A Multi-Level Benchmark for Causal Language Understanding in Social Media Discourse

Xiaohan Ding **Kaike Ping** **Buse Çarık** **Eugenia Rho**
Department of Computer Science, Virginia Tech, USA
{xiaohan,kaike,buse,eugenia}@vt.edu

Abstract

Understanding causal language in informal discourse is a core yet underexplored challenge in NLP. Existing datasets largely focus on explicit causality in structured text, providing limited support for detecting implicit causal expressions, particularly those found in informal, user-generated social media posts. We introduce **CausalTalk**, a multi-level dataset of five years of Reddit posts (2020–2024) discussing public health related to the COVID-19 pandemic, among which $10,120$ posts are annotated across four causal tasks: (1) binary causal classification, (2) explicit vs. implicit causality, (3) cause–effect span extraction, and (4) causal gist generation. Annotations comprise both gold-standard labels created by domain experts and silver-standard labels generated by GPT-4o and verified by human annotators. CausalTalk bridges fine-grained causal detection and gist-based reasoning over informal text. It enables benchmarking across both discriminative and generative models, and provides a rich resource for studying causal reasoning in social media contexts[1].

1 Introduction

Causal language understanding is a foundational yet underexplored challenge in natural language processing (NLP) (Niess et al., 2025; Yu et al., 2025; Kim et al., 2023; Ali et al., 2021; Blanco et al., 2008; Heddaya et al., 2024). Accurately identifying cause-effect relations, that is, how one event leads to another, is essential for tasks such as information extraction (Dukić et al., 2023; Hu et al., 2025), narrative understanding (Sun et al., 2023; Zhang et al., 2021), and decision support (Ding et al., 2024). While prior work has primarily focused on structured or formal texts (Caselli and Vossen, 2017; Kyriakakis et al., 2019), causal reasoning in informal or unstructured social media text remains underdeveloped, despite its growing relevance for analyzing discourse around public health (Son et al., 2018), misinformation (Adams et al., 2023), and everyday decision-making (Lin and Wu, 2008; Ding et al., 2024).

Existing causal datasets (Table 1) have made progress by annotating explicit causal links across newswire (Mirza et al., 2014a), biomedical (Mihăilă et al., 2013), and narrative texts (Mostafazadeh et al., 2016). However, they remain limited in several ways. (1) Most datasets emphasize explicit causality (Hendrickx et al., 2009; Mirza et al., 2014a), offering limited coverage of implicit relations that frequently appear in informal discourse (Hartshorne, 2014). (2) Despite growing interest in long-context reasoning, existing benchmarks lack gist-driven causal annotations that reflect how humans distill causal meaning into concise representations (Reyna, 2012; Brainerd and Reyna, 1990).

To address these gaps, we draw on fuzzy-trace theory (FTT) (Reyna, 2012; Brainerd and Reyna, 1990), which argues that texts with coherent causal relationships help people form gist-like mental representations—concise abstractions that support better understanding and memory than verbatim information. Recent work in LLM prompting has leveraged this insight to support long-context processing (Lee et al., 2024; Dhaini et al., 2024; Liu et al., 2024c), but such approaches lack suitable training or evaluation data grounded in informal, real-world discourse. Inspired by these developments, we propose a new dataset to support both causal detection and gist-based generation.

We introduce **CausalTalk**, a multi-level dataset of Reddit posts from 2020 to 2024 that capture public health discussions, with a particular focus on the COVID-19 pandemic, annotated for causal language across four tasks: (1) binary causal classification, (2) explicit vs. implicit causality detection, (3) cause–effect span extraction, and (4) causal

[1]The dataset and code are available at https://github.com/xding2/CausalTalk

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28776–28790
November 4-9, 2025

Dataset	Language	# Entries	# Causality Relations	Causality Type
SemEval-2010 Task 8 (Hendrickx et al., 2009)	English	10,674	1,325	Explicit
CausalTimeBank (Mirza et al., 2014a)	English	11,000	318	Explicit
EventStoryLine (Caselli and Vossen, 2017)	English	7,275	5,519	Explicit, Implicit
BioCausal (Mihăilă et al., 2013)	English	13,342	7,562	Explicit
BECauSE 2.0 (Dunietz et al., 2017)	English	1,803	1,634	Explicit
RED (O'Gorman et al., 2016)	English	8,731	4,969	Explicit
CaTeRS (Mostafazadeh et al., 2016)	English	2,708	488	Explicit, Implicit
MAVEN-ERE (Wang et al., 2022)	English	103,193	57,992	Explicit
HPG Incidents (Inoue et al., 2023)	Japanese	18,171	970	Explicit

Table 1: Overview of major causal relation datasets used in NLP.

gist generation. The dataset includes both *gold-standard* annotations—manually labeled and adjudicated by expert annotators—and *silver-standard* annotations—generated by GPT-4o using zero-shot prompting and subsequently verified by human annotators, following established practices from prior annotation research (Hengle et al., 2024; Li et al., 2024; Mirzakhmedova et al., 2024). In total, **CausalTalk** provides 10 million raw Reddit posts and 10,120 annotated instances across 43 public health related subreddits.

2 Related Work

2.1 Causal Datasets in NLP

Causal language understanding has long been a key challenge in information extraction. To support model development for causal relation identification, a range of datasets have been proposed across various domains and annotation schemes. A summary of these datasets is presented in Table 1.

A foundational dataset often used for causal classification is the **SemEval-2010 Task 8** dataset (Hendrickx et al., 2009), which contains 10,674 sentences annotated with semantic relations between pairs of nominals, including 1,325 instances labeled as Cause–Effect. Although the dataset was originally developed for semantic relation classification, (Kyriakakis et al., 2019) later adapted it for binary causal classification by labeling data with the Cause–Effect relation as causal and all others as non-causal.

The **CausalTimeBank** corpus (Mirza et al., 2014a) and the **EventStoryLine** dataset (Caselli and Vossen, 2017) introduce event-level causal annotations by marking causal signals (C-SIGNAL) and links (CLINK), or PLOT LINKs (CAUSES, CAUSED BY). Following (Li and Mao, 2019), (Kyriakakis et al., 2019) filtered these datasets to include only intra-sentence causal pairs, resulting in 318 and 5,519 causal instances respectively.

Other efforts have expanded the domain and annotation scope. For example, the **BioCausal** dataset provides biomedical causal annotations over 13,342 sentences from PubMed (7,562 causal), with a publicly available subset of 2,000 examples (Mihăilă et al., 2013).

To improve annotation granularity, the **BECauSE 2.0** corpus (Dunietz et al., 2017) labels 1,803 instances of causal language across 5,380 sentences, with over 90% of them including both cause and effect spans. The **Richer Event Description (RED)** corpus (O'Gorman et al., 2016) goes further to annotate 4,969 relations across 8,731 events in 95 documents, integrating causal, temporal, and coreference relations.

