The Odyssey of Commonsense Causality:
From Foundational Benchmarks to Cutting-Edge Reasoning

Shaobo Cui[1], Zhijing Jin[2], Bernhard Schölkopf[2], Boi Faltings[1]
[1]EPFL, Switzerland, [2]MPI & ETH Zürich
shaobo.cui@epfl.ch, jinzhi@ethz.ch, bs@tue.mpg.de, boi.faltings@epfl.ch

Abstract

Understanding commonsense causality is a unique mark of intelligence for humans. It helps people understand the principles of the real world better and benefits the decision-making process related to causation. For instance, commonsense causality is crucial in judging whether a defendant's action causes the plaintiff's loss in determining legal liability. Despite its significance, a systematic exploration of this topic is notably lacking. Our comprehensive survey bridges this gap by focusing on taxonomies, benchmarks, acquisition methods, qualitative reasoning, and quantitative measurements in commonsense causality, synthesizing insights from over 200 representative articles. Our work aims to provide a systematic overview, update scholars on recent advancements, provide a pragmatic guide for beginners, and highlight promising future research directions in this vital field. A summary of the related literature is available at https://github.com/cui-shaobo/causality-papers .

1 Introduction

> *We do not have knowledge of a thing until we have grasped its why, that is to say, its cause. — Aristotle, 384–322 BC*

Causality (Fisher, 1936; Rubin, 1974; Holland, 1986; Granger, 1988; Pearl, 2009; Pearl and Mackenzie, 2018) has been a cornerstone concept spanning both scientific and philosophical spheres since Aristotle's era (Hocutt, 1974). In this paper, we focus particularly on *commonsense causality*, which we formally define as the general public's intuition about how the occurrence of one event, fact, process, state, or object (the *cause*) plays a role in bringing about or contributing to the happening of another (the *effect*). For example, we know that a rainy morning precipitates traffic congestion or that eating too much leads to weight gain. This innate comprehension of cause-
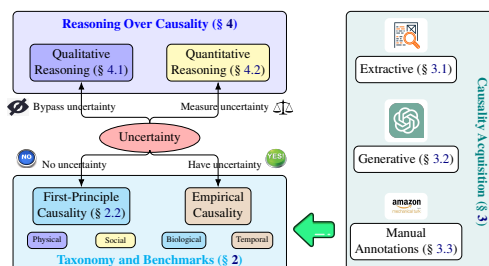

Figure 1: Different aspects of commonsense causality and their link to different sections of this survey.

and-effect dynamics is frequently termed "commonsense causality". It has applications across fields such as medical diagnosis (Richens et al., 2020), psychology (Matute et al., 2015; Eronen, 2020), behavioral science (Grunbaum, 1952), economics (Bronfenbrenner, 1981; Hoover, 2006), and legal systems (Williams, 1961; Summers, 2018) (see more applications in App. A).

Despite its significance, the field still lacks a comprehensive overview of commonsense causality [1]. While there are several survey papers on causal inference (Yao et al., 2021; Zeng and Wang, 2022; Feder et al., 2022) and commonsense knowledge (Storks et al., 2019; Bhargava and Ng, 2022), a comprehensive overview of the intersection of these two domains — commonsense causality — remains missing. The importance of this gap has been further highlighted by recent advancements in large language models (LLMs) (OpenAI et al., 2023; Touvron et al., 2023), which underscore commonsense causality as a pivotal reasoning capability for models. This emerging focus accentuates the urgent need for an in-depth overview. To fill this blank, we conduct an extensive and up-to-date survey of commonsense causality, with comprehensive coverage of its taxonomy, benchmarks, acquisition methods, as well as qualitative and quantitative

---

[1]Causality that relies on specific knowledge or counterintuitive facts, rather than commonsense, is beyond the scope of this paper.

Commonsense Causality

- **Taxonomy of Causality (§2)**
  - **First-Principle Causality (§ 2.2)**
    - **Benchmarks**: ADE (Gurulingappa et al., 2012), CauseEffectPair (Mooij et al., 2016), IHDP (Shalit et al., 2017), CRAFT (Ates et al., 2022),
  - **Empirical Causality**
    - **Text format Benchmarks**:
      Word: SemEval10-T8 (Hendrickx et al., 2010), SemEval20-T5 (Yang et al., 2020).
      Clause: Temporal-Causal (Bethard et al., 2008), EventCausality (Do et al., 2011), BioCause (Mihaila et al., 2013), AltLex (Hidey and McKeown, 2016), TCR (Ning et al., 2018), PDTB (Webber et al., 2019), CausalBank (Li et al., 2020b), SemEval20-T5 (Yang et al., 2020), SCITE (Li et al., 2021)
      Sentence: COPA (Roemmele et al., 2011), CausalTimeBank (Mirza et al., 2014), CaTeRs (Mostafazadeh et al., 2016b), BECauSE (Dunietz et al., 2017), ESL (Caselli and Vossen, 2017), TimeTravel (Qin et al., 2019), XCOPA (Ponti et al., 2020), e-CARE (Du et al., 2022), CoSIm (Kim et al., 2022), CRASS (Frohberg and Binder, 2022), δ-CAUSAL (Cui et al., 2024), COPES (Wang et al., 2023), IfQA (Yu et al., 2023b)
    - **Graph-Format Benchmarks**:
      Word: CausalNet (Luo et al., 2016)
      Phrase: ConceptNet (Speer et al., 2017), Event2Mind (Rashkin et al., 2018), CEGraph (Li et al., 2020b)
      Sentence: ATOMIC (Sap et al., 2019a), ASER (Zhang et al., 2020)
- **Taxonomy of Causality Acquisition (§ 3)**
  - **Extractive Methods (§ 3.1)**
    - **Benchmarks**: SemEval07-T4 (Girju et al., 2007), BioInfer (Pyysalo et al., 2007), CNN-extraction (Do et al., 2011), ADE (Gurulingappa et al., 2012), ESL (Caselli and Vossen, 2017), PDTB (Webber et al., 2019),
    - **Linguistic Pattern&Clues**: (Inui et al., 2003),(Inui et al., 2005), (Khoo et al., 1998) (Sakaji et al., 2008), COATIS (Garcia, 1997), Graphical (Khoo et al., 2000), (Mulkar-Mehta et al., 2011), (Bui et al., 2010), (Doan et al., 2019),
    - **Learning-Based**: (Blanco et al., 2008),CRF (Mihăilă and Ananiadou, 2013), ILP (Gao et al., 2019), Random Forest (Barik et al., 2017), Transfer (Kyriakakis et al., 2019), (Yu et al., 2019),(Hassanzadeh et al., 2020) (Dasgupta et al., 2018), BERT-MLP (Akl et al., 2020), BiLSTM-CRF (Li et al., 2021), KCNN (Li and Mao, 2019)
    - **Hybrid**: Pundit (Radinsky et al., 2012), CATENA (Mirza and Tonelli, 2016), Rule&Supervised(Son et al., 2017)
  - **Generative Methods (§ 3.2)**: Event2Mind (Rashkin et al., 2018), ATOMIC (Sap et al., 2019a), CauseWorks (Choudhry, 2020), GuidedCE (Li et al., 2020b), DISCO (Chen et al., 2023)
  - **Manual Annotation (§ 3.3 and App. F.2)**:
    General: PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016), PDTB (Prasad et al., 2008), RST (Mann and Thompson, 1988), AMT (Banarescu et al., 2013)
    Specifically designed for Causality: BioCause (Mihaila et al., 2013), TimeML (Mirza et al., 2014), RED (Ikuta et al., 2014), CaTeRs (Mostafazadeh et al., 2016b), CxG (Dunietz, 2018).
  - **Implicit/Inter-Sentential Causation Acquisition (App. F)**:
    Implicit: utilizing external knowledge base (Ittoo and Bouma, 2011; Kruengkrai et al., 2017); Learning-Based (Airola et al., 2008; Kruengkrai et al., 2017).
    Inter-Sentential: Language pattern (Wu et al., 2012; Oh et al., 2013);
- **Reasoning Over Causality (§ 4)**
  - **Qualitative Reasoning (§ 4.1)**: NLP models as causal KB: TimeTravel (Qin et al., 2019), CRM (Feng et al., 2021), Neuro-symbolic: (i) causal inference: ROCK (Zhang et al., 2022), COLA (Wang et al., 2023); (ii) temporal constraint: CCM (Ning et al., 2018); CaTeRs (Mostafazadeh et al., 2016b); (iii) logic rules: (Zhang and Foo, 2001; Bochman, 2003; Saki and Faghihi, 2022).
  - **Quantitative Measurement(§ 4.2)**: Word Co-Occurrence: WordCS (Luo et al., 2016), CEQ (Du et al., 2022), CESAR (Cui et al., 2024) Relation Words: ROCK (Zhang et al., 2022), COLA (Wang et al., 2023)
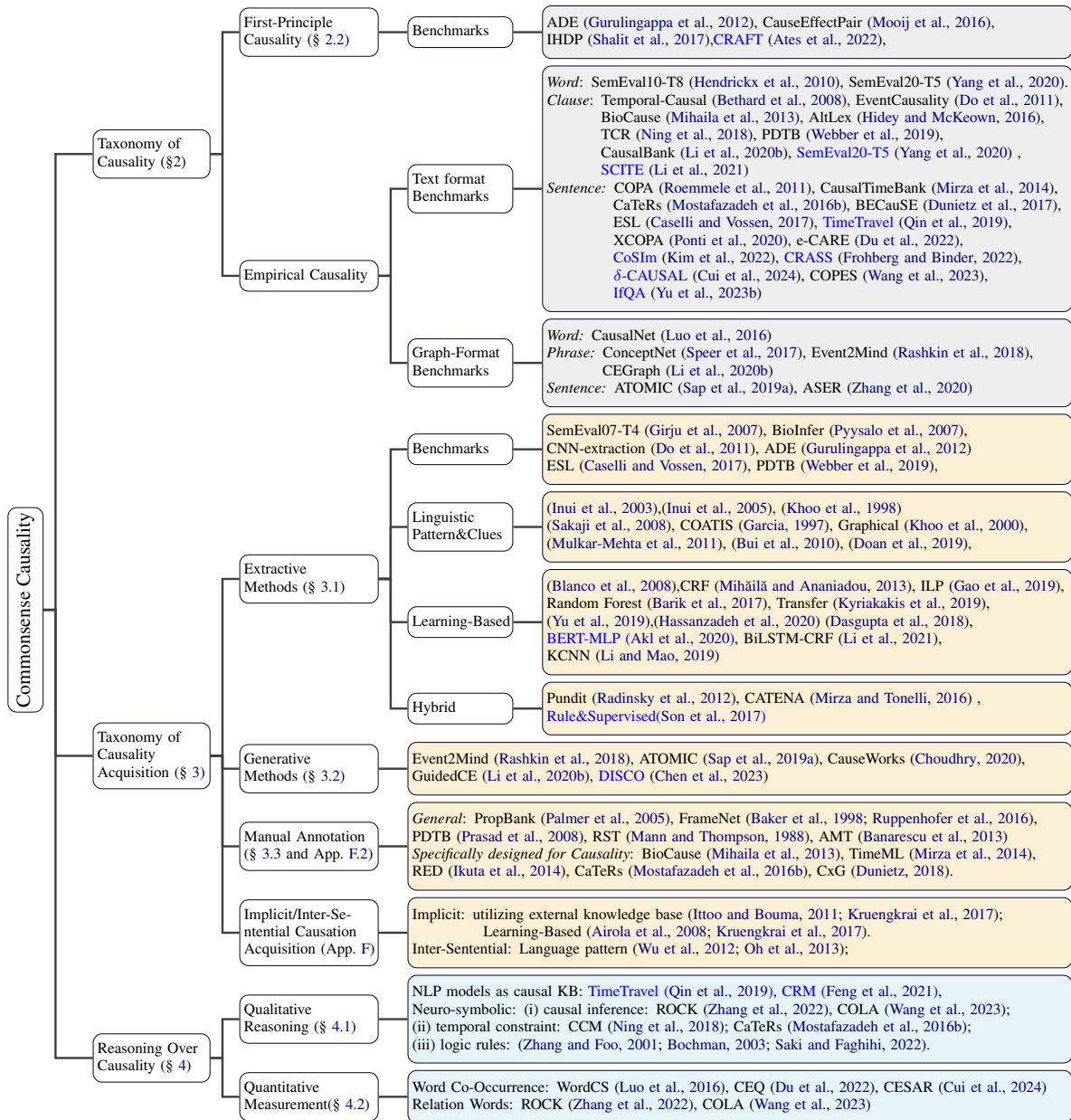
Figure 2: Taxonomy of commonsense causality in various aspects. The benchmarks, datasets, and methods in blue color are about counterfactual. Leaf nodes with different colors are associated with different sections of this survey.

reasoning approaches.

We start by presenting a taxonomy of commonsense causality based on different types of commonsense knowledge (e.g., physical, social, biological, and temporal commonsense) and different levels of uncertainty (§ 2). Leveraging this taxonomy, we methodically categorize 37 existing benchmarks to provide a structured overview. Following this, we discuss three main approaches to acquiring benchmarks conducive to commonsense causality research: extractive (§ 3.1), generative (§ 3.2), and manual annotation methods (§ 3.3). Beyond introducing each approach, we also systematically

compare the merits and demerits of these three approaches, providing insights for future work on commonsense causality acquisition.

Furthermore, we classify the existing causality reasoning methods into two categories based on their way of managing the intrinsic uncertainty within commonsense causality. The first type is qualitative approaches (§ 4.1), which simplify causal reasoning as a classification task and bypass the uncertainty. The second type is quantitative approaches (§ 4.2), which employ metrics to measure causal strength, thereby quantifying the uncertainty. This classification not only aids in understanding

the diverse methodologies but also highlights the varied strategies employed to tackle uncertainty in commonsense causality reasoning.

Lastly, we suggest several promising directions in the field of commonsense causality in § 5. These topics include the exploration of contextual nuances, the analysis of complex structures, the measurement of probabilistic causality, the understanding of temporal dynamics, and the integration of multimodal data. This exploration aims to offer a roadmap for future research.

The contributions of our survey are threefold:

- We present the first comprehensive overview of commonsense causality, synthesizing insights from over 200 representative papers to provide a broad perspective on this topic.

- We methodically review existing benchmarks, acquisition approaches, and reasoning methods by establishing an overall taxonomy, thus offering a useful road map for this field.

- We propose potential research directions for future works and provide a pragmatic handbook for researchers, along with substantial appendices covering a wide range of related topics and preliminary knowledge.[2]

Paper Selection. Our review focused on articles related to commonsense causality from leading peer-reviewed venues in NLP and AI research, such as ACL, EMNLP, NAACL, AAAI, NeurIPS, ICLR, ICML, and IJCAI. We utilized a keyword-based selection strategy, prioritizing papers featuring terms like "causality", "acquisition", "causal reasoning", and "commonsense" in their titles or abstracts. Additionally, we explored GitHub repositories related to causal NLP papers to complement our search. There are also some papers from the philosophy community that help illustrate the concepts related to causality.

The Scope of This Survey. Determining the precise end line of the scope for this survey presents a significant challenge: the domain of commonsense reasoning encompasses a vast area, within which causality plays a crucial role across a substantial

---

[2]Due to the page limit, we present a main overview of commonsense causality research in the main text. We also provide extensive supplementary information in Apps A to L, covering applications, preliminary knowledge, related survey works, other taxonomies, details of uncertainty, acquisition methods, and benchmarks, concepts of causality, NLP techniques, linguistic causality, causal inference, and handbook for beginners.

portion. Nevertheless, each dataset and reasoning methods covered in this survey explicitly incorporates the concept of causality and commonsense, either through its designation or its inherent characteristics. Exclusions are made for datasets that focus on non-causal reasoning, such as Social Chemistry 101 (Forbes et al., 2020), datasets pertaining to generic logical reasoning (e.g., ProofWriter), among others, which constitute a separate category.

