances facilitative and preventative effects.

$$\begin{aligned} \text{CS}(C, E) = & p(l = +1 \mid (C, E)) - \\ & p(l = -1 \mid (C, E)) \end{aligned} \tag{4}$$

This definition provides a balanced view of the overall impact by considering both positive and negative influences. It helps assess the net effect, which is valuable in scenarios where both facilitation and prevention matter.

- *Maximum Causal Influence:* Reflects the strongest effect among facilitation, prevention, or neutrality.

$$\begin{aligned} \text{CS}(C, E) = \max\big(& p(l = +1 \mid (C, E)), \\ & p(l = -1 \mid (C, E)), \\ & p(l = 0 \mid (C, E))\big) \end{aligned} \tag{5}$$

This approach highlights the most dominant influence of C on E, whether facilitative, preventative, or neutral. It is particularly useful when identifying the strongest effect, which is crucial, such as in decision-making processes.

The visualization of the relationship between causal uncertainty and causal strength using different formulations of causal strength is presented in Figure 7. The plots show that as causal strength increases, causal uncertainty tends to decrease, supporting the notion that they are inversely related but not strictly inverses of each other. Furthermore, it's important to note that while a high causal strength generally corresponds to lower causal uncertainty, a low causal strength does not necessarily imply high uncertainty. The uncertainty depends on the distribution of probabilities across the labels. For example, if all probabilities are low and evenly distributed, uncertainty is high, but if one probability is low and the others are negligible, uncertainty may still be low due to the dominance of one outcome.

F Impact of Causal Uncertainty on LLMs

Causal uncertainty poses a significant challenge to the reliability of causal decisions given by LLMs. Addressing this issue requires a comprehensive understanding of the different types of uncertainties' impact on LLMs and a systematic evaluation of this influence. In this section, we first highlight how each type undermines the performance of causal reasoning in LLMs. Subsequently, we propose several promising evaluation methodologies designed to quantify and mitigate the impact of these uncertainties on LLM outputs.

Types of Causal Uncertainty in LLM Outputs . Though LLMs often show impressive performance, hidden uncertainties in causality can lead to flawed causal conclusions (Zečević et al., 2023; Tang et al., 2023; Liu et al., 2023a; Jin et al., 2024; Mündler et al., 2024). By highlighting examples of aleatoric randomness, epistemic knowledge gaps, and ontological misinterpretations, we reveal how each uncertainty type can degrade LLM outputs. This perspective underlines the need for more transparent, uncertainty-aware modeling strategies in LLMs to mitigate spurious or even erroneous causal claims in practice.

- Aleatoric uncertainty: Inherent randomness in data samples can lead to LLMs' inconsistent and incorrect causal decisions (Liu et al., 2024c). For instance, even with controlled experiments, LLMs may still make erroneous decisions regarding treatment for patients, as patients' reactions to the same treatment vary. This is due to individual differences.
- Epistemic uncertainty: Missing data or model design flaws will likely cause LLMs to hallucinate over the causal links (Zečević et al., 2023; Jin et al., 2024). For example, an LLM might incorrectly link diet to specific health outcomes if critical variables like genetics are missing. This might be due to the missing data in the LLMs' pre-training corpus.
- Ontological uncertainty: LLMs might generate spurious causal connections (Mündler et al., 2024; Liu et al., 2024d), such as linking ice cream sales to shark attacks when only a confounder(confounded by hot weather) exists but no direct causation between ice cream and shark attacks.

In conclusion, these diverse forms of causal uncertainty can substantially degrade the reliability of LLM outputs. Addressing them is essential for conducting robust, trustworthy causal reasoning for LLMs.

Evaluating the Influence of Causal Uncertainty on LLM Outputs. Systematic evaluation of causal uncertainty's influence on large language models (LLMs) is critical for identifying and then mitigating erroneous causal conclusions. There are several promising approaches to effectively measure this impact: (i) Specific benchmark construction: novel benchmarks can be designed that specifically challenge LLMs with scenarios involving spurious causal links, such as those that are purely correlational or confounded. Additionally, constructing datasets annotated with explicit uncertainty-level indicators can further assess whether LLMs accurately express uncertainty or avoid unwarranted definitive claims; (ii) Quantitative assessment approaches: These quantitative assessment approach include (a) tracking hallucination rates by comparing model-generated causal attributions against established ground truths; (b) evaluating self-consistency and self-contradictions through minimally perturbed prompts for causal reasoning; (c) provide robust metrics for examining how effectively LLMs manage causal uncertainty. These methodologies constitute a comprehensive framework for diagnosing and ultimately enhancing the robustness of LLMs' causal reasoning under uncertainty.

G Future Research Directions

Numerical (or Comparative) Benchmark Construction for Causal Uncertainty. Current

benchmarks often represent causal uncertainty as uncertain factors rather than through numerical values (e.g., 0.1, 0.9) or epistemic modal expressions (e.g., "most likely", "perhaps") (Rudinger et al., 2020; Cui et al., 2024c). This limitation restricts the ability to quantitatively measure and compare uncertain factors. Therefore, creating datasets that support numerical and comparative quantification is crucial for advancing causal models and enhancing uncertainty assessment methods.