To explore both explicit and implicit causality, **CaTeRS** (Mostafazadeh et al., 2016) annotates 1,600 sentences from ROCStories, marking 488 causal and temporal relations across 2,708 events. Despite its limited size, it emphasizes narrative causality and implicit links, complementing the more structural datasets.

More recently, efforts have shifted toward large-scale and multilingual resources. **MAVEN-ERE** (Wang et al., 2022) is a unified event relation corpus with over 57,000 human-annotated causal links, 1.2M temporal relations, and 103,000 event coreference chains, offering substantial coverage and enabling joint modeling of event relations. Meanwhile, the **HPG Incidents** dataset (Inoue et al., 2023) focuses on industrial safety, with 970 annotated high-pressure gas incident reports designed for causal and named entity recognition tasks. Finally, the **MECI** corpus (Lai et al., 2022) advances event causality identification (ECI) research by offering consistent annotations across five typologically diverse languages (English, Danish, Spanish, Turkish, Urdu). With over 11,000 annotated relations, MECI supports multilingual and crosslingual causality modeling.

2.2 Causal Gist and Fuzzy-Trace Theory in NLP and LLMs

Fuzzy-trace theory (FTT) posits that people rely on simplified, essential meanings (called *gists*), rather than exact details when processing information (Brainerd and Reyna, 1990). This gist-based reasoning helps explain how individuals make decisions, assess risks (Ding et al., 2024), and comprehend long texts (Lee et al., 2024; Reyna, 2012). Research shows that gist representations are more memorable than verbatim details (Reyna, 2012), supporting the idea that cognitive development involves a shift from literal recall to abstract understanding (Reyna, 2021).

In recent years, FTT has increasingly informed developments in natural language processing and large language models (Ding et al., 2024; Lee et al., 2024; Dhaini et al., 2024; Liu et al., 2024c). Studies show that summarizing long texts using gist-level representations, combined with attention to local details, improves the efficiency of long-context reasoning (Lee et al., 2024; Reyna, 2012). For example, Lee et al. proposed *ReadAgent*, a prompting-based LLM system designed to emulate human reading strategies. The system segments long texts into smaller episodes, generates gists for each segment, and maintains contextual links to support reasoning across extended content. A similar line of work, Liu et al. 2024c uses summarization models to identify gists and integrates them into downstream models to enhance long-text understanding.

Extending these approaches, Ding et al. 2024 proposed a framework for applying causal gist generation to chain-of-thought prompting in LLMs. Their method, *Role-Based Incremental Coaching* (RBIC), breaks down complex causal reasoning—especially within fragmented or informal online discourse—into more manageable subtasks, each designed to focus on a specific causal step and summarized by a causal gist. Ding et al.'s approach treats gist as a central step in structuring causal reasoning, allowing models to interpret nuanced online discourse.

Overall, past research highlights the importance of aligning causal language modeling with human cognitive strategies, especially gist-based understanding. Motivated by these insights, our dataset is designed to support multi-level causal understanding in social media discourse, ranging from detection to the generation of causal gist.

3 Dataset Collection

3.1 Data Sources

We collected Reddit posts (including submissions and their associated comments) related to public health, with a particular focus on the COVID-19 pandemic, from January 2020 through December 2024 across 43 subreddits. The primary data sources include pandemic-focused subreddits such as `r/Coronavirus`, `r/COVID19`, `r/wuhanflu`, and `r/DebateVaccines`. These 43 subreddits were selected due to their coverage of pandemic discussions from scientific, policy-related, and personal experience perspectives. Detailed descriptions of these subreddits are provided in the Appendix A.

We used the Pushshift API to retrieve data and filter it to ensure data quality. Specifically, duplicates, extremely short posts (fewer than 20 tokens), and non-English content were excluded. Finally, we anonymise the data user id to protect user privacy.

Feature	Count
Total Submissions	239,222
Total Comments	19,138,266
Number of Subreddits	43

Table 2: Summary Statistics of Reddit COVID-19 Pandemic Dataset

3.2 Dataset Statistics

The final dataset comprises 239,222 Reddit submissions and 19,138,266 comments collected over the five-year span. Table 2 summarizes the dataset statistics.

4 Annotation Schema

To capture the diverse manifestations of causal language in social media text, we designed a hierarchical annotation schema consisting of four complementary tasks: (1) binary annotation of causal language, (2) explicit versus implicit causality classification, (3) fine-grained span annotation of causes, effects, and signals, and (4) casual gist generation. Each task incrementally builds upon the previous, facilitating both analysis of causal structure and the creation of high-quality, multi-level benchmark data to support downstream NLP modeling.

Task 1: Binary Annotation of Causal Language. Our annotators received initial training based on Sanders and Sweetser 2009's definition of causal relationships in language, as well as the *Annotation of Causal Relations* schema (Rehbein and

Ruppenhofer, 2017). Their research characterizes causal content as textual expressions indicating a relationship where one event, action, or condition directly leads to or influences another (Rehbein and Ruppenhofer, 2017; Sanders and Sweetser, 2009).

During the annotation process, each data entry was independently annotated by five individuals as either containing causal content (1) or not (0). Discrepancies were resolved through group discussions facilitated by a sixth annotator, with the final ground truth label determined by the majority vote reached during these discussions.

Task 2: Explicit vs. Implicit Causality. Our annotators were trained based on prior research to distinguish between *Explicit* and *Implicit* causality in language (Sanders and Spooren, 2009; Pickering and Majid, 2007). The definitions and references used during training are as follows:

- **Explicit causality**: Causality that is overtly expressed through causal markers such as *"because," "therefore," "as a result,"* etc. These connective cues clearly indicate a causal relationship between two events or propositions (Sanders and Spooren, 2009).
- **Implicit causality**: Causality that is not explicitly marked by connectors but is instead inferred from context, verb semantics, or the structure of events (Pickering and Majid, 2007).

Each data entry that was labeled as containing causal content in Task 1 was further examined in Task 2. Specifically, these entries were annotated by five independent annotators as expressing either **explicit causality** (1) or **implicit causality** (0). As with the previous task, disagreements were resolved through group discussion led by a sixth annotator.

Task 3: Cause–Effect Span Annotation. In this task, annotators identified specific spans of text corresponding to the cause, effect, and the causal signal (when applicable) within each data entry labeled as containing causal content in Task 1 ($n = 600$). The goal of this span-level annotation was to capture the linguistic realization of causal relationships within individual sentences.