2 Taxonomy and Benchmarks

Benchmarks	Annotation Unit	#Overall	#Causal	C.F.[1]	Type
First-principle causality					
CauseEffectPairs (Mooij et al., 2016)	Variable	108	108	◻	*
IHDP (Shalit et al., 2017)	Variable	2,000	2,000	◧	BioC
CRAFT (Ates et al., 2022)	Video	58,000	-	■	PhysC
Empirical causality in text format					
Temporal-Causal (Bethard et al., 2008)	Clause	1,000	271	◻	TempC
CW (Ferguson and Sanford, 2008)	Clause	128	128	■	*
SemEval07-T4 (Girju et al., 2007)	Phrase	220	114	◻	*
SemEval10-T8 (Hendrickx et al., 2010)	Phrase	10,717	1,331	◻	*
COPA (Roemmele et al., 2011)	Sentence	2,000	1,000	◻	*
EventCausality (Do et al., 2011)	Clause	583	583	◻	*
BioCause (Mihaila et al., 2013)	Clause	851	851	◻	BioC
CausalTimeBank (Mirza et al., 2014)	Sentence	318	318	◻	TempC
CBND (Boué et al., 2015)	Sentence	120	120	◻	BioC
CaTeRs (Mostafazadeh et al., 2016b)	Sentence	2,502	308	◻	TempC
AltLex (Hidey and McKeown, 2016)	Clause	44,240	4,595	◻	*
BECauSE (Dunietz et al., 2017)	Sentence	729	554	◻	*
ESL (Caselli and Vossen, 2017)	Sentence	2,608	2,608	◻	TempC
TCR (Ning et al., 2018)	Clause	172	172	◻	TempC
SocialIQa (Sap et al., 2019b)	Sentence	37,588	-	◻	SocC
PDTB (Webber et al., 2019)	Clause	7,991	7,991	◻	*
TimeTravel (Qin et al., 2019)	Sentence	109,964	29,849	◧	*
GLUCOSE (Mostafazadeh et al., 2020)	Clause	670K	670K	◻	SocC
XCOPA (Ponti et al., 2020)	Sentence	11,000	11,000	◻	*
CausalBank (Li et al., 2020b)	Clause	314M	314M	◻	*
SemEval20-T5 (Yang et al., 2020)	Clause	25,501	25,501	◻	*
e-CARE (Du et al., 2022)	Sentence	21,324	21,324	◻	PhysC
CoSIm (Kim et al., 2022)	Image&Text	3,500	3,500	■	*
CRASS (Frohberg and Binder, 2022)	Sentence	274	274	■	*
COPES (Wang et al., 2023)	Sentence	1,360	1,360	◻	*
IfQA (Yu et al., 2023b)	Sentence	3,800	3,800	■	SocC
CW-extended (Li et al., 2023)	Sentence	10,848	10,848	■	*
CausalQuest (Ceraolo et al., 2024)	Sentence	13,500	13,500	◧	*
δ-CAUSAL (Cui et al., 2024)	Sentence	11,245	11,245	◧	*
Empirical commonsense causality in knowledge graph format					
CausalNet (Luo et al., 2016)	Word	11M	11M	◻	*
ConceptNet (Speer et al., 2017)	Phrase	473,000	-	◻	*
Event2Mind (Rashkin et al., 2018)	Phrase	25,000	-	◻	SocC
ATOMIC (Sap et al., 2019a)	Sentence	877K	-	◧	SocC
ASER (Zhang et al., 2020)	Sentence	64M	494K	◻	*
CauseNet (Heindorf et al., 2020)	Word	11M	11M	◻	*
CEGraph (Li et al., 2020b)	Phrase	89.1M	89.1M	◻	*

Table 1: Overview of commonsense causality datasets. A more detailed version is present in App. G.

Different classification criteria lead to different taxonomies for commonsense causality. We build our criteria based on commonsense types (§ 2.1) and uncertainty levels (§ 2.2). This section corresponds to the context marked in light gray color in Figure 2.

---

[2]*C.F.* denotes whether the dataset contains *counterfactual* reasoning, which can vary from no counterfactuals (◻), a subset being counterfactuals (◧), to all counterfactuals (■). For the commonsense type (*Type*), * means that the dataset covers multiple commonsense types.

2.1 Classification by Commonsense Types

According to the commonsense types (App. B.1) on which causality is built, commonsense causality can be roughly classified into four categories: (i) *Physical causality* (PhysC) refers to the commonsense cause-effect relationships grounded in the physical world. PhysC usually covers domains such as physics, chemistry, and environmental science, with datasets such as CRAFT (Ates et al., 2022) and e-CARE (Du et al., 2022); (ii) *Social causality* (SocC) involves the understanding of social norms, cultures, human behavior, intents, and reactions. For instance, criticism (cause) leads to depression (effect) in a social context. SocC covers domains like law, culture, education, psychology, etc. Typical examples are ATOMIC (Sap et al., 2019a), GLUCOSE (Mostafazadeh et al., 2020), and IfQA (Yu et al., 2023b); (iii) *Biological causality* (BioC) relates to cause-effect pairs that govern biological processes and phenomena such as a healthy diet contributes to longevity. Typical benchmarks include BioCause (Mihaila et al., 2013), CBND (Boué et al., 2015), etc; (iv) *Temporal causality* (TempC) involves the sequential understanding that the cause must precede the effect in time (Imbens et al., 2022; Goffrier et al., 2023). This type includes Temporal-Causal (Bethard et al., 2008), CausalTimeBank (Mirza et al., 2014), CaTeRs (Mostafazadeh et al., 2016b), etc.

2.2 Classification by Uncertainty Levels

Sources of Uncertainty. Generally, commonsense causality usually involves unobserved facts and uncertainties. For instance, the claim that "eating a healthy diet and exercising regularly" leads to "a long life" does not reveal/consider the influence of other factors including genetics, access to healthcare, accidents, and so on. Based on the criteria of causal sufficiency and necessity (App. I), there are two kinds of uncertainties in commonsense causality (Yarlett and Ramscar, 2019): (i) *Factual uncertainties* refers to uncertainties caused by insufficient information. This is pervasive in commonsense causality since the knowledge humans possess is always incomplete. For instance, the claim that "rain makes roads slippery" does not reveal detailed information about the type of roads (asphalt, concrete, gravel, earth, chip seal, cobblestones, pervious concrete, etc.) and the intensity of the rain. The missing of these important information influences the validity of causality; (ii)

Causal uncertainties concerns uncertainties due to unstable observation about the cause-effect relation. One example is the claim that "smoking leads to lung cancer". Although there is overwhelming evidence that smokers have a high incidence of lung cancer, there are always some people who smoke a lot but do not develop lung cancer. See more factual and causal uncertainty details in App. E.

Categorization by Levels of Uncertainty. Depending on the level of uncertainty, commonsense causality can be categorized into two types: first-principle causality and empirical causality:

- **First-principle causality** refers to causal relationships grounded in established laws, such as the link between mass and gravity. Usually, first-principle causality is based on fully observed, well-defined, proven settings based on definite physical or mathematical facts.

- **Empirical causality** is prone to suffer from various sources of uncertainties. For instance, it is common knowledge that stepping on a banana peel causes one to slip. However, the validity of this causal relationship is influenced by factors such as the condition of the banana peel(e.g., fresh or dried, factual uncertainties), the condition of the roads (is stepping on the banana peel is the real cause or the wet or oily surface of the road is the true causes, causal uncertainties).

Existing benchmarks, categorized by the two criteria aforementioned, are summarized in Table 1. Further classifications based on skill sets and entity types are detailed in App. D.

3 Causality Acquisition

Common methods for acquiring commonsense causality benchmarks are categorized into three main approaches: extractive methods (§ 3.1), generative methods (§ 3.2), and manual annotation methods (§ 3.3). These methods are summarized in Figure 2 with a hidden orange background color

3.1 Extractive Methods

Benchmarks. The automatic extraction methods are based on annotated domain corpus: open-source text and standard benchmarks. The open-source corpus generally refers to the content available on web pages or Wikipedia. The standard benchmarks cover a variety of datasets such

Form	Connectives
Cause-Effect Connectives	
Cause-Effect	as, because, cause, since, bring about, due to, lead to, owing to, resulting in
Consequence	accordingly, as a result, consequently, for this reason, hence, so, therefore, thus
Reason	in light of, given that, on account of, by reason of, for the sake of, inasmuch as, seeing that
Intention	so that, in order to, so as to, with the aim of, for the purpose of, with this in mind, in hopes of
Conditions	if...then, provided that, assuming that, as long as, unless, in the event that
Source	arises from, stems from, comes from, originates from
Counterfactual Connectives	
Hypothetical	had...then, if it hadn't been for, had it not been for, if only
Negation	were it not for, but for, if it weren't for, without, in the absence of, lacking

Table 2: Common causality-related connectives. The presence of these connectives usually implies the existence of causal relations, which is commonly used in extractive methods.

as SemEval07-T4 (Girju et al., 2007), CNN-extraction (Do et al., 2011), ESL (Caselli and Vossen, 2017) and PDTB (Webber et al., 2019) from the general domain, as well as BioInfer (Pyysalo et al., 2007) and ADE (Gurulingappa et al., 2012) from the biomedical domain. A detailed description of these benchmarks is presented in App. G.

Linguistic Pattern Matching Methods. The methods for extracting causality from text by linguistic pattern matching can be either *clue*-based or *rule*-based. (i) The clue-based approach (Sakaji et al., 2008; Cao et al., 2014) relies on hand-crafted or automatically generated clues to detect the presence of causation. For instance, the presence of the words "cause" or "accordingly" always indicates causality. We list common causal connectives in Table 2; (ii) The pattern/rule-based approach (Girju, 2003; Cole et al., 2006; Ishii et al., 2010) predefines a specific semantic format for extracting causality from text. One common format is a *noun phrase*, a *causation verb* (see App. J.2 for a detailed list of causation verbs), and another *noun phrase or an object complement*. We provide an example sentence within this format in Figure 3.

Machine and Deep Learning-Based Methods. Machine learning-based methods use traditional machine learning models like Support Vector Ma-
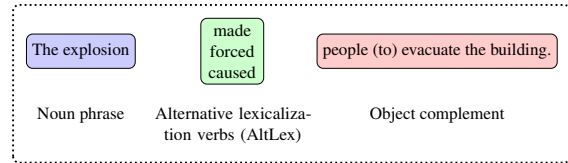
Figure 3: A template of pattern matching from AltLex (Hidey and McKeown, 2016).

chines (SVMs) (Cortes and Vapnik, 1995) or Decision Trees (DTs) (Quinlan, 1986) to detect the presence of causal relationships. The hand-crafted or automatically generated textual features, e.g., dependency parsing features, causal patterns (Girju, 2003; Blanco et al., 2008), the presence of causatives, causal connectives (Zhao et al., 2016) are taken as the input features to the machine learning models, which are then trained to learn the causal extractor. In addition to the conventional machine learning techniques, with the recent success of deep neural networks in various tasks, the deep learning models especially the pre-trained language models provide a more powerful engine for causality extraction.

3.2 Generative Methods

The rapid advance of generative language models like T5 (Raffel et al., 2020) and ChatGPT (OpenAI et al., 2023) enables the LLMs to be useful tools for generating reliable cause-effect pairs (Kim et al., 2023). Rashkin et al. (2018) utilize an encoder-decoder structure for generating intents/reactions over a range of daily events, which contains a variety of causal relationships. Cause-Works (Choudhry, 2020) is a generative method that converts causal graphs into textual narratives of causal relationships. Li et al. (2020b) firstly utilize pattern matching to build a causal graph Causal-Bank, and then employ a Sequence-to-Sequence model to generate the textual cause-effect pairs.[3]

3.3 Manual Annotation

Apart from the automatic extraction strategies, manual annotation is also an important approach for collecting commonsense causality benchmarks. There are plenty of general annotation schemes in semantic parsing that introduce the causation

---

[3]Note that although some works (Madaan et al., 2021; Robeer et al., 2021; Wu et al., 2021; Calderon et al., 2022; Chen et al., 2023) focusing on counterfactual generation, some of them are more on the side of adversarial/fake samples generation instead of the counterfactual meaning in causal reasoning.

as *one of* the semantic relations to be annotated. Some representative schemes include PropBank (Palmer et al., 2005), FrameNet (Baker et al., 1998; Ruppenhofer et al., 2016), PDTB (Prasad et al., 2008), RST (Mann and Thompson, 1988), AMT (Banarescu et al., 2013) and so on. Besides general schemes, there are schemes designed exclusively for annotating causal relations. For example, BioCause (Mihaila et al., 2013), TimeML (Mirza et al., 2014), RED (Ikuta et al., 2014), CaTeRs (Mostafazadeh et al., 2016b), and CxG (Dunietz, 2018) all fall into this framework. More discussion on annotation schemes is in App. F.2.

3.4 Comparison of Data Acquisition Methods

The summary of the pros and cons of these acquisition methods is presented in Table 3. Generally, compared to extractive and generative methods, manual annotation provides the highest quality data and is more explainable. However, it suffers from cost and efficiency issues and thus lacks scalability and coverage. We refer to App. F.3 for a more detailed comparison.

Method	Accuracy	Cost	Coverage	Explainablity
Extractive	★★★★☆	★★★★★	★★★★☆	★★★★☆
Generative	★★★☆☆	★★★★☆	★★★☆☆	★☆☆☆☆
Manual Annotation	★★★★★	★★☆☆☆	★★★★★	★★★★★

Table 3: Comparison of different commonsense causality acquisition methods. The more solid stars, the better.

The aforementioned methods are mainly targeted at explicit causality acquisition and they are more centered on causality inside the sentence. However, causality is not always explicit and may appear in different sentences. More details about implicit causal relationships and inter-sentential causality can be found in App. F.1.

4 Reasoning Over Causality

This section reviews qualitative and quantitative causal reasoning approaches for addressing uncertainty in commonsense causality, as discussed in § 2.2. Qualitative methods (§ 4.1) treat causal reasoning as a 0/1 classification task, while quantitative methods (§ 4.2) quantify causality strength numerically. This section relates to the content highlighted in pale blue color in Figure 2.

4.1 Bypassing Uncertainty by Qualitative Causal Reasoning

Scaling NLP Models as Causal Knowledge Bases. The evolution of commonsense reasoning is in parallel with the advancement of NLP models. NLP models can be used as the causal knowledge bases that are distilled from the training data or pre-training corpora. NLP models experienced four stages of development: (i) Statistical Methods: The initial approach in NLP analyzes patterns and linguistic correlation of text resources to identify causal relationships. They are solely based on term co-occurrence and thus suffer from complex causal structures; (ii) Deep Learning Methods: Methods based on neural network architectures, especially recurrent neural networks and later transformers, are more capable of capturing contextual information. Consequently, they show substantial improvements in the identification and analysis of causal relationships in text; (iii) Pre-Trained Language Models: Language models like BERT (Devlin et al., 2019) and GPT (Brown et al., 2020) that are trained on large corpora expand the reasoning ability drastically. When fine-tuned for causal/counterfactual reasoning tasks, they can not only identify the causal relationship but also comprehend the subtleties inherent in commonsense causality such as implicit causality, temporal constraints, etc; (iv) LLMs (OpenAI et al., 2023; Jiang et al., 2023; Touvron et al., 2023; Mesnard et al., 2024): We are now in the era of LLMs employed with prompting techniques (Wei et al., 2022; Yu et al., 2023a; Alkhamissi et al., 2023). They enable more accurate understanding, predictions, and explanations of causal and counterfactual scenarios.

A detailed chronological overview of these advancements and their impact on causal reasoning is provided in Figure 8 and App. H.

Neuro-Symbolic Methods. Neuro-symbolic methods represent an innovative approach to computational reasoning, overcoming the limitations of traditional NLP models that struggle with complex, non-linear causal relationships. These methods leverage the synergy of neural networks and symbolic logic, blending the pattern-recognition prowess of the former with the explicit, interpretable reasoning of the latter. We categorize these neuro-symbolic strategies into three distinct subcategories:

- *Reasoning with Causal Inference Rules:* Techniques like ROCK (Zhang et al., 2022) and

COLA (Wang et al., 2023) employ the concept of Average Treatment Effect (ATE) to assess the likelihood of one event causing another. ATE is instrumental in quantifying the effect of a treatment on an outcome, represented as $P(E_i \rightarrow E_j) = p(E_i \prec E_j) - p(\neg E_i \prec E_j)$. Furthermore, Jin et al. (2023b) integrates causal inference steps into chain-of-thought reasoning, a method pioneered by Wei et al. (2022). Preliminary of causal inference is elaborated in App. K.

- *Explicitly Incorporating Temporal Constraints*: Recognizing that the cause must precede the effect in time – a fundamental principle in science – methods like those proposed by Ning et al. (2018) introduce temporal constraints. These constraints aid in causal reasoning, reformulating the problem as an integer linear programming challenge.

- *Integrating Logic Rules*: This approach (Zhang and Foo, 2001; Bochman, 2003; Saki and Faghihi, 2022) involves embedding logic rules directly into the reasoning mechanism, thereby enhancing the model's ability to handle complex, logically-driven tasks and presenting better explainability.

4.2 Measuring Uncertainty by Quantitative Causal Reasoning

While qualitative causal reasoning focuses on distinguishing true cause-effect relationships from erroneous ones, it faces challenges due to uncertainties and the defeasible nature of commonsense causality (Marcos, 2021; Cui et al., 2024). Quantitative approaches aim to address these challenges by measuring the likelihood of a cause leading to an effect, thus providing a nuanced understanding of causality. Existing methods for quantitative causal reasoning can be roughly categorized into two types.