Causal Uncertainty Quantification. Accurate quantification of causal uncertainty is important in high-stake domains such as finance and medical diagnosis. We highlight several promising directions: (i) Quantification with causal interventions: By manipulating specific variables and measuring the effects (Zhang et al., 2020; Wang et al., 2022), we can estimate how interventions impact uncertainty levels, which is vital for robust decision-making in fields like policy-making, medicine, and economics; (ii) Quantification of uncertainty in counterfactual reasoning: Counterfactual reasoning estimates the effects of interventions by comparing actual outcomes to hypothetical scenarios. However, current methods provide a single counterfactual prediction without indicating its uncertainty. Future research is needed to quantify uncertainty in counterfactual scenarios.

Towards Better LLMs for Causal Uncertainty. LLMs have made significant strides in causal reasoning but still face limitations when handling causal uncertainty. Future work should focus on: (i) Enhancing LLMs' consistency and confidence: Improve the reliability of LLMs by ensuring their causal predictions are stable, reproducible, and consistent (Cui et al., 2024b); (ii) Achieving versatility in causal reasoning: Enable LLMs to effectively handle various types of causality, including associational, interventional, counterfactual, data-based, and commonsense causality (Yang et al., 2024; Kiciman et al., 2024; Liu et al., 2023c, 2024b); (iii) Accurate estimation of uncertain influences: Enhance LLMs' ability to identify and quantify the impact of uncertain factors across all uncertainty types discussed in § 2. More details on the desired characteristics for LLMs in causality are provided in § 4.3.

H Glossary of Terms and Concepts

This section presents definitions and examples of key terms and concepts involved in this survey, helping readers unfamiliar with these terminologies. The definitions and illustrative examples of these terminologies are provided in Table 5, including causal uncertainty, causal strength, aleatoric uncertainty, epistemic uncertainty, ontological uncertainty, and predictive uncertainty.

Term	Definition	Example(s)
Causal Uncertainty	Causal uncertainty refers to the ambiguity and unknown factors in identifying, reasoning, or quantifying causal relationships. It encompasses three types of uncertainty: aleatoric, epistemic, and ontological uncertainty.	When investigating the impact of medicine on illness treatment, causal uncertainty may arise from limited clinical data, potential confounders, and difficulty in establishing a solid cause-effect relationship.
Causal Strength	Causal strength is a measure that gauges how strongly a cause leads to the occurrence of its effect.	Under the context of probability raising theory, causal strength could measure smoking's effect on lung cancer by assessing how smoking increases the likelihood of lung cancer.
Aleatoric Uncertainty	Generally, aleatoric uncertainty arises from inherent randomness or unpredictability in the data. It is irreducible with more data. In causality, the aleatoric causal uncertainty is mainly due to natural variability in causal effects, as described in § 2.1.	In medical research, smoking's effect on lung cancer varies from person to person. This variation is irreducible even with more study cases. Similar aleatoric uncertainty examples include that a drug designed to regulate blood sugar levels may have varying effects on different patients due to biological variability.
Epistemic Uncertainty	Epistemic uncertainty refers to the uncertainty caused by incomplete knowledge about causal mechanisms or limitations inherent to the model structure and parameters.	A model predicting medication effects has epistemic uncertainty if it lacks information on patients' diets. Recognizing this helps improve causal models by identifying where additional data could enhance accuracy.
Ontological Uncertainty	Different from aleatoric and epistemic uncertainty, ontological uncertainty is exclusive to causality due to the existential uncertainty of the causal link.	In causality, ontological uncertainty is about the existence or validity of a causal relationship. Examples include the correlation between ice cream sales and drowning incidents(the confounder is hot weather).
Predictive Uncertainty	Predictive uncertainty primarily estimate how uncertain an AI agent is in its prediction (Malinin and Gales, 2018; Ulmer et al., 2022). In § 3, we adapt the uncertainty quantification formula as the method for quantifying causal uncertainty.	Predictive uncertainty has been studied in text classification (Van Landeghem et al., 2022), conditional language generation (Xiao and Wang, 2021), ensemble models (Lakshminarayanan et al., 2017), etc.
Conformal Prediction	Conformal prediction quantifies uncertainty by providing a set of possible outcomes or intervals, expressing uncertainty by giving a range within which the true outcome is likely to lie.	For medical diagnosis, conformal prediction generates a set of potential diagnoses with confidence levels. It aids robust decision-making by providing confidence regions around causal estimates.
Ladder of Causation	The original Judea Pearl's ladder of causation (Pearl, 2009; Pearl and Mackenzie, 2018) consists of three levels: (i) association ($p(e\|c)$); (ii) intervention ($p(y\|\text{do}(x), z)$), and (iii) counterfactual ($p(e_c\|c', e')$).	Association (Seeing): How does watching more TV correlate with students' academic performance? Intervention (Doing): If we restrict the TV time for students, will students' academic performance improve? Counterfactual (Imagining): If I had not watched too much TV, would my academic performance have progressed differently?

Table 5: Glossary of key terms in causal uncertainty.