For instances of *explicit causality*, annotators typically selected spans surrounding overt causal markers (e.g., "because," "as a result"). In contrast, for *implicit causality*, annotators relied on contextual cues, event structure, and verb semantics to infer and label the appropriate cause and effect spans, even when no explicit connective was present (Rehbein and Ruppenhofer, 2017). Below are examples of annotated sentences demonstrating both explicit and implicit causal structures:

Explicit causal signal:
If <cause>someone sneezes in your face </cause> it can cause <effect> respiratory droplets that expose you to their virus-laden viruses </effect>

Implicit causality (no signal):
<cause>Long-term isolation in the home</cause> at least <effect>eliminates cross-contamination </effect>

As in previous tasks, all annotations were performed independently by five annotators. Discrepancies were discussed and resolved by a sixth annotator.

Task 4: Gist Generation from Causal Relationships. In the final task, annotators were asked to synthesize a **causal gist** that captures the core causal relationship(s) expressed in each sentence identified as containing causal content. The gist generation process was conducted in two stages. First, individual annotators independently drafted candidate gists based on the previously annotated `<cause>` and `<effect>` spans. Second, all annotators participated in a group discussion to review, refine, and consolidate these gists into a single agreed-upon version per instance.

4.1 Annotation Process

We employed a dual annotation strategy to build both a high-quality gold-standard dataset and a large-scale silver-standard corpus.

Gold Annotations. Five annotators with backgrounds in causal linguistics and public health independently labeled 1,320 Reddit posts, randomly selected to ensure a balanced sample across 43 subreddits. Disagreements were resolved through group discussions led by a sixth annotator, resulting in consensus gold-standard labels for all annotation tasks. These tasks included causal classification (600 posts labeled as causal, 720 as non-causal), explicit vs. implicit causality distinction (243 implicit, 357 explicit), span extraction, and gist generation ($n = 600$ gists).

Silver Annotations. Building upon prior research in human–AI collaborative annotation (Li et al., 2024; Mirzakhmedova et al., 2024; Chen et al., 2024), we utilized GPT-4o with zero-shot learning to generate silver-standard labels for $8,800$ Reddit posts (randomly selected to ensure a balanced sample across 43 subreddits). The model was prompted using gold-standard exemplars within an RBIC-based (Ding et al., 2024) and chain-of-thought (Wei et al., 2022) instruction format (see Appendix B for prompt details). All model-generated annotations were subsequently reviewed and refined by the same team of six expert annotators. Specifically, to assess the quality of the silver-standard annotations, the full set of 8,800 posts was evaluated by human annotators according to the following dimensions:

- **Causality Accuracy**: Whether a causal relationship is correctly identified (binary: Yes/No).
- **Causality Type Accuracy**: Correct classification of the causal relation as explicit (1) or implicit (0).
- **Relevance (Span Extraction)**: Degree to which the extracted cause and effect spans capture the core causal relationship (5-point Likert scale).
- **Conciseness (Gist Generation)**: Degree to which the generated gist succinctly and coherently summarizes the identified causal relationship (5-point Likert scale).

To minimize error propagation across evaluation stages, annotations were assessed sequentially. Annotators first determined the presence and type of causal relationships. Posts identified as non-causal were excluded from subsequent evaluations. For posts with valid causal content, annotators then rated the relevance of the extracted cause-effect spans. If deemed relevant, they proceeded to assess the conciseness and coherence of the generated gist. Each evaluation criterion was rated independently by multiple annotators. Inter-annotator agreement was consistently high across all dimensions, as measured by Fleiss' Kappa (Fleiss, 1971) (See Table 3).

Following the evaluations, all annotators participated in a group adjudication session to conduct error analysis and resolve disagreements. During this process, annotators reviewed a sample of annotations with low agreement or identified errors and collectively discussed the rationale behind differing judgments. Final decisions were reached through consensus, and all inconsistencies were resolved manually to ensure the integrity of the silver-standard dataset.

Evaluation Criterion	**Score**	**Fleiss' κ**
Causality Accuracy	$ACC_{avg} = 0.902$	0.892
Causality Type Accuracy	$ACC_{avg} = 0.702$	0.780
Relevance (Span Extraction)	Mean = 4.30	0.839
Conciseness (Gist Generation)	Mean = 4.50	0.864

Table 3: Evaluation scores and inter-annotator agreement (Fleiss' κ) for each annotation dimension in the silver-standard dataset.

The finalized dataset includes 8,800 posts, with 3,054 labeled as causal ($1,484$ implicit, $1,570$ explicit) and 5,746 as non-causal. For each causal post, our dataset also provides cause-effect spans and corresponding gist summaries ($n = 3{,}054$).

5 Experimental Setup

To ensure fair evaluation, we randomly split each dataset (gold and silver) into 80% training and 20% testing sets, maintaining class balance in the testing set.

5.1 Models

We benchmark CausalTalk across four core tasks using a combination of discriminative and generative language models.

For Tasks 1–3 (binary classification, causality type classification, and span extraction), we fine-tune the following transformer-based discriminative models on both the gold-standard and silver-standard datasets: BERT-base (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019), XLNet-base (Yang et al., 2019), and DeBERTa-v3 (He et al., 2021), which are general-purpose pre-trained encoders commonly used for text classification. We also include SpanBERT (Joshi et al., 2020), a span-aware model optimized specifically for span-level prediction, making it particularly well-suited for the extraction of cause-effect relations.

For Task 4 (causal gist generation), we evaluate two paradigms: supervised fine-tuning and instruction-based prompting. In the supervised setting, we fine-tune T5-base (Raffel et al., 2020), FLAN-T5-base (Chung et al., 2024), GPT-2 (Radford et al., 2019), and BART-base (Lewis et al., 2020) on both the gold and silver datasets. For instruction-based evaluation, we employ instruction-tuned large language models, including LLaMA-3.2-3B (Grattafiori et al., 2024), Google

Gemini 2.0 Flash (Google, 2025), DeepSeek-V3 (Liu et al., 2024a), and Claude 3.5 Haiku (Anthropic, 2023). Each model is evaluated under both zero-shot and few-shot settings using prompts derived from gold-standard exemplars.

5.2 Evaluation Criteria

We assess model performance across four tasks that capture key dimensions of causal language understanding in social media.

Task 1: Causal Classification. To determine whether a post expresses a causal relationship, we evaluate models using weighted precision, recall, and F1 score, which account for class imbalance in binary classification.

Task 2: Explicit vs. Implicit Causality Detection. For posts identified as causal, this task classifies whether the causality is expressed explicitly or implicitly. We adopt the same set of metrics as in Task 1 to ensure consistency in evaluation under class-imbalanced conditions.

Task 3: Cause–Effect Span Extraction. We frame this as a sequence labeling task where models extract labeled spans corresponding to `<cause>` and `<effect>`. Following prior work on span extraction, we report token-level and span-level precision, recall, and F1 score.

Task 4: Causal Gist Generation. This generation task requires models to produce concise and coherent summaries of the underlying causal relationship. We adopt the ROUGE suite (ROUGE-1, ROUGE-2, ROUGE-L) (Lin, 2004), which measures n-gram overlap between model outputs and references. To assess semantic fidelity, we additionally report BERTScore (Zhang et al., 2019), which captures contextual similarity using pre-trained embeddings.