Measurement Based on Event Probability. This body of work adopts a probabilistic perspective on causality, positing that a cause *increases* the likelihood of an effect occurring. This perspective is framed by two principal probability constraints[4]:

$$\begin{cases} P(E|C) > P(E) \\ P(E|C) > P(E|\neg C) \end{cases} \quad (1)$$

---

[4]These probabilistic constraints clear off the challenges of *imperfect regularity* and *irrelevance* but still struggle with the challenges of *asymmetry* and *spurious regularities*. More details can be referred to (Hitchcock, 1997).

where C represents the cause, E denotes the effect, and $\neg C$ signifies any event other than C. These constraints argue that the presence of C elevates the likelihood of E compared to the absence of C or the presence of any alternative event $\neg C$. We summarized several key metrics developed from these two constraints in Table 4. Although these met-

	Formulation
(Good, 1961)	$\log \frac{1 - P(E|\neg C)}{1 - P(E|C)}$
(Suppes, 1973)	$P(E|C) - P(E)$
(Eells, 1991)	$P(E|C) - P(E|\neg C)$
(Pearl, 2009)	$P(E|C)$

Table 4: Probabilistic causal strength metrics.

rics appear intuitive and easy to understand at first glance, they are actually difficult to characterize in practice for the following two reasons. Firstly, accurately estimating the conditional probabilities $P(E|C)$ and $P(E|\neg C)$ is challenging due to linguistic variability. Secondly, the solution space for $\neg C$ is vast and cannot be exhaustively explored. The comparison of these causal strength metrics is illustrated in Figure 4.
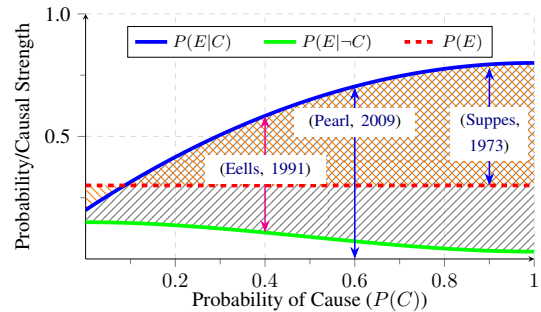

Figure 4: Comparison of different causal strength metrics (Suppes, 1973; Eells, 1991; Pearl, 2009).

Measurement Based on Word Co-occurrences. This approach conceptualizes the causal strength between two events as the cumulative effect of word-level causal strengths of word pairs within these events. The word-level causal strength is measured based on the frequency of word co-occurrences. One example metric CEQ (Luo et al., 2016), which estimates the sentence-level causality by synthesizing the word-level causality.

$$\mathcal{CS}_{\text{CEQ}}(E_1, E_2) = \frac{1}{N_{E_1} + N_{E_2}} \sum_{w_i \in E_1, w_j \in E_2} \text{cs}(w_i, w_j) \quad (2)$$

where N_{E_1} and N_{E_2} are respectively the number of words within the sentences corresponding to the

events E_1 and E_2. $\mathrm{cs}(w_i, w_j)$ is the causal strength between the word w_i and w_j. This word-level causal strength is derived based on the estimation from a large-scale web corpus proposed in (Luo et al., 2016). In contrast to the simple average of word-level causal strengths in CEQ, CESAR (Cui et al., 2024) adopt a weighted aggregation strategy to emphasize word pairs with strong causal indicators, such as "CO_2" and "warming":

$$\mathcal{CS}_{\text{CESAR}}(C, E) = \sum_{e_i \in C} \sum_{e_j \in E} a_{ij} \frac{|e_i^T e_j|}{\|e_i\| \|e_j\|} \quad (3)$$

where e_i and e_j are the causal embeddings for tokens in C and E, respectively. And a_{ij} is the weighting factor. These causal embeddings are generated by a BERT encoder model that is trained on a causal reasoning dataset, which incorporates considerations of uncertainty.

Comparison of Qualitative and Quantitative Causal Reasoning Approaches. We compare the qualitative and quantitative causal reasoning methods from their objectives, applications, merits, and limitations. Please see details in Table 5.

Aspect	Qualitative Reasoning	Quantitative Reasoning
Objectives	To identify the causal relationship between variables.	To provide precise estimates of causal effects.
Merits	(i) Intuitive understanding; (ii) Easy to use.	(i) Precise estimation; (ii) Good comparability across different cause-effect pairs;
Limitations	(i) Lack of precision; (ii) Oversimplication, e.g., confounders are not considered;	(i) Challenges in estimating probabilistic terms like $P(E|C)$, $P(E|\neg C)$, etc; (ii) Need for large amount of high-quality causality data.
Applications	(i) Identification of potential causal relationship between variables; (ii) Simple decision-making tasks where actions are determined by straightforward cause-and-effect determination.	(i) Quantification of uncertainty factor; (ii) Robust decision making; (iii) Comparable analysis for fine-grained causality.

Table 5: Comparison of qualitative and quantitative causal reasoning approaches.

5 Future Research Directions

Contextual Nuances: Exploring Context-Dependent Commonsense Causality. Contextual commonsense causality refers to the phenomenon where cause-effect relationships are valid within specific contexts but may not apply univer-

sally. For instance, while exercise typically benefits health, it can pose risks for individuals with heart conditions, potentially leading to severe consequences. This variability underscores the importance of understanding the contextual dynamics influencing causality. Dupré (1984) introduced the concept of contextual-unanimity causality to capture these contextual nuances:

$$\sum_{B \in \mathbb{B}} P(E|C, B) \times P(B) > \sum_{B \in \mathbb{B}} P(E|\neg C, B) \times P(B)$$
$$(4)$$

where $\mathbb{B}$ represents the set of all potential conditions, contexts, or backgrounds. According to this formulation, the presence of C should increase the average likelihood of E conditional on all conceivable contexts B. Although this formula provides us with the basic idea of describing contextual causality, it contains several quantities that are difficult to obtain. More work is needed in the future to address these issues: (i) Estimation of $P(B)$ and $\mathbb{B}$: Identifying a comprehensive set of conditions $\mathbb{B}$ and characterizing $P(B)$ precisely to minimize contextual unpredictability in commonsense cause-effect relationships; (ii) Partial Contextual Models: Instead of accounting for all possible contexts $\mathbb{B}$, these partial contextual models focus on a subset of contexts $\mathbb{B}' \subseteq \mathbb{B}$ that are deemed most relevant or have the most significant impact on the cause-effect relationship. The objective is to find an optimal $\mathbb{B}'$ such that the model balances between accuracy (in terms of explaining the causality between C and E) and simplicity (minimizing the size of $\mathbb{B}'$). This can be formalized as an optimization problem:

$$\max_{\mathbb{B}' \subseteq \mathbb{B}} \left\{ \sum_{B' \in \mathbb{B}'} P(E|C, B') \times P(B') - \lambda \cdot |\mathbb{B}'| \right\} \quad (5)$$

where λ is a regularization parameter that controls the trade-off between the model's complexity (the number of contexts considered) and its explanatory power for commonsense causality.

Unveiling Complex Structures: Understanding Complex Commonsense Causality. In the domain of commonsense causality, reality often extends beyond simple, direct cause-and-effect relationships to encompass richer, more intricate structures such as confounders, colliders, causal chains, and cyclic causality. Such complex causal frameworks, detailed further in App. K.1, underscore the intricate nature of commonsense causality where multiple variables interact to influence outcomes. Promising topics in this domain include (i) Development of Complex Structure Commonsense

Causality Benchmarks: Creating comprehensive benchmarks that capture the richness of complex structural commonsense causality is the cornerstone of our study for understanding the complexity of real-world causal relationships; (ii) Theoretical Frameworks for Complex Structures Analysis: More efforts should be put into developing theoretical frameworks that are capable of modeling these sophisticated structures. For example, the confounders C_{ij} — variables that influence both the cause X_i and the effect Y_j — can be identified by a structural equation model: $C_{ij} = f(X_i, Y_j)$.

Temporal Dynamics: Unraveling the Role of Time in Commonsense Causality. Temporal dynamics are fundamental to causality, requiring that causes must precede effects. Despite its apparent simplicity, temporal dynamics offer rich future research avenues: (i) Optimal Timing for Intervention: This research aims to determine the best times for interventions that prevent negative outcomes, using causal insights to proactively mitigate risks; (ii) Temporal Patterns of Causal Effects: This direction studies how the impact/effect of a cause varies over time, from immediate, mid-term to long-term effects. This research is vital for informed decision-making, allowing for consideration of an action's extended consequences in the long run.

Beyond Binary: Expanding Probabilistic Perspectives in Causality Measurement. As highlighted in Section 2.2, commonsense causality transcends deterministic frameworks, embodying inherent uncertainties. To navigate and quantify these uncertainties, we suggest two promising research directions that employ a probabilistic perspective: (i) Probabilistic Graphical Models: Developing probabilistic graphical models, such as Bayesian Networks (Heckerman, 2008) or Markov Random Fields, to model probabilistic commonsense causality. The focus would be on characterizing conditional probability distributions $P(E|C)$ that quantify the probabilities of cause-effect relationships; (ii) Dynamic Probabilistic Causal Models with Temporality: This path delves into dynamic causal models that integrate the dimension of time, thereby enhancing the understanding of how causation probabilities evolve over time. This direction might entail the use of differential equations or discrete-time models that estimate $P(E_t|C_{t-\delta})$ — the probability of an effect E at time t given a cause C at a preceding time $t - \delta$.

Expanding Horizons: Advancing Multimodal

Approaches for Commonsense Causality. Multimodal commonsense causality refers to cause-effect pairs whose entities are converted beyond text such as audio, image, and video. The burgeoning availability of multimodal data coupled with advancements in multimodal models (Lu et al., 2019; Chen et al., 2020; Li et al., 2020a) has made the study of commonsense causality both more urgent and achievable. Here we provide several prospective research topics: (i) Advancing Acquisition and Reasoning for Multimodal Commonsense Causality: This topic focuses on developing refined methodologies for the collection and analysis of multimodal data to identify and reason cause-effect relationships within commonsense knowledge; (ii) Cross-Modal Cause-Effect Pair Alignment: It focuses on synchronizing cause-effect pairs across modalities. For example, the cause is a text narrator about deforestation in the Amazon rainforest, while the effect is in videos of trucks carrying logs and the resulting habitat loss for indigenous species. Key challenges involve creating techniques for cross-modal representation and developing robust evaluation metrics for alignment accuracy.

6 Conclusion

In this survey, we present an overview of commonsense causality, including its taxonomy, benchmarks, and data acquisition methods, along with qualitative and quantitative reasoning approaches. Furthermore, we shed light on several future promising research directions. Our work, drawing on insights from over 200 articles, aims to provide a thorough understanding of commonsense causality in the era of LLMs. Additionally, we include a pragmatic handbook in App. L for researchers interested in further exploration of this field.

Limitations

In this study, we provide a survey of commonsense causality in the context of natural language processing. We try our best to provide a bird's-eye view of commonsense causality in an 8-page paper. Notwithstanding our best efforts, this paper still has some limitations. Firstly, it is difficult to cover every aspect of commonsense causality due to the page limit. We choose to focus on specific subtopics including benchmarks, acquisition, qualitative reasoning, and quantitative measurement while the other areas receive less attention. Besides, we focus more on papers already being pub-

lished while not capturing the unpublished works. Notwithstanding our best efforts and an extraordinarily detailed appendix, some relevant work may be unintentionally omitted. Furthermore, commonsense causality is an interdisciplinary area requiring expertise in linguistics, psychology, philosophy, and NLP. It is difficult to delve into each area in a survey paper. We are compelled to engage in prioritization and compromise. We place a greater emphasis on the NLP domain, with the employed methodologies predominantly leaning towards the realm of NLP.

Ethical Considerations

As a survey paper on a commonly addressed NLP task, there are no foreseeable major ethical concerns. All the investigated benchmarks or methods are clearly cited and used in their intended purpose. A minor concern is that while we analyzed the benchmarks, we found that some dataset papers did not provide licenses for using their data, which may cause concerns about ethical usage. Besides, for a broad topic like commonsense causality, oversimplification for certain theories or resources is likely to happen due to the limitation of coverage as well as the concerns raised in the previous limitation section.

Acknowledgements

References

Antti Airola, Sampo Pyysalo, Jari Björne, Tapio Pahikkala, Filip Ginter, and Tapio Salakoski. 2008. A graph kernel for protein-protein interaction extraction. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*, pages 1–9, Columbus, Ohio. Association for Computational Linguistics. 2, 27

Hanna Abi Akl, Dominique Mariko, and Estelle Labidurie. 2020. Semeval-2020 task 5: Detecting counterfactuals by disambiguation. *CoRR*, abs/2005.08519. 2

Badr Alkhamissi, Siddharth Verma, Ping Yu, Zhijing Jin, Asli Celikyilmaz, and Mona Diab. 2023. OPT-R: Exploring the role of explanations in finetuning and prompting for reasoning skills of large language models. In *Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations*

(NLRSE), pages 128–138, Toronto, Canada. Association for Computational Linguistics. 6

Holger Andreas and Mario Guenther. 2021. Regularity and Inferential Theories of Causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. Metaphysics Research Lab, Stanford University. 37

Nabiha Asghar. 2016. Automatic extraction of causal relations from natural language texts: A comprehensive survey. *CoRR*, abs/1605.07895. 25

Fatemeh Torabi Asr and Vera Demberg. 2012. Implicitness of discourse relations. In *International Conference on Computational Linguistics*. 27

Tayfun Ates, M. Ateşoğlu, Çağatay Yiğit, Ilker Kesen, Mert Kobas, Erkut Erdem, Aykut Erdem, Tilbe Goksun, and Deniz Yuret. 2022. CRAFT: A benchmark for causal reasoning about forces and inTeractions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2602–2627, Dublin, Ireland. Association for Computational Linguistics. 2, 3, 4, 30

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*. 2, 6

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics. 2, 6

Biswanath Barik, Erwin Marsi, and Pinar Öztürk. 2017. Extracting causal relations among complex events in natural science literature. In *Natural Language Processing and Information Systems - 22nd International Conference on Applications of Natural Language to Information Systems, NLDB 2017, Liège, Belgium, June 21-23, 2017, Proceedings*, volume 10260 of *Lecture Notes in Computer Science*, pages 131–137. Springer. 2

Steven Bethard, William Corvey, Sara Klingenstein, and James H. Martin. 2008. Building a corpus of temporal-causal structure. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). 2, 3, 4, 30

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pretrained language models: A survey. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The*

Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, pages 12317–12325. AAAI Press. 1, 24, 25

Eduardo Blanco, Nuria Castell, and Dan Moldovan. 2008. Causal relation extraction. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). 2, 5

Alexander Bochman. 2003. A logic for causal reasoning. In *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, pages 141–146. Morgan Kaufmann. 2, 7

Kenneth A Bollen and Judea Pearl. 2013. Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*, pages 301–328. Springer. 40

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258. 35

Léon Bottou, Jonas Peters, Joaquin Quiñonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y. Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: the example of computational advertising. *J. Mach. Learn. Res.*, 14(1):3207–3260. 24

Stéphanie Boué, Marja Talikka, Jurjen Willem Westra, William Hayes, Anselmo Di Fabio, Jennifer Park, Walter K. Schlage, Alain Sewer, Brett Fields, Sam Ansari, Florian Martin, Emilija Veljkovic, Renee Kenney, Manuel C. Peitsch, and Julia Hoeng. 2015. Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database*, 2015:bav030. 3, 4

Martin Bronfenbrenner. 1981. Causality in economics. john hicks. *Economic Development and Cultural Change*, 29:860–863. 1, 23, 41, 42

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 6, 36, 42

Quoc-Chinh Bui, Breanndán Ó Nualláin, Charles A. Boucher, and Peter M. A. Sloot. 2010. Extracting causal relations on HIV drug resistance from literature. *BMC Bioinform.*, 11:101. 2

Nitay Calderon, Eyal Ben-David, Amir Feder, and Roi Reichart. 2022. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland. Association for Computational Linguistics. 5

Angela Cao, Gregor Williamson, and Jinho D. Choi. 2022. A cognitive approach to annotating causal constructions in a cross-genre corpus. In *Proceedings of the 16th Linguistic Annotation Workshop (LAW-XVI) within LREC2022*, pages 151–159, Marseille, France. European Language Resources Association. 28

Yanan Cao, Peng Zhang, Jing Guo, and Li Guo. 2014. Mining large-scale event knowledge from web text. In *Proceedings of the International Conference on Computational Science, ICCS 2014, Cairns, Queensland, Australia, 10-12 June, 2014*, volume 29 of *Procedia Computer Science*, pages 478–487. Elsevier. 5

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics. 2, 3, 5, 28, 31, 42

Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. 2024. Causalquest: Collecting natural causal questions for AI agents. *CoRR*, abs/2405.20318. 3, 33