6 Result and Analysis

6.1 Task 1: Causal Classification

We fine-tuned four transformer-based models—BERT-base, RoBERTa-base, XLNet-base, and DeBERTa-v3—to classify Reddit posts as either causal or non-causal. Each model was trained and evaluated on both the gold-standard dataset and the silver-standard dataset. As shown in Table 4, all models perform consistently better on the silver dataset, likely due to its larger size, which provides stronger training signals despite being machine-annotated and subsequently human-verified.

Dataset	Model	Precision	Recall	F1 Score
Gold	BERT-base	$0.76_{0.023}$	$0.74_{0.023}$	$0.75_{0.024}$
	RoBERTa-base	$0.81_{0.021}$	$0.80_{0.020}$	$0.80_{0.021}$
	XLNet-base	$0.80_{0.021}$	$0.78_{0.020}$	$0.80_{0.021}$
	DeBERTa-v3*	$\mathbf{0.82_{0.021}}$	$\mathbf{0.80_{0.021}}$	$\mathbf{0.83_{0.022}}$
Silver	BERT-base	$0.81_{0.025}$	$0.79_{0.024}$	$0.80_{0.027}$
	RoBERTa-base	$0.85_{0.020}$	$0.83_{0.020}$	$0.84_{0.020}$
	XLNet-base	$0.84_{0.024}$	$0.82_{0.023}$	$0.83_{0.024}$
	DeBERTa-v3†	$\mathbf{0.87_{0.025}}$	$\mathbf{0.86_{0.024}}$	$\mathbf{0.87_{0.027}}$
$\Delta_{\text{model}^\dagger-\text{model}^*}$		↑ 0.05	↑ 0.06	↑ 0.04

Table 4: **Performance on Task 1 (Causal Classification) across gold and silver datasets.** Results are reported on the respective held-out test sets (20% of each dataset), with mean ± standard deviation over five random seeds. *Best performing model on Gold dataset; †Best performing model on Silver dataset. Green arrows indicate the improvement of the silver-trained best model relative to the gold-trained best model.

Across both datasets, DeBERTa-v3 outperforms all other models, demonstrating robustness to both limited training data and the inherent noise in silver annotations.

6.2 Task 2: Explicit vs. Implicit Detection

For posts labeled as causal, we evaluated model performance on identifying whether the causal relationship is conveyed explicitly or implicitly. This task was conducted on both the gold-standard and silver-standard datasets. As shown in Table 5, performance is consistently higher on the silver dataset, which is likely due to its greater size and coverage. Nevertheless, DeBERTa-v3 again has better performance across both datasets.

Dataset	Model	Precision	Recall	F1 Score
Gold	BERT-base	$0.61_{0.021}$	$0.59_{0.024}$	$0.58_{0.027}$
	RoBERTa-base	$0.61_{0.019}$	$0.60_{0.020}$	$0.60_{0.021}$
	XLNet-base	$0.63_{0.022}$	$0.62_{0.018}$	$0.63_{0.020}$
	DeBERTa-v3*	$\mathbf{0.68}_{0.017}$	$\mathbf{0.68}_{0.015}$	$\mathbf{0.69}_{0.016}$
Silver	BERT-base	$0.66_{0.026}$	$0.65_{0.025}$	$0.65_{0.027}$
	RoBERTa-base	$0.68_{0.022}$	$0.66_{0.023}$	$0.67_{0.019}$
	XLNet-base	$0.70_{0.021}$	$0.69_{0.019}$	$0.69_{0.020}$
	DeBERTa-v3†	$\mathbf{0.75}_{0.016}$	$\mathbf{0.74}_{0.015}$	$\mathbf{0.74}_{0.014}$
—	DeBERTa-v3‡	$0.69_{0.018}$	$0.70_{0.017}$	$0.70_{0.018}$

Table 5: Performance on Task 2 (Explicit vs. Implicit Causality Classification). Results are mean ± standard deviation over five random seeds on the respective held-out test sets (20% of each dataset). *Best model on the Gold test set; †Best model on the Silver test set; ‡*Silver-trained* DeBERTa-v3 evaluated on the *Gold* test set (cross-evaluation).

6.3 Task 3: Cause–Effect Span Extraction

We frame span extraction as a token classification task, where models identify tokens corresponding to the <cause> and <effect> spans. We evaluate performance using two complementary metrics: token-level F1 and span-level overlap, capturing both fine-grained tagging accuracy and holistic span correctness.

Dataset	Model	Precision	Recall	F1
Gold	**BERT-base**			
	- Token	$0.82_{0.012}$	$0.83_{0.014}$	$0.82_{0.013}$
	- Span	$0.71_{0.015}$	$0.69_{0.015}$	$0.70_{0.014}$
	SpanBERT			
	- Token	$0.84_{0.011}$	$0.85_{0.013}$	$0.84_{0.012}$
	- Span	$0.75_{0.014}$	$0.73_{0.015}$	$0.74_{0.014}$
	RoBERTa-base			
	- Token	$0.87_{0.010}$	$0.87_{0.011}$	$0.87_{0.010}$
	- Span	$0.79_{0.013}$	$0.77_{0.014}$	$0.78_{0.013}$
	DeBERTa-v3			
	- Token*	$\mathbf{0.89}_{0.010}$	$\mathbf{0.89}_{0.010}$	$\mathbf{0.89}_{0.010}$
	- Span⋆	$\mathbf{0.82}_{0.012}$	$\mathbf{0.80}_{0.013}$	$\mathbf{0.81}_{0.012}$
Silver	**BERT-base**			
	- Token	$0.89_{0.011}$	$0.90_{0.012}$	$0.88_{0.011}$
	- Span	$0.78_{0.014}$	$0.76_{0.015}$	$0.77_{0.014}$
	SpanBERT			
	- Token	$0.91_{0.010}$	$0.92_{0.011}$	$0.91_{0.010}$
	- Span	$0.82_{0.013}$	$0.80_{0.014}$	$0.81_{0.013}$
	RoBERTa-base			
	- Token	$0.94_{0.010}$	$0.94_{0.010}$	$0.94_{0.010}$
	- Span	$0.86_{0.012}$	$0.84_{0.013}$	$0.85_{0.012}$
	DeBERTa-v3			
	- Token†	$\mathbf{0.95}_{0.010}$	$\mathbf{0.95}_{0.010}$	$\mathbf{0.95}_{0.010}$
	- Span‡	$\mathbf{0.89}_{0.011}$	$\mathbf{0.87}_{0.012}$	$\mathbf{0.88}_{0.011}$
$\Delta_{\text{Token}^\dagger-\text{Token}^*}$		↑ 0.06	↑ 0.06	↑ 0.06
$\Delta_{\text{Span}^\ddagger-\text{Span}^\star}$		↑ 0.07	↑ 0.07	↑ 0.07

Table 6: Performance on Task 3 (Cause–Effect Span Extraction) between gold and silver standard datasets. Each model is evaluated using both token-level and span-level metrics. *Best token-level model on Gold dataset; ⋆Best span-level model on Gold dataset; †Best token-level model on Silver dataset; ‡Best span-level model on Silver dataset. Green arrows indicate performance improvements of Silver over Gold dataset.