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics. 35

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and

Jingjing Liu. 2020. UNITER: universal image-text representation learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer. 9

Zeming Chen, Qiyue Gao, Antoine Bosselut, Ashish Sabharwal, and Kyle Richardson. 2023. DISCO: Distilling counterfactuals with large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5514–5528, Toronto, Canada. Association for Computational Linguistics. 2, 5

Arnaud Chiolero. 2019. Causality in public health: one word is not enough. *American Journal of Public Health*, 109(10):1319. 23

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics. 35

Arjun Choudhry. 2020. *Narrative Generation to Support Causal Exploration of Directed Graphs*. Ph.D. thesis, Virginia Tech. 2, 5

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113. 35

Stephen V Cole, Matthew D Royal, Marco G Valtorta, Michael N Huhns, and John B Bowles. 2006. A lightweight tool for automatically extracting causal relationships from text. In *Proceedings of the IEEE SoutheastCon 2006*, pages 125–129. IEEE. 5

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Mach. Learn.*, 20(3):273–297. 5

Shaobo Cui, Lazar Milikic, Yiyang Feng, Mete Ismayilzada, Debjit Paul, Antoine Bosselut, and Boi Faltings. 2024. 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. 2, 3, 7, 8, 33

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. 2

Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Comput. Surv.*, 56(4). 24, 25

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. 6, 35, 42

Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active prompting with chain-of-thought for large language models. *CoRR*, abs/2302.12246. 36

Quang Do, Yee Seng Chan, and Dan Roth. 2011. Minimally supervised event causality identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 294–303, Edinburgh, Scotland, UK. Association for Computational Linguistics. 2, 3, 5, 25, 30

Son Doan, Elly W. Yang, Sameer S. Tilak, Peter W. Li, Daniel S. Zisook, and Manabu Torii. 2019. Extracting health-related causality from twitter messages using natural language processing. *BMC Medical Informatics Decis. Mak.*, 19-S(3):71–77. 2

Brett Drury, Hugo Gonçalo Oliveira, and Alneu de Andrade Lopes. 2022. A survey of the extraction and applications of causal relations. *Nat. Lang. Eng.*, 28(3):361–400. 25

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. 2, 3, 4, 24, 32

Jesse Dunietz. 2018. Annotating and automatically tagging constructions of causal language. *CARNEGIE MELON UNIVERSITY DISSERTATION*. 2, 6, 28, 42

Jesse Dunietz, Lori Levin, and Jaime Carbonell. 2017. The BECauSE corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104, Valencia, Spain. Association for Computational Linguistics. 2, 3, 31

John Dupré. 1984. Probabilistic causality emancipated. *Midwest Studies in Philosophy*, 9:169–175. 8, 42

Ellery Eells. 1991. *Probabilistic causality*, volume 1. Cambridge University Press. 7, 37, 42

Markus I. Eronen. 2020. Causal discovery and the problem of psychological interventions. *New Ideas in Psychology*, 59:100785. 1, 23, 41, 42

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158. 1, 25

Fuli Feng, Jizhi Zhang, Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2021. Empowering language understanding with counterfactual reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2226–2236, Online. Association for Computational Linguistics. 2

Heather J Ferguson and Anthony J Sanford. 2008. Anomalies in real and counterfactual worlds: An eye-movement investigation. *Journal of Memory and Language*, 58(3):609–626. 3, 30, 33

Richard Fisher. 1936. Design of experiments. *British Medical Journal*, 1(3923):554. 1

Branden Fitelson and Christopher Hitchcock. 2011. Probabilistic measures of causal strength. *Causality in the Sciences*, pages 600–627. 25

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics. 3

Jörg Frohberg and Frank Binder. 2022. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2126–2140, Marseille, France. European Language Resources Association. 2, 3, 32

Lei Gao, Prafulla Kumar Choubey, and Ruihong Huang. 2019. Modeling document-level causal structures for event causal relation identification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1808–1817, Minneapolis, Minnesota. Association for Computational Linguistics. 2

Daniela Garcia. 1997. Coatis, an NLP system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management, 10th European Workshop, EKAW'97, Sant Feliu de Guixols, Catalonia, Spain, October 15-18, 1997, Proceedings*, volume 1319 of *Lecture Notes in Computer Science*, pages 347–352. Springer. 2

Dan Geiger, Thomas Verma, and Judea Pearl. 1989. d-separation: From theorems to algorithms. In *UAI '89: Proceedings of the Fifth Annual Conference on Uncertainty in Artificial Intelligence, Windsor, Ontario, Canada, August 18-20, 1989*, pages 139–148. North-Holland. 39

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics. 5

Roxana Girju, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, and Deniz Yuret. 2007. SemEval-2007 task 04: Classification of semantic relations between nominals. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18, Prague, Czech Republic. Association for Computational Linguistics. 2, 3, 5, 30

Clark Glymour, Kun Zhang, and Peter Spirtes. 2019. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524. 25

Graham Van Goffrier, Lucas Maystre, and Ciarán Mark Gilligan-Lee. 2023. Estimating long-term causal effects from short-term experiments and long-term observational data with unobserved confounding. In *Conference on Causal Learning and Reasoning, CLeaR 2023, 11-14 April 2023, Amazon Development Center, Tübingen, Germany, April 11-14, 2023*, volume 213 of *Proceedings of Machine Learning Research*, pages 791–813. PMLR. 4

Irving J Good. 1961. A causal calculus (i). *The British journal for the philosophy of science*, 11(44):305–318. 7, 25, 42

Nelson Goodman. 1947. The problem of counterfactual conditionals. *The Journal of Philosophy*, 44(5):113–128. 24

Clive Granger. 1988. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1-2):199–211. 1

Adolf Grunbaum. 1952. Causality and the science of human behavior. *American Scientist*, 40(4):665–689. 1, 23, 41, 42

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J. Biomed. Informatics*, 45(5):885–892. 2, 5

Joseph F Hair Jr, G Tomas M Hult, Christian M Ringle, Marko Sarstedt, Nicholas P Danks, Soumya Ray, Joseph F Hair, G Tomas M Hult, Christian M Ringle, Marko Sarstedt, et al. 2021. An introduction to structural equation modeling. *Partial least squares structural equation modeling (PLS-SEM) using R: a workbook*, pages 1–29. 40

Joshua K. Hartshorne. 2014. What is implicit causality? *Language, Cognition and Neuroscience*, 29(7):804–824. 27

Oktie Hassanzadeh, Debarun Bhattacharjya, Mark Feblowitz, Kavitha Srinivas, Michael Perrone, Shirin Sohrabi, and Michael Katz. 2020. Causal knowledge extraction through large-scale text mining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13610–13611. AAAI Press. 2

Leslie Hayduk, Greta Cummings, Rainer Stratkotter, Melanie Nimmo, Kostyantyn Grygoryev, Donna Dosman, Michael Gillespie, Hannah Pazderka-Robinson, and Kwame Boadu. 2003. Pearl's d-separation: One more step into causal thinking. *Structural Equation Modeling: A Multidisciplinary Journal*, 10(2):289–311. 39

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. 35

David Heckerman. 2008. A tutorial on learning with bayesian networks. In Dawn E. Holmes and Lakhmi C. Jain, editors, *Innovations in Bayesian Networks: Theory and Applications*, volume 156 of *Studies in Computational Intelligence*, pages 33–82. Springer. 9, 42

Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. Causenet: Towards a causality graph extracted from the web. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 3023–3030. ACM. 3, 34

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multiway classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics. 2, 3, 25, 30

Miguel A Hernán and James M Robins. 2010. Causal inference. 25

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics. 2, 3, 5, 25, 31

Christopher Hitchcock. 1997. Probabilistic causation. 7, 37, 42

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780. 35

Max Hocutt. 1974. Aristotle's four becauses. *Philosophy*, 49(190):385–399. 1

Rinke Hoekstra and Joost Breuker. 2007. Commonsense causal explanation in a legal domain. *Artif. Intell. Law*, 15(3):281–299. 23

Thomas Höfer, Hildegard Przyrembel, and Silvia Verleger. 2004. New evidence for the theory of the stork. *Paediatric and perinatal epidemiology*, 18(1):88–92. 39

Paul W Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960. 1

Kevin D. Hoover. 2006. Causality in economics and econometrics. *History of Finance eJournal*. 1, 23, 41, 42

Rei Ikuta, Will Styler, Mariah Hamang, Tim O'Gorman, and Martha Palmer. 2014. Challenges of adding causation to richer event descriptions. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 12–20, Baltimore, Maryland, USA. Association for Computational Linguistics. 2, 6

Guido Imbens, Nathan Kallus, Xiaojie Mao, and Yuhao Wang. 2022. Long-term causal inference under persistent confounding via data combination. *arXiv preprint arXiv:2202.07234*. 4

Takashi Inui, Kentaro Inui, and Yuji Matsumoto. 2003. What kinds and amounts of causal knowledge can be acquired from text by using connective markers as clues? In *Discovery Science, 6th International Conference, DS 2003, Sapporo, Japan, October 17-19,2003, Proceedings*, volume 2843 of *Lecture Notes in Computer Science*, pages 180–193. Springer. 2

Takashi Inui, Kentaro Inui, and Yuji Matsumoto. 2005. Acquiring causal knowledge from text using the connective marker *tame*. *ACM Trans. Asian Lang. Inf. Process.*, 4(4):435–474. 2

Hiroshi Ishii, Qiang Ma, and Masatoshi Yoshikawa. 2010. Causal network construction to support understanding of news. In *43rd Hawaii International International Conference on Systems Science (HICSS-43 2010), Proceedings, 5-8 January 2010, Koloa, Kauai, HI, USA*, pages 1–10. IEEE Computer Society. 5

Ashwin Ittoo and Gosse Bouma. 2011. Extracting explicit and implicit causal relations from sparse, domain-specific texts. In *Natural Language Processing and Information Systems - 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain, June 28-30, 2011. Proceedings*, volume 6716 of *Lecture Notes in Computer Science*, pages 52–63. Springer. 2, 27

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825. 6

Xianxian Jin, Xinzhi Wang, Xiangfeng Luo, Subin Huang, and Shengwei Gu. 2020. Inter-sentence and implicit causality extraction from chinese corpus. In *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part I*, volume 12084 of *Lecture Notes in Computer Science*, pages 739–751. Springer. 28

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. 2023a. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*. 25

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023b. CLadder: Assessing causal reasoning in language models. In *NeurIPS*. 7

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 25

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, ICLR 2024*. OpenReview.net. 25

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361. 36

Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. Extracting causal knowledge from a medical database using graphical patterns. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 336–343, Hong Kong. Association for Computational Linguistics. 2

Christopher SG Khoo, Jaklin Kornfilt, Robert N Oddy, and Sung Hyon Myaeng. 1998. Automatic extraction of cause-effect information from newspaper text without knowledge-based inferencing. *Literary and linguistic computing*, 13(4):177–186. 2

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *CoRR*, abs/2305.00050. 24, 25

Hyounghun Kim, Abhay Zala, and Mohit Bansal. 2022. CoSIm: Commonsense reasoning for counterfactual scene imagination. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 911–923, Seattle, United States. Association for Computational Linguistics. 2, 3, 32

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics. 5

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics. 35

MV Kratenko. 2022. The problem of uncertainty of causality in" medical cases" and ways to solve it.(regarding the evidence level of expert opinion). *Sudebno-meditsinskaia Ekspertiza*, 65(1):62–66. 27

Canasai Kruengkrai, Kentaro Torisawa, Chikara Hashimoto, Julien Kloetzer, Jong-Hoon Oh, and Masahiro Tanaka. 2017. Improving event causality recognition with multiple background knowledge

sources using multi-column convolutional neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3466–3473. AAAI Press. 2, 27

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, and Joan Ginés i Ametllé. 2019. Transfer learning for causal sentence detection. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 292–297, Florence, Italy. Association for Computational Linguistics. 2

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. 35

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240. 35

David K. Lewis. 1973. *Counterfactuals*. Blackwell, Malden, Mass. 37, 42

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. 35, 42

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. 3, 33

Pengfei Li and Kezhi Mao. 2019. Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. *Expert Syst. Appl.*, 115:512–523. 2

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020a. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer. 9

Zhaoning Li, Qi Li, Xiaotian Zou, and Jiangtao Ren. 2021. Causality extraction based on self-attentive

bilstm-crf with transferred embeddings. *Neurocomputing*, 423:207–219. 2

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020b. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3629–3636. ijcai.org. 2, 3, 5, 32, 34

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110. 35

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692. 35

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23. 9

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. 2, 3, 7, 8, 33, 34

Nishtha Madaan, Inkit Padhi, Naveen Panwar, and Diptikalyan Saha. 2021. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13516–13524. AAAI Press. 5

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281. 2, 6

Henrique Marcos. 2021. A study on defeasibility and defeaters in international law: Process or procedure distinction against the non-discrimination rule. *International Courts and the Guarantee of Social Rights*. 7

Fabienne Martin. 2018. Time in probabilistic causation: Direct vs. indirect uses of lexical causative verbs. In *Proceedings of Sinn und Bedeutung*, volume 22, pages 107–124. 38

Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A. Vadillo, and Itxaso Barberia. 2015. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, 6. 1, 23, 41, 42

Peter Menzies. 1989. Probabilistic causation and causal processes: A critique of lewis. *Philosophy of Science*, 56:642 – 663. 42

Peter Menzies and Helen Beebee. 2001. Counterfactual theories of causation. In *The Stanford Encyclopedia of Philosophy*. 37

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Cristian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, and et al. 2024. Gemma: Open models based on gemini research and technology. *CoRR*, abs/2403.08295. 6

Claudiu Mihăilă and Sophia Ananiadou. 2013. What causes a causal relation? detecting causal triggers in biomedical scientific discourse. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 38–45, Sofia, Bulgaria. Association for Computational Linguistics. 2

Claudiu Mihaila, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC Bioinform.*, 14:2. 2, 3, 4, 6, 31

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*. 35

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the TempEval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19, Gothenburg, Sweden. Association for Computational Linguistics. 2, 3, 4, 6, 31

Paramita Mirza and Sara Tonelli. 2016. CATENA: CAusal and TEmporal relation extraction from NAtural language texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 64–75, Osaka, Japan. The COLING 2016 Organizing Committee. 2

Joris M. Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. 2016. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.*, 17:32:1–32:102. 2, 3, 30

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016a. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics. 31, 32

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016b. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics. 2, 3, 4, 6, 28, 31, 42

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics. 3, 4

Rutu Mulkar-Mehta, Christopher A. Welty, Jerry R. Hobbs, and Eduard H. Hovy. 2011. Using part-of relations for discovering causality. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, May 18-20, 2011, Palm Beach, Florida, USA*. AAAI Press. 2

Prerna Nadathur and Sven Lauer. 2020. Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa*, 5:49. 36, 42

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 2278–2288, Melbourne, Australia. Association for Computational Linguistics. 2, 3, 7, 24, 31

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and intersentential causal relations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1733–1743, Sofia, Bulgaria. Association for Computational Linguistics. 2

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David

Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. 1, 5, 6

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106. 2, 6

Judea Pearl. 2000. Causality: Models, reasoning and inference. 40, 42

Judea Pearl. 2009. *Causality*. Cambridge university press. 1, 7, 40, 42

Judea Pearl. 2012. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, August 14-18, 2012*, pages 3–11. AUAI Press. 40

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons. 25

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books. 1

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word

representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics. 35

M Hashem Pesaran and Ron P Smith. 2016. Counterfactual analysis in macroeconometrics: An empirical investigation into the effects of quantitative easing. *Research in Economics*, 70(2):262–280. 24

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: foundations and learning algorithms*. The MIT Press. 25

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. 2, 3, 32

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA). 2, 6

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. Bioinfer: a corpus for information extraction in the biomedical domain. *BMC Bioinform.*, 8. 2, 5

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with language model prompting: A survey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5368–5393, Toronto, Canada. Association for Computational Linguistics. 24, 25

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. 2, 3, 31

J. Ross Quinlan. 1986. Induction of decision trees. *Mach. Learn.*, 1(1):81–106. 5

Kira Radinsky, Sagie Davidovich, and Shaul Markovitch. 2012. Learning causality for news events prediction. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*, pages 909–918. ACM. 2

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67. 5, 35

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics. 2, 3, 5, 34

Hans Reichenbach. 1956. *The Direction of Time*. Dover Publications, Mineola, N.Y. 25

Jonathan G. Richens, Ciarán M. Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, 11. 1, 23, 41, 42

Marcel Robeer, Floris Bex, and Ad Feelders. 2021. Generating realistic natural language counterfactuals. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3611–3625, Punta Cana, Dominican Republic. Association for Computational Linguistics. 5