As shown in Table 6, model performance on the silver-standard dataset is comparable to that on the gold-standard dataset, with only minor variations across evaluation metrics. This suggests that well-verified silver data can serve as a viable alternative to gold annotations, even for structured prediction tasks requiring fine-grained span supervision. Across both datasets and evaluation levels, DeBERTa-v3 yields the highest scores, followed closely by RoBERTa-base and SpanBERT.

To further contextualize our findings, we benchmark CausalTalk against a range of recent causal datasets (Mirza et al., 2014b; Wang et al., 2022; Caselli and Vossen, 2017; Mihăilă et al., 2013; Lai et al., 2022). As shown in Appendix Table 15, our span extraction task achieves competitive performance, particularly on implicit causality, while also complementing datasets that focus on event-level or reasoning-oriented tasks.

6.4 Task 4: Causal Gist Generation

We evaluate causal gist generation using four metrics: ROUGE-1, ROUGE-2, ROUGE-L, and BERTScore. The task involves generating concise summaries that capture the core causal relationship within each Reddit post. In this section, we only report results for the silver standard dataset, due to the size of the dataset.

As shown in Table 7, we compare two model categories: (1) supervised fine-tuned models (SFT) trained on the silver dataset, and (2) instruction-tuned large language models (LLMs) evaluated using zero-shot and few-shot prompting. Among fine-tuned models, FLAN-T5-base performs best across most metrics. Among LLMs, Google Gemini 2.0 Flash, DeepSeek-V3, and Claude 3.5 Haiku demonstrate strong zero-shot and few-shot performance, often surpassing supervised baselines. In contrast, Claude 3.5 Haiku and GPT-2 underperform on all metrics.

6.5 Error Analysis

Our error analysis revealed that models often miss implicitly causal instances (in Task 1), which we explain through confusion matrices and representative examples. Beyond this, Task 2 requires fine-grained semantic distinctions, while Tasks 3 and 4 involve structured or generative outputs in which aggregate metrics (e.g., ROUGE, BERTScore) may obscure deeper errors. We therefore conduct a qualitative analysis of Tasks 2–4 to expose systematic failure modes and guide future improvements.

Task 1: Causal vs. Non-Causal. Although classification scores are stable, confusion matrices (Tables 13) reveal the weakness: some implicitly causal sentences are not recognized as causal, leading to false negatives. These cases typically involve causal meaning inferred from context (e.g., presupposed health changes, life events) rather than explicit connectives. This weakness shows that current models struggle to use context and discourse cues, which limits their ability to detect causality when it is not indicated by explicit words.

Model	Type	Causal Gist Generation			
		ROUGE-1	ROUGE-2	ROUGE-L	BERTScore
T5-base	SFT	$0.429_{0.012}$	$0.334_{0.009}$	$0.512_{0.013}$	$0.670_{0.016}$
FLAN-T5-base*	SFT	$\mathbf{0.559}_{0.007}$	$0.354_{0.011}$	$0.521_{0.008}$	$0.704_{0.012}$
GPT-2	SFT	$0.281_{0.014}$	$0.089_{0.019}$	$0.235_{0.017}$	$0.305_{0.016}$
BART-base	SFT	$0.442_{0.015}$	$0.261_{0.010}$	$0.400_{0.009}$	$0.528_{0.013}$
LLaMA-3.2-3B	zero-shot	$0.432_{0.013}$	$0.243_{0.017}$	$0.400_{0.014}$	$0.532_{0.011}$
	few-shot	$0.448_{0.012}$	$0.235_{0.016}$	$0.417_{0.013}$	$0.551_{0.012}$
Google Gemini†	zero-shot	$0.557_{0.010}$	$\mathbf{0.436}_{0.015}$	$\mathbf{0.588}_{0.012}$	$\mathbf{0.764}_{0.009}$
	few-shot	$0.545_{0.011}$	$0.427_{0.014}$	$0.574_{0.011}$	$0.745_{0.010}$
DeepSeek-V3	zero-shot	$0.526_{0.008}$	$0.411_{0.016}$	$0.568_{0.012}$	$0.731_{0.014}$
	few-shot	$0.537_{0.009}$	$0.422_{0.015}$	$0.549_{0.013}$	$0.715_{0.013}$
Claude 3.5 Haiku	zero-shot	$0.436_{0.019}$	$0.210_{0.018}$	$0.356_{0.016}$	$0.462_{0.017}$
	few-shot	$0.423_{0.018}$	$0.221_{0.016}$	$0.366_{0.015}$	$0.475_{0.016}$
$\Delta_{\text{Gemini}^\dagger-\text{FLAN-T5}^\star}$		$\downarrow$ 0.002	$\uparrow$ 0.082	$\uparrow$ 0.067	$\uparrow$ 0.060

Table 7: Task 4: Performance of causal gist generation on the silver-standard dataset. The upper section includes supervised fine-tuned models (SFT), while the lower section shows zero-shot and few-shot prompting results from instruction-tuned LLMs. *Best SFT model; †Best overall model. Green arrows indicate performance improvements of Gemini over FLAN-T5, while red arrows indicate performance decreases.

Task 2: Explicit vs. Implicit. The misclassifications observed in Task 1 motivated a closer examination of implicit cases in Task 2, as this task strictly assumes causal instances as input. Models frequently misclassify implicitly causal statements, especially when causality is conveyed by discourse context or pragmatic presupposition rather than explicit connectives (e.g., *because*, *so*). We observe (i) implicit cases labeled as explicit, reflecting reliance on surface connective-like signals and the well-documented label-shift between explicit and implicit data; and (ii) the converse, where explicit connectives functioning rhetorically or temporally are down-weighted, leading to implicit labels. These patterns echo known challenges in implicit relation recognition—ambiguity among senses (Liu et al., 2024b; Lin et al., 2009), the need for context and world knowledge, and the pitfalls of connective reliance.

Task 3: Cause–Effect Span Extraction. Span-level F1 drops whenever causal expressions are long or nested. Under-extended spans miss essential modifiers (e.g., temporal clauses), whereas over-extended spans absorb irrelevant material from coordinate or subordinate clauses. In complex sentences, models often identify the correct tokens but fail to align exact boundaries, suggesting that token-wise tagging alone is insufficient for precise span delimitation.

Task 4: Causal Gist Generation. Despite the automatic scores, generated gists sometimes omit the causal link or hallucinate unsupported content. Omissions are most common when multiple events occur in a single post; the model selects a salient event but loses the causal connection. Hallucinations typically emerge in long posts with ambiguous discourse structure, where the model invents a cause or effect to produce a fluent but semantically erroneous summary. To illustrate, we provide representative examples of systematic failures in Appendix Table 16.

7 Conclusion

CausalTalk addresses a critical gap in NLP resources by providing a comprehensive multi-level dataset for causal language analysis. Our benchmark results demonstrate both the utility of the dataset and the continued challenges in automated causal reasoning, particularly for implicit causality. The four interconnected tasks enable research that bridges detection, classification, extraction, and generation of causal information. By releasing this resource, we aim to accelerate research in causal reasoning and natural language understanding across domains.

Limitations

While CausalTalk represents an advancement in resources for causal language analysis, several limitations should be acknowledged. First, Reddit's user base is not demographically representative of the general population, potentially limiting the diversity of causal expressions. Second, our annotation focus on English excludes causal language patterns that may be unique to other languages. Third, despite verification procedures, our silver annotations may contain systemic biases from the underlying GPT-4o model. Finally, our focus on submission-level annotations may obscure some nuances in causal relationships that depend on broader conversational context and trajectories.

In future work, we plan to expand the dataset by including posts from multilingual forums (e.g., r/AskEurope, r/mexico), supplement annotations with demographically diverse annotator pools, and pilot a context-aware annotation tool that allows annotators to view preceding and following comments to better preserve causal coherence. We also aim to systematically compare zero-shot and few-shot prompting strategies to better understand their trade-offs in annotation quality, stability, and bias.

8 Ethics Statement

All data used in this study were collected from publicly available Reddit posts. To protect user privacy, we removed any personally identifiable information and anonymized user IDs. We restricted our dataset to English-language posts and applied filters to exclude duplicates and extremely short entries.

References

Zoë Adams, Magda Osman, Christos Bechlivanidis, and Björn Meder. 2023. (why) is misinformation a problem? *Perspectives on Psychological Science*, 18(6):1436–1463.

Wajid Ali, Wanli Zuo, Rahman Ali, Xianglin Zuo, and Gohar Rahman. 2021. Causality mining in natural languages using machine and deep learning techniques: A survey. *Applied Sciences*, 11(21):10064.

Anthropic. 2023. Claude (oct 8 version). https://www.anthropic.com/. Large language model.

Eduardo Blanco, Nuria Castell, and Dan I Moldovan. 2008. Causal relation extraction. In *Lrec*, volume 66, page 74.

Charles J Brainerd and Valerie F Reyna. 1990. Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review*, 10(1):3–47.

Tommaso Caselli and Piek Vossen. 2017. Event storyline corpus: A new resource for causal and temporal relation extraction. *Proceedings of the Events and Stories in the News Workshop*.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17772–17780.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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.

Mahdi Dhaini, Ege Erdogan, Smarth Bakshi, and Gjergji Kasneci. 2024. Explainability meets text summarization: A survey. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 631–645.

Xiaohan Ding, Buse Carik, Uma Sushmitha Gunturi, Valerie Reyna, and Eugenia Ha Rim Rho. 2024. Leveraging prompt-based large language models: predicting pandemic health decisions and outcomes through social media language. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–20.

David Dukić, Kiril Gashteovski, Goran Glavaš, and Jan Šnajder. 2023. Leveraging open information extraction for more robust domain transfer of event trigger detection. *arXiv preprint arXiv:2305.14163*.

Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Google. 2025. Gemini (2.0 flash). https://gemini.google.com/. Large language model.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Mourad Heddaya, Qingcheng Zeng, Chenhao Tan, Rob Voigt, and Alexander Zentefis. 2024. Causal micro-narratives. *arXiv preprint arXiv:2410.05252*.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O'Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*.

Amey Hengle, Atharva Kulkarni, Shantanu Patankar, Madhumitha Chandrasekaran, Sneha D'Silva, Jemima Jacob, and Rashmi Gupta. 2024. Still not quite there! evaluating large language models for comorbid mental health diagnosis. *arXiv preprint arXiv:2410.03908*.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Jiafeng Guo, and Xueqi Cheng. 2025. Large language model-based event relation extraction with rationales. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7484–7496.

Shumpei Inoue, Minh-Tien Nguyen, Hiroki Mizokuchi, Tuan-Anh D Nguyen, Huu-Hiep Nguyen, and Dung Tien Le. 2023. Towards safer operations: An expert-involved dataset of high-pressure gas incidents for preventing future failures. *arXiv preprint arXiv:2310.12074*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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

Manolis Kyriakakis, Ion Androutsopoulos, Artur Saudabayev, and 1 others. 2019. Transfer learning for causal sentence detection. *arXiv preprint arXiv:1906.07544*.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Meci: A multilingual dataset for event causality identification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2346–2356.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*.

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.

Chunyang Li, Hao Peng, Xiaozhi Wang, Yunjia Qi, Lei Hou, Bin Xu, and Juanzi Li. 2024. MAVEN-FACT: A large-scale event factuality detection dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11140–11158, Miami, Florida, USA. Association for Computational Linguistics.

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

Chi-Jen Lin and Wei-Wen Wu. 2008. A causal analytical method for group decision-making under fuzzy environment. *Expert Systems with Applications*, 34(1):205–213.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Wei Liu, Stephen Wan, and Michael Strube. 2024b. What causes the failure of explicit to implicit discourse relation recognition? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2738–2753, Mexico City, Mexico. Association for Computational Linguistics.

Yan Liu, Yazheng Yang, and Xiaokang Chen. 2024c. Improving long text understanding with knowledge distilled from summarization model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11776–11780. IEEE.

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. *arXiv preprint arXiv:1907.11692*.

Claudiu Mihăilă, Tomoko Ohta, Sampo Pyysalo, and Sophia Ananiadou. 2013. Biocause: Annotating and analysing causality in the biomedical domain. *BMC bioinformatics*, 14:1–18.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014a. 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.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014b. 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.

Nailia Mirzakhmedova, Marcel Gohsen, Chia Hao Chang, and Benno Stein. 2024. Are large language models reliable argument quality annotators? In *Conference on Advances in Robust Argumentation Machines*, pages 129–146. Springer.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination in causal q&a: The efficacy of fine-tuning over prompting in llms. In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 253–258.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56.

Martin J Pickering and Asifa Majid. 2007. What are implicit causality and consequentiality? *Language and Cognitive Processes*, 22(5):780–788.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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. *Journal of machine learning research*, 21(140):1–67.

Ines Rehbein and Josef Ruppenhofer. 2017. Catching the common cause: Extraction and annotation of causal relations and their participants. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 105–114, Valencia, Spain. Association for Computational Linguistics.