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI. 2, 3, 24, 30

Neal J Roese. 1994. The functional basis of counterfactual thinking. *Journal of personality and Social Psychology*, 66(5):805. 24

Neal J Roese and Mike Morrison. 2009. The psychology of counterfactual thinking. *Historical Social Research/Historische Sozialforschung*, pages 16–26. 24

Julia M Rohrer. 2018. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in methods and practices in psychological science*, 1(1):27–42. 38

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688. 1

Josef Ruppenhofer, Michael Ellsworth, Myriam Schwarzer-Petruck, Christopher R Johnson, and Jan Scheffczyk. 2016. Framenet ii: Extended theory and practice. Technical report, International Computer Science Institute. 2, 6

Federica Russo and Jon Williamson. 2007. Interpreting causality in the health sciences. *International Studies in the Philosophy of Science*, 21(2):157–170. 23

Hiroki Sakaji, Satoshi Sekine, and Shigeru Masuyama. 2008. Extracting causal knowledge using clue phrases and syntactic patterns. In *Practical Aspects of Knowledge Management, 7th International Conference, PAKM 2008, Yokohama, Japan, November 22-23, 2008. Proceedings*, volume 5345 of *Lecture Notes in Computer Science*, pages 111–122. Springer. 2, 5

Amir Saki and Usef Faghihi. 2022. A fundamental probabilistic fuzzy logic framework suitable for causal reasoning. *CoRR*, abs/2205.15016. 2, 7

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108. 35

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019a. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press. 2, 3, 4, 34

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics. 3

Uri Shalit, Fredrik D. Johansson, and David A. Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3076–3085. PMLR. 2, 3, 30

Brian Skyrms. 1981. Causal necessity. *Philosophy of Science*, 48(2):329–335. 37

Youngseo Son, Anneke Buffone, Joe Raso, Allegra Larche, Anthony Janocko, Kevin Zembroski, H Andrew Schwartz, and Lyle Ungar. 2017. Recognizing counterfactual thinking in social media texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 654–658, Vancouver, Canada. Association for Computational Linguistics. 2

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First*

AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, pages 4444–4451. AAAI Press. 2, 3, 24, 34

Shane Storks, Qiaozi Gao, and Joyce Y. Chai. 2019. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *CoRR*, abs/1904.01172. 1, 25

Andrew Summers. 2018. Common-Sense Causation in the Law. *Oxford Journal of Legal Studies*, 38(4):793–821. 1, 23, 41, 42

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223. 35

Patrick Suppes. 1973. A probabilistic theory of causality. *British Journal for the Philosophy of Science*, 24(4). 7, 25, 42

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112. 35

Charles Sutton and Andrew McCallum. 2012. An introduction to conditional random fields. *Found. Trends Mach. Learn.*, 4(4):267–373. 35

Kumutha Swampillai and Mark Stevenson. 2011. Extracting relations within and across sentences. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 25–32, Hissar, Bulgaria. Association for Computational Linguistics. 28

Leonard Talmy. 1988. Force dynamics in language and cognition. *Cogn. Sci.*, 12(1):49–100. 37, 42

Philip E Tetlock and Aaron Belkin. 1996. *Counterfactual thought experiments in world politics: Logical, methodological, and psychological perspectives*. Princeton University Press. 24

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Kathleen S. Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Agüera

y Arcas, Claire Cui, Marian Croak, Ed H. Chi, and Quoc Le. 2022. Lamda: Language models for dialog applications. *CoRR*, abs/2201.08239. 35

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971. 1, 6, 35

Robert R. Tucci. 2013. Introduction to judea pearl's do-calculus. *CoRR*, abs/1305.5506. 40

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008. 35

Ingo Venzke. 2018. What if? counterfactual (hi) stories of international law. *Asian journal of international law*, 8(2):403–431. 24

T. Vigen. 2015. *Spurious Correlations*. Hachette Books. 39

Edward Volchok. 2015. Three Levels of Causation — media.acc.qcc.cuny.edu. http://media.acc.qcc.cuny.edu/faculty/volchok/causalMR/CausalMR3.html. [Accessed 26-06-2024]. 36

Zhaowei Wang, Quyet V. Do, Hongming Zhang, Jiayao Zhang, Weiqi Wang, Tianqing Fang, Yangqiu Song, Ginny Wong, and Simon See. 2023. 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. 2, 3, 7, 32

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108. 2, 3, 5, 28, 31, 37, 42

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*. 6, 7, 35, 36, 42

Ernest J Weinrib. 2016. Causal uncertainty. *Oxford Journal of Legal Studies*, 36(1):135–164. 27

Glanville Llewelyn Williams. 1961. Causation in the law. *The Cambridge Law Journal*, 19:62 – 85. 1, 23, 41, 42

Jon Williamson. 2009. Probabilistic theories of causality. 25

Phillip Wolff. 2007. Representing causation. *Journal of experimental psychology: General*, 136(1):82. 28

Phillip Wolff and Jason Shepard. 2013. Causation, touch, and the perception of force. In Brian H. Ross, editor, *Psychology of learning and motivation*, volume 58 of *Psychology of Learning and Motivation*, pages 167–202. Academic Press. 28

Phillip Wolff and Robert Thorstad. 2017. Force dynamics. *The Oxford handbook of causal reasoning*, pages 147–168. 37

Jheng-Long Wu, Liang-Chih Yu, and Pei-Chann Chang. 2012. Detecting causality from online psychiatric texts using inter-sentential language patterns. *BMC Medical Informatics Decis. Mak.*, 12:72. 2, 27

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving 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 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics. 5

Jinghang Xu, Wanli Zuo, Shining Liang, and Xianglin Zuo. 2020. A review of dataset and labeling methods for causality extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1519–1531, Barcelona, Spain (Online). International Committee on Computational Linguistics. 25

Jun Xu, Yonghui Wu, Yaoyun Zhang, Jingqi Wang, Hee-Jin Lee, and Hua Xu. 2016. CD-REST: a system for extracting chemical-induced disease relation in literature. *Database J. Biol. Databases Curation*, 2016. 27

Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowl. Inf. Syst.*, 64(5):1161–1186. 25

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. 2020. SemEval-2020 task 5: Counterfactual recognition. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 322–335, Barcelona (online). International Committee for Computational Linguistics. 2, 3, 32

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764. 35

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):74:1–74:46. 1, 25

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *CoRR*, abs/2305.10601. 42

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. 4, 26

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. 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 4664–4674, Hong Kong, China. Association for Computational Linguistics. 2

Ping Yu, Tianlu Wang, Olga Golovneva, Badr AlKhamissi, Siddharth Verma, Zhijing Jin, Gargi Ghosh, Mona Diab, and Asli Celikyilmaz. 2023a. ALERT: adapting language models to reasoning tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Toronto, Canada. Association for Computational Linguistics. 6

Wenhao Yu, Meng Jiang, Peter Clark, and Ashish Sabharwal. 2023b. IfQA: A dataset for open-domain question answering under counterfactual presuppositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8276–8288, Singapore. Association for Computational Linguistics. 2, 3, 4, 33

Liangjun Zang, Cong Cao, Yanan Cao, Yuming Wu, and Cungen Cao. 2013. A survey of commonsense knowledge acquisition. *J. Comput. Sci. Technol.*, 28(4):689–719. 25

Jingying Zeng and Run Wang. 2022. A survey of causal inference frameworks. *arXiv preprint arXiv:2209.00869*. 1, 25

Dongmo Zhang and Norman Y. Foo. 2001. EPDL: A logic for causal reasoning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence, IJCAI 2001, Seattle, Washington, USA, August 4-10, 2001*, pages 131–138. Morgan Kaufmann. 2, 7

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2. 2, 3, 34

Jiayao Zhang, Hongming Zhang, Weijie J. Su, and Dan Roth. 2022. ROCK: causal inference principles for reasoning about commonsense causality. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 26750–26771. PMLR. 2, 6

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2023. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. 36

Sendong Zhao, Ting Liu, Sicheng Zhao, Yiheng Chen, and Jian-Yun Nie. 2016. Event causality extraction based on connectives analysis. *Neurocomputing*, 173:1943–1950. 5

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. 36

A Real-World Applications of Commonsense Causality

Commonsense causality has a wide range of applications in domains like medical diagnosis (Richens et al., 2020), psychology (Matute et al., 2015; Eronen, 2020), behavioral science (Grunbaum, 1952), economics (Bronfenbrenner, 1981; Hoover, 2006), legal systems (Williams, 1961; Summers, 2018). Here we mainly detail two of them which include healthcare assistance (App. A.1) and forensic analysis (App. A.2).

A.1 Healthcare and Medical Assistance

The cornerstones for medicine or healthcare are the investigation of (Russo and Williamson, 2007):

1. What *causes* diseases and pandemics to develop?

2. What medicine and policy could *stop* or *prevent* the disease or pandemic?

For these two core objectives, commonsense causality assists in various aspects:

- Medical Diagnosis: Medical professionals use commonsense to interpret symptoms and link them to particular diseases (Richens et al., 2020)

- Disease Treatment and Prevention Program: A deep comprehension of causal relationships between certain lifestyles and diseases helps people to make better treatment and prevention plans. For instance, knowing that a sedentary lifestyle leads to Type 2 Diabetes will motivate people to exercise more to prevent illness.

- Public Health Strategy: Commonsense causality is important for prudent public health strategy making (Chiolero, 2019). For example, the causal relationship between air pollution and increasing numbers of pulmonary disease patients pushes the government to restrict emissions and promote clean energy.

A.2 Legal and Forensic Analysis

One of the most important applications of commonsense causality is understanding legal causation. As mentioned in Section 2.A of (Summers, 2018), commonsense has been a useful tool in determining legal causation. As Lord Reid put in *Stapley v Gypsum Mines*:

> To determine what caused an accident from the point of view of legal liability is a most difficult task. If there is any valid logical or scientific theory of causation it is quite irrelevant in this connection . . . The question must be determined by <u>applying common sense</u> to the facts of each particular case.

There are various legal scenarios where commonsense causality plays an important role:

- Determining Legal Liability: Establishing causality is crucial for determining legal liability. Commonsense causality is helpful in judging whether a defendant's action leads to the plaintiff's loss (Williams, 1961; Summers, 2018; Hoekstra and Breuker, 2007).

- Investigation of Criminal Intent and Motive: A comprehension of the causal relationship helps to understand the criminal motive. This assists judges with the sentencing of defendants and makes fair decisions. For instance, if one driver hits another car parked on the side of the road, commonsense causality helps to attribute the cause of the incident to the driver.

B Preliminaries and Definitions

In this section, we mainly introduce the preliminary knowledge about commonsense in App. B.1 and then describe the qualitative reasoning tasks in App. B.2. Other more specific preliminary knowledge such as language models, causal concepts, linguistic causality, and causal inference is described in App. H, I, J, and K, respectively. To help readers refer back to the main body of the paper, this section corresponds to § 2.1.

B.1 Commonsense

What Is Commonsense? Commonsense in the domain of NLP refers to widely accepted knowledge that helps the majority of people understand the real world better like "water flows from high to low" and "rain leads to slippery roads". There are some aspects of commonsense: (i) World Knowledge Reasoning: Information about daily life such as "When you are hungry you need to eat food"; (ii) Commonsense Causal Reasoning: Understanding the cause-effect relationship such as "rain makes roads slippery"; (iii) Commonsense Temporal Reasoning: Understanding sequences of events and the concept of time order, e.g., "Dessert usually comes

after the main course"; (iv) Commonsense Spatial Reasoning: Understanding the physical concept of space, e.g., "a ball is placed inside a box instead of a bowl" and "a basketball is usually larger than a table tennis ball"; (v) Social Context: Comprehending the social norm, i.e., the accepted behaviors, practices, and values within a society. For instance, it's customary to bring a small gift when visiting someone's house; (vi) Counterfactual Reasoning: Reasoning over scenarios that didn't happen but could have. For instance, "Had I noticed the 'Wet Floor' sign, I wouldn't have slipped".

Characteristics of Commonsense. Commonsense, by its inherent nature and definition, has distinctiveness like intuitiveness and universality. Beyond that, there are some aspects that are commonly ignored: (i) Contextual Dependency: The applicability of commonsense varies depending on the context. What is considered as commonsense in one culture may not seen in the same way, e.g., the thumbs-up gesture 👍 is viewed as approval in one culture but impoliteness in some other cultures; (ii) Time-Sensitiveness: Commonsense is evolving over time. What was perceived as commonsense previously is not commonsense now. A great example is the understanding of the solar system, it was commonsense to posit that Earth was the center around celestial bodies, i.e., the geocentric model. However, the heliocentric model become common nowadays, which believes that the Sun, rather than the Earth, is at the center; (iii) Error-Proneness or Inherent Uncertainties: Due to the aforementioned time-sensitiveness and contextual dependency, we can easily tell that there are inherent uncertainties in commonsense causality and it is prone to claim fake commonsense causality.

What Is Not Commonsense. In contrast to commonsense, non-commonsense knowledge includes (i) Specialized Knowledge: Knowledge acquired via specific education, training, or experience is not within the realm of commonsense. For instance, comprehension of complex theories of mathematics or legal principles; (ii) Individual Subjectivity: Individual experience on certain cause-and-effect cannot be viewed as commonsense causality. For instance, if a person feels sleepy after drinking milk. Nevertheless, we cannot draw a causal relationship between milk drinking and being sleepy; (iii) Counterintuitive Facts: Some scientific facts are not commonsense knowledge during a certain period. For instance, the Earth revolving around the Sun was once a counterintuitive idea before the 14th century.

B.2 Qualitative Reasoning Tasks Related to Commonsense Causality

Causal Reasoning. Commonsense causal reasoning (CCR) is the task of capturing causal dependencies between one event (the cause) and the other (the effect) based on human knowledge. Generally, these events are in textual format. Datasets like COPA (Roemmele et al., 2011), TCR (Ning et al., 2018), and e-CARE (Du et al., 2022) follows the following format. Each question consists of a premise and two alternatives and the goal is to select the more plausible cause (or effect) of the given premise.

> **Example of Causal Reasoning**
>
> Premise: The man broke his toe. What was the CAUSE of this?
> Alternative 1: He got a hole in his sock.
> Alternative 2: He dropped a hammer on his foot.

Counterfactual Reasoning. Counterfactual reasoning (Goodman, 1947; Bottou et al., 2013) describe possible outcomes that could happen had certain events happened, e.g., "Had I brought an umbrella, I would not get wet". It has been studied in various domains such as Psychology (Roese, 1994; Roese and Morrison, 2009), Law (Speer et al., 2017; Venzke, 2018), Economics (Pesaran and Smith, 2016), Social Science (Tetlock and Belkin, 1996).

C Related Survey

We provide different lines of surveys related to commonsense causality in Table 6. The related surveys can be categorized into five types:

- Surveys of Commonsense Reasoning: These surveys cover works from benchmarks (Davis, 2023) to methods (Bhargava and Ng, 2022; Qiao et al., 2023) about reasoning with commonsense.

- Surveys of Causal Knowledge Acquisition: Existing works cover datasets, methods, and evaluation metrics of the causality acquisition task.

- Surveys of Causal Reasoning With Language Models: Kiciman et al. (2023) examine the

ability of large language models in causal tasks like causal discovery, counterfactual inference, discerning necessary and sufficient causality via solely natural language input.

- Surveys of Causal Inference: Except for textbooks of (Hernán and Robins, 2010; Pearl et al., 2016; Peters et al., 2017), there are surveys (Yao et al., 2021; Zeng and Wang, 2022) covering the benchmarks, application, and frameworks of causal inference.

- Surveys of Probabilistic View of Causality: (Williamson, 2009) review existing probabilistic theories of causality and analyze their failure examples critically.

Our survey sets itself apart by offering a comprehensive exploration of commonsense causality from a language perspective. Unlike the aforementioned surveys focusing on particular aspects, our works provide an overview of commonsense causality, covering the dimensions of benchmarks, taxonomies, acquisition methods, and both qualitative and quantitative measurements.

D More Taxonomies of Commonsense Causality

Different criteria for categorizing commonsense causality lead to the development of distinct taxonomies, each offering a unique perspective on the organization and relationships of commonsense causality. Here we further supplement with taxonomies by skill sets (App. D.1) and the nature of entities involved (App. D.2). This section refers back to § 2.