Valerie F Reyna. 2012. A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making*, 7(3):332–359.

Valerie F Reyna. 2021. A scientific theory of gist communication and misinformation resistance, with implications for health, education, and policy. *Proceedings of the National Academy of Sciences*, 118(15):e1912441117.

Ted Sanders and Wilbert Spooren. 2009. Causal categories in discourse: Converging evidence from language use. *Causal categories in discourse and cognition*, pages 205–246.

Ted Sanders and Eve Sweetser. 2009. Introduction: Causality in language and cognition–what causal connectives and causal verbs reveal about the way we think. *Causal categories in discourse and cognition*, pages 1–18.

Youngseo Son, Nipun Bayas, and H Andrew Schwartz. 2018. Causal explanation analysis on social media. *arXiv preprint arXiv:1809.01202*.

Yidan Sun, Qin Chao, and Boyang Li. 2023. Event causality is key to computational story understanding. *arXiv preprint arXiv:2311.09648*.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, and 1 others. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. *arXiv preprint arXiv:2211.07342*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Longxuan Yu, Delin Chen, Siheng Xiong, Qingyang Wu, Dawei Li, Zhikai Chen, Xiaoze Liu, and Liangming Pan. 2025. Causaleval: Towards better causal reasoning in language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12512–12540.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Xiyang Zhang, Muhao Chen, and Jonathan May. 2021. Salience-aware event chain modeling for narrative understanding. *arXiv preprint arXiv:2109.10475*.

A Subreddit Descriptions

We selected 43 subreddits collaboratively, guided by prior literature (Ding et al., 2024) and a systematic review of subreddit activity and relevance. The chosen subreddits reflect a diverse range of public health discussions—especially those related to the COVID-19 pandemic—covering core medical communities, regional updates, vaccine-related experiences, commentary and satire, and conspiracy or anti-lockdown narratives.

#	Subreddit	Description
01	COVID19	General global discussion of COVID-19
02	COVID19_Pandemic	Global news and data tracking
03	COVID19_commentary	Commentary and policy-related discourse
04	COVID19_support	Emotional support and mutual aid
05	COVID19positive	First-hand accounts from those who tested positive
06	Coronavirus	Largest global COVID-19 discussion hub
07	CoronaVirus_2019_nCoV	Early-stage discussion board on COVID-19
08	covid19_testimonials	Personal recovery and infection experiences

Table 8: Core COVID-19 Discussion Subreddits

#	Subreddit	Description
09	CoronaVirusPA	COVID-19 updates in Pennsylvania
10	CoronaVirusTX	COVID-19 updates in Texas
11	CoronavirusAZ	Arizona pandemic coverage
12	CoronavirusCA	California-specific COVID-19 news
13	CoronavirusColorado	COVID-19 in Colorado
14	CoronavirusGA	COVID-19 in Georgia
15	CoronavirusIllinois	Illinois pandemic news
16	CoronavirusMN	Minnesota pandemic updates
17	CoronavirusMa	Massachusetts-specific content
18	CoronavirusMichigan	Michigan COVID-19 discussion
19	CoronavirusNewYork	New York State pandemic updates
20	CoronavirusOC	Orange County, CA updates
21	CoronavirusUS	National-level U.S. COVID-19 discourse
22	CoronavirusWA	Washington State discussion
23	CoronavirusWI	Wisconsin COVID-19 updates
24	Covid19_Ohio	Ohio-specific coverage
25	FloridaCoronavirus	Florida COVID-19 updates

Table 9: U.S. Region-Specific COVID-19 Subreddits

#	Subreddit	Description
26	ChurchOfCOVID	Satirical religious-style meme community
27	CoronavirusCirclejerk	Meme-based COVID-19 commentary
28	CoronavirusMemes	Humorous content about the pandemic
29	CoronavirusRecession	Discussion of COVID-induced economic impact
30	CoronavirusFOS	Niche community with unclear focus

Table 10: Commentary, Satire, and Meme Subreddits

#	Subreddit	Description
31	DebateVaccines	Vaccine controversies and open debate
32	covidvaccinateduncut	Side effect reports and dissenting opinions
33	vaccinelonghaulers	Long-term vaccine symptom experiences
34	vaccinelong_haulers	Variant spelling of above with similar content
35	VAERSreports	Discussion based on VAERS adverse event data

Table 11: Vaccine Experience and Side Effect Subreddits

#	Subreddit	Description
36	FauciForPrison	Criticism and distrust of Dr. Anthony Fauci
37	LockdownSkepticism	Opposition to global lockdown policies
38	NoNewNormal	Rejection of pandemic-era norms and mandates
39	TrueAntiVaccination	Strongly anti-vaccine discourse
40	CovidIsAFraud	Claims that COVID-19 is a hoax
41	Covid19Origin	Discussion of COVID origins (e.g., lab leak)
42	ivermectin	Advocacy of ivermectin for COVID treatment
43	wuhanflu	Politicized and discriminatory COVID-19 subreddit

Table 12: Anti-lockdown / Conspiracy / Anti-vaccine Subreddits

B Causal Reasoning Prompt Flow with Explicitness Identification

For silver-standard annotation, we used OpenAI's GPT-4o model via the OpenAI API with the following configuration: temperature: 0.0, top_p: 1.0, and stop: None.

In this prompting flow, each P (e.g., P1–P5) represents a user-issued prompt, and each O (e.g., O1–O5) denotes the AI model's response. The structure follows a step-by-step Chain-of-Thought (CoT) reasoning process to guide the model through understanding, identifying, and explaining cause-effect relationships in social media posts.

- **P1.** Your role is to understand the cause-effect relationships in social media posts. Can you provide a brief definition of what a cause-effect relationship is?

- **O1.** Certainly! A cause-effect relationship is a relationship between two events or variables where one is understood to be a consequence of the other.

- **P2.** Based on your role, can you explain the term, "causal gist" in relation to sentences that have causal coherence?

- **O2.** Of course. The term "causal gist" refers to the fundamental meaning or essence of a sentence or text that expresses a causal relationship, usually highlighting the cause, effect, and the direction of influence.

- **P3.** So, given the sentence: *I took the vaccine yesterday. I'm really sick now.*
  Is there a cause-effect relationship in this given sentence?
  - If yes, just answer: "Yes"
  - If no, just answer: "No"
  - Don't give me any explanations

- **O3.** Yes

- **P3.5.** Given that there is a cause-effect relationship in the sentence, please classify it as either:
  - `"Explicit"`: if the causal connection is clearly stated using connectives or causal language;
  - `"Implicit"`: if the causal relationship is implied but not directly stated.

  Just answer with one word: `Explicit` or `Implicit`

- **O3.5.** Implicit

- **P4A.** Indeed, there is a cause-effect relationship in the given sentence.
  Then extract the corresponding cause phrase and effect phrase in the given sentence.
  Just respond in JSON format: `{"Cause": "", "Effect": ""}`

- **O4.** Sure: `{"Cause": "took the vaccine", "Effect": "really sick now"}`

- **P5.** Generate a reasonable and clear causal gist based on `{"Cause": "took the vaccine", "Effect": "really sick now"}` and your understanding of the sentence with the cause-effect relationship.

- **O5.** Taking the vaccine yesterday caused the person to become sick.

C Error Analysis for Task 1

In Task 1 (Causal vs. Non-Causal classification), we examined model errors by inspecting the confusion matrices on both the gold-standard and silver-standard test sets (Table 13).

	Pred. Non-Causal	Pred. Causal	Total
Gold-test			
Non-Causal	119	13	132
Causal	23	109	132
Total	142	122	264
Silver-test			
Non-Causal	792	88	880
Causal	153	727	880
Total	945	815	1760

Table 13: DeBERTa confusion matrices for Task 1 on the Gold-test and Silver-test datasets.

D CausalTalk Dataset Examples

We provide five annotated examples from the CausalTalk dataset to illustrate the annotation schema for the four causal tasks: causal gist generation. These examples demonstrate how causal relationships are identified and analyzed in informal social media posts related to public health topics, particularly in the context of COVID-19 discussions.

These examples illustrate how the CausalTalk dataset bridges fine-grained causal detection and gist-based reasoning over informal text.

E Cross-Dataset Benchmarking

Table 15 presents comparative results across multiple cause–effect span extraction benchmarks using a consistent DeBERTa model. While datasets such as Causal-TimeBank and MAVEN-ERE emphasize event-event causal links (Wang et al., 2022; Mirza et al., 2014b), CausalTalk combines span extraction with implicit vs. explicit causality, achieving strong overall performance.

F Error Analysis for Task 4

Post	Type	Cause	Effect	Causal Gist
My neighbor attended a large wedding party last week. Now their entire family is in isolation at home. They were supposed to visit us this weekend, but all our plans had to be canceled.	Implicit	attended a large wedding party last week	entire family is in isolation at home	Attending the large wedding party led to the family needing to isolate.
Due to the new mask mandate at our local hospital, I had to reschedule my non-emergency surgery until I could get a negative test result. The hospital staff explained that because of the recent outbreak in our county, they've implemented stricter protocols for all patients.	Explicit	the new mask mandate at our local hospital	had to reschedule my non-emergency surgery	The new mask mandate at the local hospital caused the rescheduling of the non-emergency surgery.
I started taking vitamin D supplements three months ago after reading about its immune benefits. Haven't had even a minor cold since then, despite everyone in my office getting sick.	Implicit	started taking vitamin D supplements three months ago	Haven't had even a minor cold since then	Taking vitamin D supplements prevented the person from getting even a minor cold.
My roommate had to quit his job at the restaurant because their new policy requires all staff to be vaccinated and he refuses to get the shot. Since he lost his income, I'm now covering most of our rent and utilities until he finds something else.	Explicit	their new policy requires all staff to be vaccinated and he refuses to get the shot	had to quit his job at the restaurant	The restaurant's vaccination requirement combined with the roommate's refusal to get vaccinated caused him to quit his job.
Schools in our district switched to online learning last Monday. My kids haven't been able to focus on their assignments, and their grades are dropping dramatically. I've had to take time off work to supervise them during class hours, which is putting additional stress on our family finances.	Implicit	Schools in our district switched to online learning	kids haven't been able to focus on their assignments, and their grades are dropping dramatically	The switch to online learning resulted in the children's inability to focus and declining academic performance.

Table 14: Examples from the CausalTalk dataset showing different types of causality in social media posts. For each example, we provide the original post text, the type of causality (explicit or implicit), the extracted cause and effect spans, and the generated causal gist that concisely summarizes the causal relationship.

Dataset	Task Type	Implicit Causes	Model	Precision	Recall	F1
CausalTalk (Ours)	Span extraction + Implicit/Explicit	Yes	DeBERTa	0.75	0.74	0.74
Causal-TimeBank	Event-event causal links (span)	No	DeBERTa	0.70	0.76	0.73
MAVEN-ERE	Large-scale event causality	No	DeBERTa	0.73	0.78	0.75
MECI (EN)	Event-level causality (span)	Yes	DeBERTa	0.72	0.77	0.74
BioCause	Biomedical cause-effect (span)	No	DeBERTa	0.58	0.52	0.55
EventStoryLine	Document-level event causality	Yes	DeBERTa	0.35	0.42	0.38

Table 15: Cross-dataset performance comparison for cause–effect span extraction tasks using DeBERTa.

Post: Ever since I recovered from COVID, climbing stairs feels like a chore. I used to be active, but now even short walks tire me out. **Model's Generated Gist:** The user has become lazy after recovering from COVID. **Human Analysis:** After recovering from COVID, the user experiences fatigue during physical activity. **Error Analysis:** Implicit causality is overlooked; the model hallucinates a subjective explanation ("lazy") that distorts the causal structure.
Post: I was laid off in 2020 and had to move back in with my parents. While living there, I started binge eating and gained 30 pounds. **Model's Generated Gist:** The user binge ate and gained weight while living with their parents. **Human Analysis:** Being laid off led the user to move back home, which contributed to emotional eating and weight gain. **Error Analysis:** The model compresses the chain to "living with parents → binge eating," missing the prior causal link from job loss.
Post: I haven't seen my friends in months. Most days I just lie in bed, scrolling through my phone, not feeling like doing anything. The lockdowns really changed my routine. **Model's Generated Gist:** The user feels unmotivated as a result of spending most days in bed. **Human Analysis:** The user felt unmotivated because of the routine changes caused by COVID lockdowns. **Error Analysis:** The model omits the true cause ("lockdowns"), attributing lack of motivation to a symptom rather than the underlying event.
Post: Since the pandemic started, my anxiety has gotten worse. I can't focus on work and my performance reviews have slipped. **Model's Generated Gist:** The user performs poorly at work because they are not motivated. **Human Analysis:** The stress and anxiety caused by the pandemic reduced the user's ability to focus, which in turn affected their job performance. **Error Analysis:** The model fails to capture the implicit causal chain (pandemic → anxiety → reduced focus → poor performance), instead attributing the outcome to lack of motivation.
Post: After months of working from home, I've gained weight. I barely walk anymore and end up snacking all day. **Model's Generated Gist:** The user gained weight because they eat too much. **Human Analysis:** The shift to remote work during the pandemic reduced daily activity and increased snacking, which together led to weight gain. **Error Analysis:** The model captures only the surface factor ("eating") and misses the underlying contextual cause (remote work → sedentary lifestyle → snacking → weight gain).

Table 16: Representative examples of causal reasoning errors.