D.1 Classification by Skill Sets

We can classify the skill sets required by causal reasoning into two high-level types: (1) *Closed book* causality means tasks that can be completed by only looking at the given text, but not recalling external knowledge. This category can test skills such as (a) proper linguistic understanding of the given text, as in information extraction, such as causal relation extraction (Do et al., 2011; Hidey and McKeown, 2016), counterfactual statement identification (Hendrickx et al., 2010), or (b) formal reasoning on the given conditions and statistics, using skills such as causal inference (Jin et al., 2023a), and causal discovery (Jin et al., 2024, 2022). (2) *Open book* causality refers to tasks that require external

Citation	Summary
Commonsense Reasoning	
(Storks et al., 2019)	A survey of existing benchmarks and methods for commonsense reasoning.
(Bhargava and Ng, 2022)	Survey about methods of utilizing pretrained language model for commonsense knowledge reasoning and acquisition.
(Qiao et al., 2023)	Survey of different prompting methods for commonsense reasoning.
(Davis, 2023)	Survey of 139 commonsense benchmarks: 102 text-based, 18 image-based, 12 video-based, and 7 physical simulation-based. Furthermore, this survey presents the definition and role of commonsense in AI, discusses the desirable nature of a commonsense benchmark, and shows the flaws of existing commonsense benchmarks.
Causal Knowledge Acquisition	
(Zang et al., 2013)	Survey about the methods and evaluation of existing commonsense knowledge acquisition systems.
(Drury et al., 2022)	Survey about extraction of causal relationships from text.
(Xu et al., 2020)	Survey of datasets and labeling methods for causality extraction from text.
(Feder et al., 2022)	Survey for adapting important causal inference concepts into textual format.
(Fitelson and Hitchcock, 2011)	Survey of methods for analyzing causal strength via probability.
(Glymour et al., 2019)	A brief review of computational methods for causal discovery including constraint-based, score-based, and functional causal model-based methods.
(Yang et al., 2022)	Survey of causality extraction including taxonomies of causality extraction, benchmark datasets, and extraction techniques.
(Asghar, 2016)	Survey of automatic extraction of causal relationship from natural language.
Causal Reasoning	
(Kiciman et al., 2023)	Survey of large language models' ability in performing causal discovery, which includes effect inference, attribution, and actual causality, and understanding actual causality, which includes counterfactual reasoning, identifying necessary and sufficient causes.
Causal Inference	
(Yao et al., 2021)	A survey about causal inference under the potential outcome framework, benchmarks, and applications.
(Zeng and Wang, 2022)	A review of past works that focus on outcomes framework and causal graphical models of causal inference.
Probabilistic View of Causality	
(Williamson, 2009)	Survey of probabilistic theories of causality, which includes the theories of Reichenbach (Reichenbach, 1956), Good (Good, 1961), and Suppes (Suppes, 1973).

Table 6: Related surveys.

knowledge out of the provided text, which usually includes (a) questions about a causal relation directly, such as asking about the relation between two events, the effect given the cause, or the cause given the effect, or (b) counterfactual reasoning, where an alternative condition is given and asks for the outcome. As indicated in Figure 5, open book causality requires memorization skills and reasoning.
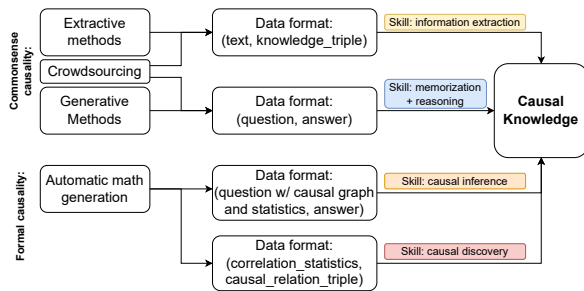

Figure 5: Overview of causal NLP tasks and required skill sets.

D.2 Classification by Nature of Entities Involved

Based on the nature of the entities involved, commonsense causality can be further classified into physical commonsense causality and social commonsense causality. Physical commonsense causality usually involves non-human entities like inanimate objects or natural phenomena. However, social commonsense causality always involves humans, human behavior, social norms, cultures, etc.

- Physical Commonsense Causality: It usually occurs in the context of the physical or natural world and is governed by the laws and principles of mathematics, physics, biology, and physics. Generally, it is more predictable and context-free.

- Social Commonsense Causality: Different from physical causality, social causality is governed by social background, cultural norms, etc. It is less predictable and relies heavily on social context. It is often observed in the domains of sociology, psychology, and related disciplines.

There are many other taxonomies for commonsense causality, which is beyond the scope of this survey.

E Uncertainty in Commonsense Causality

Uncertainty is almost everywhere, no exception for commonsense causality. We summarize all sources of uncertainties over commonsense causality into two categories (Yarlett and Ramscar, 2019): factual uncertainties (App. E.1) and causal uncertainties (App. E.2). This section corresponds to § 2.2.

E.1 Factual Uncertainties

Factual uncertainties are due to the principle that the observation or description of contextual information of the cause or effect can never be complete. The factual uncertainties can be further classified into the following subcategories:

- Incomplete Observation: The observation of the world is hardly complete. For instance, it is the commonsense knowledge that exercise leads to fatigue. However, a small amount of exercise actually makes people more energetic rather than exhausted.

- Contextual Uncertainty: It arises when the context of the cause or effect introduces ambiguity about the facts. For instance, when determining the cause of certain symptoms, the symptom descriptions heavily depend on the medical diagnosis equipment, which causes uncertainty in the determination of the true cause for diagnosis.

- Temporal Uncertainty: Due to the time-sensitive characteristic of commonsense, commonsense is inherently vulnerable to temporal uncertainty. For instance, historically, the need for light (the cause) leads to using candles(the effect). However, after the widespread adoption of electricity and bulbs, this causal relationship doesn't hold anymore.

E.2 Causal Uncertainties

Causal uncertainties arise in cases where the cause is not invariably followed by the effect. For instance, we all know that smoking contributes to the occurrence of lung cancers. However, some people smoke a lot but do not suffer from lung cancer. Similar situations can be found in examples like "clouds lead to rain", but there are days there are a lot of clouds but no rain at all. The causal uncertainties can be further divided into the following subcategories:

- Probabilistic Causation: It refers to the situation wherein causes increase the likelihood of but do not guarantee the occurrence of effects. This is also the focus in the § 4.2. Examples include "not all smokers get lung cancer", "a healthy diet does not guarantee longevity", etc.

- Complex Interaction: Complex causal structures like co-founder, collider, causal chain, triangular causality, and the combination of these basic structures lead to significant complexity and introduce additional uncertainties.

- Causal Loops: Though causal loops can be included in the category of complex interaction, we define them separately, hoping it draws particular attention. There are scenarios where the effect also influences the cause, forming a causal loop. For example, poverty results in poor education opportunities, which in turn aggravates poverty. A similar example in the domain of the environment is the feedback loop between global warming and ice glacier melting. Global warming speeds up the melting of ice glaciers. Without ice to reflect back the sunlight, more solar energy research to the surface of the Earth and thus perpetuate global warming. This phenomenon is also observed in the marketing area: high-quality products reinforce the marketing share, which in turn empowers companies' ability to develop better products.

Besides, the uncertainty of causality in other domains like medical (Kratenko, 2022) and legal domains (Weinrib, 2016) is also investigated. However, due to the page limit, we will not discuss these topics in this survey.

F More Topics on Causality Acquisition

In this section, we cover some supplementary topics related to commonsense causality acquisition, including extraction methods for implicit and inter-sentential causality (App. F.1), and details of manual annotation schemes (App. F.2). This section corresponds to § 3 that is about causality acquisition methods.

F.1 Extraction of Different Kinds of Causality

Extraction of Implicit Causality. Since causality can be expressed in various ways, the extraction of implicit causality (Hartshorne, 2014; Asr and Demberg, 2012) [5] is even more challenging than the extraction of explicit causality with linguistic indicators such as "because", "due to", "lead to", etc.

> **Example of implicit causality**
>
> Tom got caught in a heavy rain yesterday and worked with a fever today.

For implicit causality, it is infeasible to use linguistic patterns to detect the presence of causality. There are two approaches to extracting implicit causality:

- Utilizing External Knowledge Bases: These works (Ittoo and Bouma, 2011; Kruengkrai et al., 2017) utilize external knowledge to enhance implicit causality extraction and alleviate the need for manually annotated data. Xu et al. (2016) used document-level classifier,

- Learning-Based Approach (Airola et al., 2008; Kruengkrai et al., 2017): They use background knowledge and the original sentences as the features to train models for extracting causality. The key limitation is the lack of supervised learning data for model training.

Extraction of Inter-Sentential Causality. Besides, different from intra-sentential causality, wherein inter-sentential causality, the cause and the effect lie in two sentences. As the following example shows, the inter-sentential causal relation between "'paper deadline" and "went to sleep earlier than before" is difficult to identify due to the lack of causal connectives.

> **Example of Inter-Sentential causality**
>
> I was tired last night due to a paper deadline. I went to sleep earlier than before.

For inter-sentential causality, there are two extraction approaches [6]

- Linguistic Pattern Matching: Ittoo and Bouma (2011); Wu et al. (2012) extend the pattern

---

[5]The boundary between explicit causality and implicit causality is unclear. Here, we refer to causality that lacks explicit indicators such as "because", "due to", etc., as implicit causality.

[6]Most of the intra-sentential causality extraction methods still apply to inter-sentential causality well. Here, we only name several methods specifically designed for inter-sentential causality extraction.

Method	Accuracy/Quality	Cost/Efficiency	Coverage	Adaptability	Scalability	Explainablity
Extractive	★★★★☆	★★★★★	★★★★☆	★★★★☆	★★★★★	★★★★☆
Generative	★★★☆☆	★★★★☆	★★★★★	★★★☆☆	★★★★★	★☆☆☆☆
Manual Annotation	★★★★★	★★☆☆☆	★★☆☆☆	★★★★★	★☆☆☆☆	★★★★★

Table 7: Comparison of different commonsense causality acquisition methods. The more solid stars, the better.

matching methods for causality detection to the inter-sentential causality. Jin et al. (2020) propose a cascaded multi-Structure Neural Network (CSNN) to extract inter-sentential causality without dependency on external knowledge.

- Learning-Based Approach: Swampillai and Stevenson (2011) propose an approach that works for both intra-sentential and inter-sentential causality extraction. They use adapted features and techniques to deal with the special issues due to the inter-sentential cases.

F.2 Manual Annotation Schemes of Causation

Existing manual annotation schemes can be roughly classified into three types (Cao et al., 2022):

- Trigger Scheme: A manual annotation scheme based on the template of *cause argument, trigger, effect argument*. Inside the template, triggers usually are conjunctions, adverbials, and causation verbs that indicate causation. Manual annotation schemes like BECausSE (Dunietz, 2018), PDTB (Webber et al., 2019) fall into this category.

- CEP Scheme: A manual annotation scheme based on CAUSE, ENABLE, PREVENT (CEP) causal relationship. CEP scheme is based on the force dynamics theory of causation (Wolff, 2007; Wolff and Shepard, 2013). This category covers manual annotation schemes including CCEP (Cao et al., 2022), CaTeRS (Mostafazadeh et al., 2016b), and BECauSE (Dunietz, 2018).

- Joint Scheme: A manual annotation scheme that jointly annotates causality and temporality. The annotation methods like CaTeRS (Mostafazadeh et al., 2016b), ESL (Caselli and Vossen, 2017) are included in this category.

The relations between these three manual annotation schemes can be seen in Figure 6.
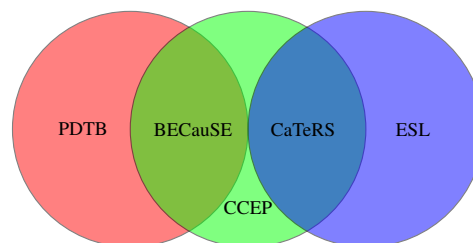
Figure 6: Relation of different manual annotation schemes.

F.3 Strengths and Weaknesses of Different Causality Acquisition Methods

As shown in Table 7, we analyze the strengths and weaknesses of the extractive methods, generative methods, and manual annotation from different aspects:

- Quality: Generative methods may give poor quality output, even generative LLMs are still suffering from hallucination problems. Extractive methods highly depend on the quality of the source data and are influenced by the extraction methods. However, humans have the capacity to perceive nuanced causal relationships and thus contribute high-quality commonsense causality.

- Collection Cost: Manual annotation is the most labor-intensive and costly due to human labor. The extractive methods can process a large amount of sources. The generative model, however, is a bit more costly than extractive methods due to the training cost of generative models, even invoking the API can become costly if the datasets are large.

- Collection Efficiency: It is self-evident that manual annotation is quite slow. Extractive methods are the most efficient while the generative methods are between the two regarding collection efficiency.

- Coverage: The scale of generative datasets can be very large due to the flexibility of generative methods. The scale of extractive datasets

is subject to the size of the source data. Due to the cost and efficiency concerns, the scale of manually annotated datasets is relatively small compared with extractive or generative methods.

- Adaptability: Generative methods are the least adaptive methods as they heavily rely on the domain of training datasets. Extractive methods are more adaptable but are limited by pre-defined patterns, which can vary across different domains. Manual annotations, however, are the most adaptable as humans can easily adapt to new domains and emerging commonsense knowledge.

- Scalability: It is obvious that the scalability of manual annotation is poor due to the cost and efficiency concerns while both generative and extractive are more scalable and are free from these concerns.

- Explainability: It is well-known that the generative methods lack interpretability and explainability due to the block-box characteristic of large models. Extractive methods are better as the matching patterns are explicit and defined by users. Manual annotation is the most explainable as humans can well explain the causal relationships they create.

G   Details About Commonsense Causality Benchmarks

We list the details of these benchmarks in Table 8 including the annotation unit, number of causation in the whole dataset, brief introduction, and the license for more responsible research. This section corresponds back to the benchmark introduction in § 2.

	Annotation Unit	#Overall	#Causal	C.F. [1]	Brief introduction	License
First-Principle Causality						
CauseEffectPairs (Mooij et al., 2016)	Variable	108	108	◻	108 different cause-effect pairs selected from 37 datasets which cover domains like meteorology, economy, medicine, engineering, biology. It focuses on the causal discovery problem whose goal is to decide whether X causes Y or Y causes X, given the co-existence of two variables X and Y.	FreeBSD
IHDP (Shalit et al., 2017)	Variable	2,000	2,000	◧	IHDP, the Infant Health and Development Program dataset, is about the effect of home visit on cognitive test scores for infants.	Custom Dataset Terms
CRAFT (Ates et al., 2022)	Video	58,000	-	◼	A new video question answering dataset that needs comprehension of physical forces and object interactions. CRAFT contains descriptive and counterfactual questions.	MIT
Commonsense Causality in Text Format						
Temporal-Causal (Bethard et al., 2008)	Clause	1,000	271	◻	A corpus of 1,000 event pairs for both temporal and causal relations.	Missing
CW (Ferguson and Sanford, 2008)	Clause	128	128	◼	CW, Counterfactual-World, is collected from existing psycholinguistic experiments.	Missing
SemEval07-T4 (Girju et al., 2007)	Phrase	220	114	◻	SemEval07-T4 is not specific for causal relations. It focuses on semantic analysis, i.e., automatic recognition of relations between pairs of words, of which causal relation exists.	Missing
SemEval10-T8 (Hendrickx et al., 2010)	Phrase	10,717	1,331	◻	Similar as the dataset in SemEval07-T4, it focuses on the automatic classification of semantic relations between pairs of nominals, which covers the cause-effect relations.	CC BY 3.0 Unported
COPA (Roemmele et al., 2011)	Sentence	2,000	1,000	◻	Each question consists of a premise and two plausible causes or effect, where the correct one is more plausible than the other.	BSD 2-Clause
EventCausality (Do et al., 2011)	Clause	583	583	◻	(Do et al., 2011) used the discourse connectives and the particular discourse relation to detect causality between events and built a causality corpus.	Missing

BioCause (Mihaila et al., 2013)	Clause	851	851	▭	BioCause contains 851 causal relations collected from 19 biomedical journal articles in the domain of infectious disease.	Creative Commons
CausalTimeBank (Mirza et al., 2014)	Sentence	318	318	▭	CausalTimeBank is the Timebank corpus with causal samples taken from TempEval-3 corpus.	CC BY-NC-SA 3.0
CaTeRs (Mostafazadeh et al., 2016b)	Sentence	2,502	308	▭	Annotation of a total of 1,600 sentences in the context of five-sentences storied selected from ROCStories corpus (Mostafazadeh et al., 2016a). CaTeRS is the name of the annotation scheme.	Missing
AltLex (Hidey and McKeown, 2016)	Clause	44,240	4,595	▭	AltLex is an open class of markers that contains causality.	Missing
BECauSE 2.0 (Dunietz et al., 2017)	Sentence	729	554	▭	BECauSE focus on annotations with causal relations and other relations that co-exists with causation.	MIT
ESL (Caselli and Vossen, 2017)	Sentence	2,608	2,608	▭	ESL is a corpus for causal and temporal relation detection.	CC BY 3.0 Unported
TCR (Ning et al., 2018)	Clause	172	172	▭	Similar as CaTeRS, TCR also focuses on temporal and causal relations. The difference here is that CaTeRs specifically focus on ROCStory dataset.	Missing
PDTB (Webber et al., 2019)	Clause	7,991	7,991	▭	PDTB is more general dataset that marks the discourse relations which include causation. The core idea of PDTB is that discourse relations are grounded in explicit words or phrases or the adjacency of two sentences.	LDC User Agreement for Non-Members
TimeTravel (Qin et al., 2019)	Sentence	109,964	29,849	◨	TimeTravel originates from the ROCStories (Mostafazadeh et al., 2016a). Each sample consists of the original story, a counterfactual fact, and a new storyline compatible with the counterfactual fact.	MIT
GLUCOSE	Clause	670K	670K	▭	Given a short story as the context and a sentence X in the story, GLUCOSE annotates 10 dimensions of causal explanation for X, which concerns events, location, possession, etc. These dimensions are usually implicit causes and effects of X.	Creative Commons Attribution NonCommercial 4.0 International Public License

XCOPA (Ponti et al., 2020)	Sentence	11,000	11,000	▭	The multilingual version (11 languages) of the COPA dataset.	CC BY 4.0
SemEval20-T5 (Yang et al., 2020)	Clause	25,501	25,501	▬	This is the dataset associated with SemEval-2020 Task 5. There are two subtasks. The first subtask is about determining whether a statement is a counterfactual or not while the second subtask is to extract the antecedent and consequent in a given counterfactual statement.	Missing
CausalBank (Li et al., 2020b)	Clause	314M	314M	▭	CausalBank consists of cause-effect statements collected from the Common Crawl corpus using causal lexical patterns.	
e-CARE (Du et al., 2022)	Sentence	21,324	21,324	▭	Each sample in e-CARE consists of a cause-effect pair and a conceptual explanation for the causation.	MIT
CoSIm (Kim et al., 2022)	Image + Text	3,500	3,500	▬	CoSIm is a multimodal counterfactual reasoning dataset. CoSIm focuses on commonsense reasoning for counterfactual scene imagination. Each sample in CoSIM consists of an image, an original question-answer pair, an imagined scene change in text format, and a question-answer pair under this imagined scene change. CoSIm stands as a touchstone for models' multimodal counterfactual ability.	MIT
CRASS (Frohberg and Binder, 2022)	Sentence	274	274	▬	It focuses on counterfactual reasoning setting that is in the format of question-answering.	Apache 2.0
COPES (Wang et al., 2023)	Sentence	1,360	1,360	▭	COPES is a corpus collected from ROCStories (Mostafazadeh et al., 2016a) for studying contextualized commonsense causal reasoning. Each sample consists of five chronologically ordered events. Annotators annotate whether the first four events contribute to the occurrence of the last event or not. COPES can study causal reasoning when contextual information is given.	MIT

Dataset	Type				Description	License
IfQA (Yu et al., 2023b)	Sentence	3,800	3,800	▬	IfQA is a counterfactual open-domain question-answering datasets. Each question is a counterfactual proposition with an *if* clause. A typical example is "What would be the time difference between Los Angeles and Paris if Los Angeles was on the east coast of the U.S.?"	Missing
CW-extended (Li et al., 2023)	Sentence	10,848	10,848	▬	CW-extended is a counterfactual dataset that is the augmentation of the original CW (Ferguson and Sanford, 2008) by words replacement. For instance, in the counterfactual statement that "If cats were vegetarians, families would feed them with cabbages." Replacement of *cats* with *dogs* would obtain a new counterfactual statement "If dogs were vegetarians, families would feed them with cabbages."	Missing
CausalQuest (Ceraolo et al., 2024)	Sentence	13,500	13,500	◧	CausalQuest is a dataset comprising natural causal questions collected from social networks, search engines, and AI assistants. It encompasses questions commonly asked by humans seeking to understand causation, categorized into queries about effects, causes, and causal relationships.	Apache 2.0
δ-CAUSAL (Cui et al., 2024)	Sentence	11,245	11,245	◧	δ-CAUSAL is a causal dataset studying the feasibility in causality. Each sample in δ-CAUSAL consists of a cause-effect pair associated with its supporting and opposing arguments. This dataset introduces the idea of uncertainty/defeasibility inherent in commonsense causality.	MIT
Commonsense Causality in Knowledge Graph Format						
CausalNet (Luo et al., 2016)	Word	11M	11M	▭	CausalNet consists of a vast amount of causal relationships from Bing web pages. Each causal relationship is a triple in the format of *(cause_word, effect_word, frequency)*.	Missing

ConceptNet (Speer et al., 2017)	Phrase	473,000	-	▭	ConceptNet is the knowledge graph version of the Open Mind Common Sense project. There are causal relations like *causes, entails, derived from and so on* inside ConceptNet relation set.	CC BY-SA 4.0
Event2Mind (Rashkin et al., 2018)	Phrase	25,000	-	▭	Each sample in Event2Mind consists of a given event, and its intent and reaction. For instance, given the event "PersonX reads PersonY's diary", X's intent is "to be nosey, know secrets", X's reaction is "guilty, curious", and Y's reaction is "angry, violated, betrayed". Some of the samples in Event2Mind contain causal relationships.	MIT
ATOMIC (Sap et al., 2019a)	Sentence	877K	-	◨	ATOMIC collect commonsense knowledge in the format of *if-then* relations. For instance, "if X pays Y a compliment, then Y will likely return the compliment". Obviously, ATOMIC contains event pairs that are in causal relations.	CC BY 4.0
ASER (Zhang et al., 2020)	Sentence	64M	494K	▭	ASER is an eventuality knowledge graph extracted from more than 11-billion-token unstructured textual data. It contains **a**ctivities, **s**tates, **e**vents, and their **r**elations.	MIT
CauseNet (Heindorf et al., 2020)	Word	11M	11M	▭	CauseNet consists of causal relations extracted from sem- and unstructured web sources(web pages and wikipedia) with precision of 83%.	CC BY 4.0
CEGraph (Li et al., 2020b)	Phrase	89.1M	89.1M	▭	A large lexical causal knowledge graphs (Cause Effect Graph) associated with CausalBank. Similar as CausalNet (Luo et al., 2016), the nodes in the graph are lemmatized terms, and a directed edge between two terms represents a causal relation. The weights of each edge is cooccurrence frequency.	Missing

Table 8: Overview of causal datasets.

H Language Models and Prompt Engineering for Large Language Models

We mention the important role of language models in qualitative causal reasoning. In this section, we cover preliminary knowledge of language models (App. H.1), large language models (App. H.2), and prompting techniques for LLMs (App. H.3). This section elaborates more details about § 4.1.

H.1 Language Models

Before the era of transformer-based pre-trained language models, there are various task-specific models such as CRF (Sutton and McCallum, 2012), HMM, LSTM (Hochreiter and Schmidhuber, 1997), GRU (Cho et al., 2014), SeqSeq (Sutskever et al., 2014), Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), CNN for text classification (Kim, 2014), etc. However, these models are unsatisfactory in performance and are mostly designed for specific tasks.

Starting from BERT (Devlin et al., 2019), transformer-based (Vaswani et al., 2017) pre-trained models have become the foundation models that have omnipotent capacities in language, vision, robotics, and reasoning. These foundation models handle surprisingly well data in the format of text, images, speech, structured data, 3D signals, etc (Bommasani et al., 2021). There are various kinds of taxonomy for transformer-based pre-trained language models. Here we classify these models by their model structure:

- Unidirectional Models: This series of language models is also known as decoder-based models. It includes the series of GPTs(GPT, GPT-2.0, GPT-3, GPT-3.5, GPT-4.0, GPT-Neo, etc), which is the model where ChatGPT builds upon, XLNet (Yang et al., 2019), LLaMA (Touvron et al., 2023), etc.

- Bidirectional Models: This series of language models is also known as encoder-based models. It includes the series of models derived from BERT (Devlin et al., 2019) such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021), lite BERT like ALBERT (Lan et al., 2020) and DistilBERT (Sanh et al., 2019), and domain-specific models like BioBERT (Lee et al., 2019), LegalBERT (Chalkidis et al., 2020),

language specific models like ERNIE (Sun et al., 2019), etc.

- Sequence-to-Sequence Models: This series of language models is also known as encoder-decoder-based models. Typical examples are T5 (Raffel et al., 2020), BART (Lewis et al., 2020),

H.2 Large Language Models

The relation between NLP models, language models, and large language models (LLMs) is shown in Figure 7. Multiple researches (Wei et al., 2022; Liang et al., 2022) have revealed the fact that "with the increase in model size, the reasoning ability of the model has undergone a change from quantitative to qualitative." During these two years, various LLMs have been proposed such as GPT-3.5, GPT-4.0, LaMDA (Thoppilan et al., 2022), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023).
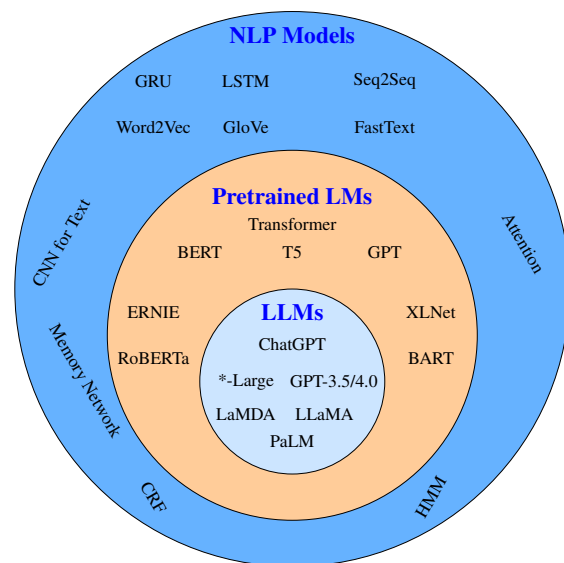
Figure 7: Relation of NLP models.

H.3 Prompting Techniques

As LLMs have become the foundation models of NLP tasks, how to fully utilize the power of these LLMs is a key point. Prompting is one of the key techniques. A prompt usually contains one or more of the following elements:

- Instruction: The description of what you expect the LLM to perform.

- Context: External or contextual information that provides more detailed guidance besides instruction.

- Input Data: The input question we want to have an answer to.

- Output Expectation: The format and requirement of the expected output.

Useful techniques for prompting include zero-shot prompting, few-shot prompting (Brown et al., 2020; Kaplan et al., 2020), chain-of-thought prompting (Wei et al., 2022; Zhang et al., 2023), automatic or active prompting (Zhou et al., 2023; Diao et al., 2023). Here we mainly elaborate on the prompting process of chain-of-thought prompting. The following example shows the key idea of chain-of-thought prompting [7]. Standard few-shot prompting is like:

> Q: A farmer collects 56 eggs from his hens. If he packs them in trays that hold 6 eggs each, how many trays does he need?
> A: The answer is 10 trays.
> Q: A classroom has 24 students. If 3 students were absent today, and 2 new students joined the class, how many students are in the class now?

The model gives an incorrect answer.

> A: The answer is 25 students.

However, chain-of-thought prompting is a series of intermediate reasoning steps:

> Q: A farmer collects 56 eggs from his hens. If he packs them in trays that hold 6 eggs each, how many trays does he need?
> A: There are 56 eggs. Each tray holds 6 eggs. To find out how many trays, divide 56 by 6, which gives 9 remainder 2. So he needs 9 full trays and 1 more tray for the remainder, making 10 trays.
> Q: A classroom has 24 students. If 3 students were absent today, and 2 new students joined the class, how many students are in the class now?

With the techniques of chain-of-thought prompting, the LLM's output is

---

[7] The example follow the same idea of the example provided in the original paper of CoT (Wei et al., 2022)

> A: The classroom originally had 24 students. With 3 students absent today, there were only 21 students present. However, 2 new students joined, bringing the total back to 23 students in class today. The total number of students enrolled now is 24 (original) - 3 (absent) + 2 (new) = 23 students. The answer is 23 students.

This simple technique has seen great success in complex reasoning tasks such as arithmetic reasoning, commonsense reasoning, symbolic reasoning, and so on. Due to the page limit, we omit the other topics such as alignment techniques for Large Language Models.

I Causality in Depth: From Concepts to Theories

When we mention uncertainty in § 2.2 and contextual nuance in § 5, it involves the understanding of causal sufficiency and necessity (Nadathur and Lauer, 2020), which we detail in App. I.1. Then, we talk about different theories about causation, which are related to the quantitative measurement of causal strength in § 4.2 and annotation schemes in § 3.3.

I.1 Causal Sufficiency and Necessity

Causal Sufficiency. Causal sufficiency refers to a situation where a cause is sufficient enough to lead to a particular effect. Namely, if a cause is present, it is capable of bringing about the effect on its own, without the need for any additional causes or factors.

Causal Necessity. Causal necessity refers to the idea that a cause is necessary for the occurrence of a specific effect. Namely, without the cause, the effect would not happen.

I.2 Three Levels of Causality Based on Causal Sufficiency and Necessity

From the perspectives of causal sufficiency and necessity, there are three levels of causality (Volchok, 2015).

Absolute Causality. Absolute causality refers to the fact that the cause is necessary and sufficient to the happening of the effect. Typically, this kind of causality is only found in the physics domain, such as "Water flows from high places to low places on Earth". However, causality in other domains
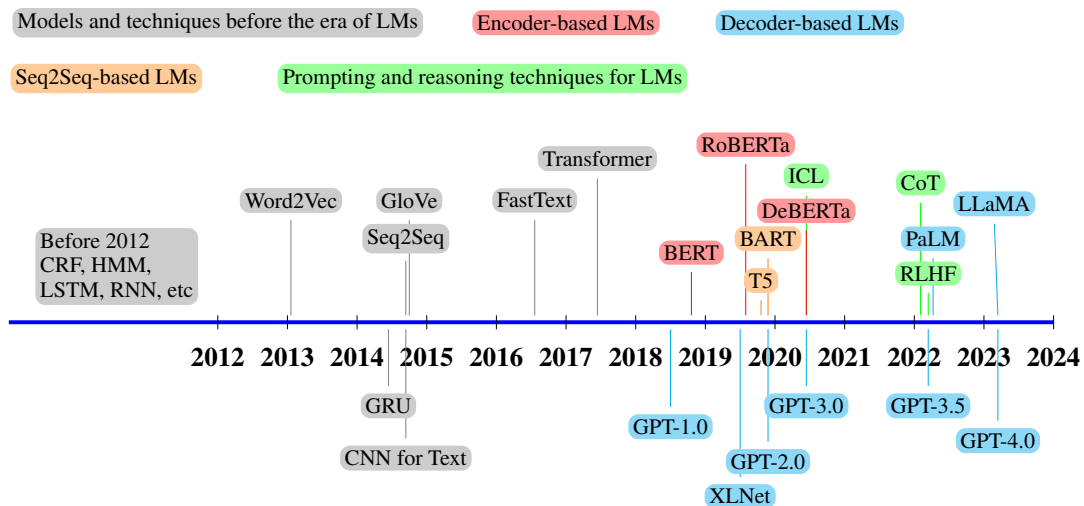
Figure 8: Timeline of major NLP models and techniques for reasoning. We take the publication(or arxiv) time of these works as the date of their appearance(best viewed in color).

like law, economics, ecology, social science, and so on is not absolute. For example, "the increase of carbon dioxide emission" leads to "rise of sea level". Most causality falls in the categories of *conditional causality* and *contributory causality*.

Conditional Causality. Conditional causality suggests that a cause is necessary, but not sufficient to trigger an effect. For instance, receiving a vaccine is a necessary but not sufficient cause for preventing a particular disease.

Contributory Causality. Contributory causality means that a cause is not the primary or sole cause of the effect. However, combined with other causes, they collaboratively lead to the occurrence of the effect. For instance, the presence of "a pedestrian jaywalks at a red light" and "a high-speed moving car" contributes to the effect that "the car hit the pedestrian", each cause on its own wouldn't lead to the occurrence of the accident.

I.3 Theories of Causation

Generally speaking, there are three theories related to causation:

- Counterfactual Theory (Lewis, 1973; Menzies and Beebee, 2001): The basic idea of the counterfactual theory of causation is that the causation between the cause and the effect can be supported by the argument that "Had *cause* not occurred, *effect* would not have occurred". The counterfactual theory is a revision of the regularity theory of causation (Andreas and Guenther, 2021).

- Probabilistic Theory (Skyrms, 1981; Eells,

1991; Hitchcock, 1997): Probabilistic causation theories assume that causes change the probabilities of their effects. They use the tools of probability theory to characterize the relationships between causes and effects.

- Force Dynamics Model (Talmy, 1988): This theory decomposes the causation into various types of force formulations including the imparting of force, resistance to force, overcoming resistance, and removal of a force. It analyzes causing into finer primitives including "letting", "hindering", "helping", and "intending" (Wolff and Thorstad, 2017). This force dynamic approach accounts for cognitive causation phenomena including causation in language, causal chains, cognitive causation, etc. It is the theoretical background for CEP annotation schemes (App. F.2).

These theories provide the theory ground for annotations, acquisition, and reasoning of causality.

J Linguistic Causality

The acquisition methods for causality in § 4 need linguistic expertise.For instance, the linguistic pattern matching methods need linguistic expertise in causal connectives and verbs, which we elaborate on in App. J.1 and J.2 respectively.

J.1 Explicit Causal Connectives

Penn Discourse Treebank (PDTB) (Webber et al., 2019) focuses on the annotation of discourse cues and does not postulate any structure con-

straints on discourse relations. There are 30 explicit causal connectives under the label of *Contingency.Cause.Result* in PDTB. The explicit causal connectives include *accordingly, and, as, as a consequence, as a result, as it turns out, as such, because of that, but, consequently, finally, for that reason, furthermore, hence, in fact, in other words, in response, in short, in the end, indeed, so, so as, so that, that is, then, therefore, thus, thus being, to this end, ultimately*.

J.2 Causation Verbs

We summarize the most common verbs that indicate causation (Martin, 2018) in Table 9. We classify these causal verbs into different types based on their causal roles including

- Direct: It means the subject directly leads to or causes the occurrence of the object. Examples are "cause", "lead to", etc.

- Preventative: It means that the subject prevents or stops the happening of the object. Typical examples include verbs like "prevent", "stop", and "prohibit".

- Facilitative: This indicates that the subject facilitates or contributes to the occurrence of the effect. Examples are "enable", "faciliatate", etc.

- Consequential: These verbs emphasize on the consequences of the subject's action brings to the object. Examples includes "drive", "force", etc.

- Influential: These influential verbs imply that the subjects influence the object. Examples include "influence", "swing", "sway", etc.

K Preliminary of Causal Inference

As mentioned in § 4.1, causal inference is a useful tool for neuro-symbolic approaches. In this section, we give a brief preliminary of causal inference, including key concepts (App. K.1), tools (App. K.2), and structural causal causal models (SCM, App. K.3).

K.1 Concepts

In this subsection, we introduce some fundamental concepts regarding causal inference, which include the distinction between causation and correlation, complex causality structures like confounder and

Type	Examples
Direct	Make: Hard work makes success. Cause: The drugs tend to cause slight drowsiness. Lead to: A poor diet will lead to severe illness.
Preventative	Prevent: The vaccine prevents the disease. Stop: The medicine stops the pain. Prohibit: The law prohibits the liquor trade.
Facilitative	Enable: The software enables the team to process the data three times faster, speeding up the project's progress. Facilitate: The peace talks facilitated a ceasefire, reducing the hostilities between these two countries that have hated each other for a long time. Assist: The teachers' detailed slides assist the students in grasping mathematical concepts.
Resultative	Drive: Ambition drives people to succeed in their careers. Force: The strong winds forced the tree to topple. Compell: Peer pressure compels one to follow group norms.
Influential	Influence: His family background influences his career choice. Swing: The judge's final decision swung the verdict in favor of the defendant. Sway: The compelling argument presented by the defense swayed the jury.

Table 9: Taxonomy of causation verbs and examples.

collider, and causal modeling concepts such as D-separation, average treatment effect, etc. These concepts are crucial for understanding the intricate nature of cause-and-effect relationships and for developing a robust framework for causal analysis.

Correlation and Causation. We may all heard the maxim that "correlation does not imply causation" (Rohrer, 2018). We present several examples to better illustrate this concept:

- Ice Cream Sales and Drowning Incidents: There is a positive correlation between ice

cream sales and the frequency of drowning incidents. These two kinds of events share the confounding factor *hot weather*. Namely, hot weather drives more people to buy more ice cream and swim, which leads to more drowning incidents.

- Stork Population Around Cities and the Deliveries Outside City Hospitals (Höfer et al., 2004): There is a positive correlation between the number of storks in the German state of Lower Saxony between 1970 and 1985 and the deliveries in that area. Both human birth rates and the population of storks are influenced by the availability of suitable nesting areas in rural areas.

- Margarine Consumption and Divorce Rates: The correlation between the per capita consumption of margarine and the diverse rate in Maine State is spurious. This is more coincidence rather than correlation, as shown in (Vigen, 2015).

From these examples, we can tell that correlation suggests a potential causal relationship but does not establish a deterministic one. To determine a causal relationship, spurious correlation, controlled experiments, and confounding variables should be considered.

Confounder. Confounders or confounding variables are external factors affecting both the cause and the effect, which is one of the major reasons for false causal inferences between an independent variable and a dependent variable. The following are several examples of confounding variables in causal inference:

- The causal relationship between vaccination status and infection rate may be overridden by a confounder of health awareness. Individuals who are more mentally aware of health are more likely to get vaccinated and take other activities to reduce infection rates.

- The causal relationship between college degree attainment and lifetime earnings is spurious due to the confounder of family background. The background of a wealthy family increases both the likelihood of college degree attainment and high earnings due to networking and social capital.

- The correlation between social media usage and depression levels may be confounded by the factor of age. Younger people tend to use social media more and are vulnerable to depression.

Collider. By contrast, colliders are variables causal by at least two other variables. Namely, a collider has more than one arrow pointing to it in a causal graph. What follows are several collider examples:

- Hypertension usually is the collider node of genetic predisposition and a high-salt diet.

- The admission to a prestigious university is the collider node of parental socioeconomic background and the quality of primary education.

- The state of one's health is the collider node of genetic factors and postnatal factors such as exercise and nutrition.

D-Separation. D-separation (Geiger et al., 1989; Hayduk et al., 2003) is a criterion used in a directed acyclic graph (DAG) to determine whether a set of variables X is independent of another set of variables Y, given a third set of variables Z. Two (sets of) nodes X and Y are d-separated by a set of nodes Z if all of the paths between (any node in) X and (any node in) Y are blocked by Z.

Treatment Group and Control Group. In a causal study, we usually divide individuals into two groups: the treatment group and the control group. The treatment group receives the intervention or treatment while the control group does not.

Average Treatment Effect. In a causal study, we choose an outcome variable influenced by the treatment. The outcome is measured in both the treatment and control groups. The Average Treatment Effect (ATE) is defined as the average difference between the outcomes of the control group and the control group. Mathematically,

$$\text{ATE} = \mathbb{E}[Y_1 - Y_0] \tag{6}$$

where the Y_1 is the outcome of the treatment group while Y_0 is the outcome of the control group. The expectation is taken over all individuals in the groups.

K.2 Tools for Causal Inference Analysis

Directed Acyclic Graphs (DAGs). DAGs provide a structured and visual representation of causal relationships between variables. Nodes in DAGs represent the variables of interest and a directed edge from A to B implies that A causally impacts B. The acyclicity ensures that there are no loops in causality. With DAGs, it is easier to detect special nodes like common ancestral nodes and collider nodes in a complex causal system. Besides, DAGs also help to identify confounding variables and analyze counterfactuals and interventions.

***Do*-Calculus.** The do-calculus (Pearl, 2000, 2009, 2012; Tucci, 2013) is an axiomatic system for replacing probability formulas containing the do operator with ordinary conditional probabilities. *Do*-calculus is to estimate the causal effect. Namely,

$$P(Y|do(X = x)). \quad (7)$$

It returns the probability of Y if we intervene X as x.

K.3 Structural Causal Models

Structural Equation. Before we start, we distinguish two different kinds of variables within a causal system. Endogenous variables are influenced by other variables within the same system. They are the variables of interest in a structural causal model. Exogenous variables, however, are not influenced by other variables within the same system. Namely, there is no causal parent for the exogenous variables in the system. In a DAG, exogenous variables are nodes without incoming edges while endogenous variables are nodes with incoming edges.

Let us suppose that X causally influences Y and X is an exogenous variable. The structural equation is

$$X = U_X \quad (8)$$

where U_X is the unobserved factors affecting X. For endogenous variable Y (it is influenced by other variables within the same system), the structural equation becomes

$$Y := f(X) + U_Y \quad (9)$$

where $f(X)$ is the causal function between X and Y. U_Y represents the unobserved factors Y.

Structural Causal Models. Structural Causal Models (SCMs)(Bollen and Pearl, 2013; Hair Jr et al., 2021) are mathematical tools to analyze
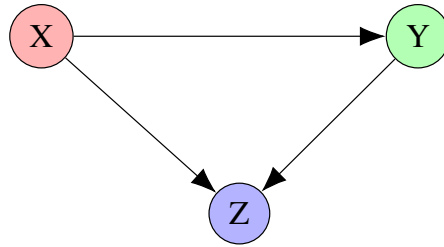
Figure 9: Timeline of major NLP models and techniques for reasoning. We take the publication(or arxiv) time of these works as the date of their appearance(best viewed in color).

DAGs. A SCM consists of a set of exogenous variables (U) and a set of endogenous variables (V) connected by a set of functions (F). Each SCM is associated with a DAG where a node is a variable in U or V while an edge is a function f in F. We show an example of a DAG in Figure 9.

Now structural equations turn into:

$$\begin{cases} Y = f_Y(X) + U_Y \\ Z = f_Z(X, Y) + U_Z \end{cases} \quad (10)$$

where f_Y are causal function between X and Y while U_Y is the unobserved factor affecting Y. f_Z are causal function between X and Y as a unit and Z while U_Z is the unobserved factor affecting Z.

L Handbook for Beginners to Learn Commonsense Causality

We believe that a detailed and accessible handbook is invaluable for junior researchers to get familiar with research topics. We list step-by-step and comprehensive practices for junior researchers in the domain and hope that it is helpful for this field's future progress.

- First, we kindly suggest researchers interested in this domain start with the background and preliminary knowledge which covers linguistic causality (App. J), NLP preliminary including pretrained language models and prompting techniques (App. H), benchmarks of commonsense causality (§ 2 and App. G), the taxonomy of causation (§ 2 and App. D), theories of causation (App. I), probabilistic causation (§ 4.2), reading of related surveys (App. C), and causal inference preliminary (App. K).

- Then, there are two branches for selection: causality acquisition (§ 3 and App. F) and

reasoning over causality (§ 4). For causality acquisition, there are various approaches such as extractive methods (§ 3.1), generative methods (§ 3.2), and manual annotations (§ 3.3). For causality reasoning, both qualitative reasoning (causal reasoning and counterfactual reasoning) and quantitative measurement of causal strength (§ 4.2, App. E) are promising topics for further research.

- Following this, there are some more challenging but significant prospective topics for exploring. We highlight the contextual nuances, complex structures, temporal dynamics, probabilistic approaches, and multimodal data of commonsense causality for future study. Specifically, we first underscore the importance of developing models that address the contextual dependency of cause-effect relationships and intricate causal structures. Additionally, we underline the temporal dynamics of commonsense causality, pointing out the urgency of identifying the most effective timing for interventions and understanding the temporal sequences of causal impacts for future study. Furthermore, we emphasize the necessity of employing probabilistic perspectives to manage uncertainties in commonsense causality. Finally, we outline several challenging topics for multimodal causality analysis including the reasoning of multimodal commonsense causality and the alignment of cross-modal commonsense causality. For a comprehensive discussion of these promising future directions, please refer to Section 5.

- Finally, for researchers who are more interested in the application side of commonsense causality (App. A), the combination of commonsense causality and fields like medical diagnosis (Richens et al., 2020), psychology (Matute et al., 2015; Eronen, 2020), behavioral science (Grunbaum, 1952), economics (Bronfenbrenner, 1981; Hoover, 2006), legal systems (Williams, 1961; Summers, 2018) are optimistic topics.

More details can be found in Figure 10. Through this detailed handbook, we hope that the topic of commonsense causality is more accessible to researchers who are interested in this domain.

Step I: Background and Preliminary

Linguistic Causality (App. J)
- Familiarity with linguistic causal lexical patterns.
- Causal connectives and verbs.

NLP Preliminary (App. H)
- Concepts and definition of commonsense in the NLP community (App. B.1).
- Pretrained language models like BERT (Devlin et al., 2019), BART (Lewis et al., 2020), GPT (Brown et al., 2020), ChatGPT, etc.
- In-context learning and prompt engineering techniques like Chain-of-Thought (Wei et al., 2022), Tree-of-Thought (Yao et al., 2023), etc.

Taxonomy of Causation (§ 2, App. D, and App. I)
- Causal sufficiency and causal necessity (Nadathur and Lauer, 2020).
- Absolute causality, conditional causality, and contributory causality.
- Counterfactual causality.

Probabilisitc Theory of Causation
- Basic probability and Simpson's paradox.
- Probability-raising theories of causation (Menzies, 1989; Hitchcock, 1997; Pearl, 2000).
- Related theories including counterfactual theory (Lewis, 1973), dynamic force model (Talmy, 1988).

Preliminary of Causal Inference (App. K)
- Concepts related to causal inference, including correlation, confounder, collider, d-separation, treatment/control group, and average treatment effect (App. K.1).
- Tools for causal inference analysis, including DAG and Do-calculus (App. K.2).
- Structural causal models (App. K.3).

Benchmarks (§ 2 and App. G)
- Benchmarks for first-principle causality, commonsense causality, and causality in graph format (Table 1 and 8).
- Annotation schemes of causality (App. F.2).

Survey reading (App. C)
- Surveys about commonsense reasoning, causal knowledge acquisition, causal inference, probabilistic causation, etc.
- Surveys about preliminary knowledge.

Others

Step II: Causality Acquisition (§ 3)

Extraction Methods (§ 3.1)
- Benchmarks and datasets for causality extraction (§ 3.1).
- Linguistic pattern matching and learning-based approaches (§ 3.1).
- Implicit and inter-sentential causality extraction (App. F).

Generative Methods (§ 3.2)
- Language model-based approach.
- From causal graph to narrative causality.

Manual Annotation (§ 3.3)
- Theories of causation (App. I.3).
- Manual annotation schemes including schemes based on the template with triggers (BECauSE (Dunietz, 2018) and PDTB (Webber et al., 2019)), schemes based on the CEP scheme (CaTeRS (Mostafazadeh et al., 2016b), and BECauSE (Dunietz, 2018)), and schemes that annotating temporality jointly (CaTeRS (Mostafazadeh et al., 2016b) and ESL (Caselli and Vossen, 2017)).

Others

Step III: Reasoning Over Causality (§ 4)

Qualitative Causal Reasoning (§ 4.1)
- Benchmarks for causal reasoning reasoning (Table 1 and 8).
- Scaling NLP models as causal knowledge bases (Figure 8 and App. H).
- Neuro-symbolic methods based on principles of causal inference, temporal constraint, and logic rules.

Counterfactual Reasoning
- Benchmarks for counterfactual commonsense causality (Table 1 and 8).
- Reasoning methods for counterfactual commonsense causality.

Quantitative Causal Reasoning (§ 4.2)
- Factual and causal uncertainty and probabilistic causality (§ 2.2 and App. E).
- Causal strength metrics based on word co-occurrence.
- Probabilistic causal strength metrics based on event probabability (Good, 1961; Suppes, 1973; Eells, 1991; Pearl, 2009).

Others

Step V: Applications (App. A)

Healthcare and Medical Assistance:
- Medical diagnosis.
- Disease treatment and prevention program.
- Public health strategy.

Legal and Forensic Analysis
- Determining legal liability.
- Investigation of criminal intent and motive.

Others: medical diagnosis (Richens et al., 2020), psychology (Matute et al., 2015; Eronen, 2020), behavioral science (Grunbaum, 1952), economics (Bronfenbrenner, 1981; Hoover, 2006), legal systems (Williams, 1961; Summers, 2018).

Step IV: Future Directions (§ 5)

Contextual Commonsense Causality
- Identification of a complete set of causality contexts and estimation of the distribution of contexts in contextual unanimity causality (Dupré, 1984).
- Partial contextual models that balance between the model's complexity and its explanatory power.

Multimodal Commonsense Causality
- Acquisition, reasoning, and measurement of commonsense causality beyond text.
- Alignment of same cause-effect pairs in different modalities.

Complex Commonsense Causality
- Understanding of rich causal structures like confounder, collider, causal chain, and cycle (App. K.1).
- Development of complex structure commonsense causality benchmarks.
- Theoretical frameworks for complex structures analysis.

Probabilistic View of Causality Strength
- Probabilistic graphical models such as Bayesian Networks (Heckerman, 2008), Markov Random Fields, etc.
- Dynamic probabilistic causal models with temporality that estimate $P(E_t|C_{t-\delta})$.

Role of Time in Commonsense Causality
- Determining optimal timing for intervention to prevent negative outcomes.
- Investigating temporal patterns of causal effects to decide how the effects of causes vary over time.

Others

Figure 10: Handbook for beginner researchers interested in commonsense causality.