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

EMNLP 2025
Suzhou, China | 中国苏州

VIRGINIA TECH

Xiaohan Ding, Kaike Ping, Buse Carik, Eugenia Rho

{xiaohan, pkk, buse, eugenia} @ vt.edu

Background & Motivation - *Why Causal Language Understanding?*

- Causal reasoning is central to NLP but underexplored in informal social media text.
- Existing datasets focus on explicit causality in structured domains (e.g., news, biomedical).
- Social media: messy, implicit, gist-driven → critical for public health, misinformation, decision-making.
- <u>**Gap**</u>: lack of datasets that connect detection, classification, extraction, and gist generation

Dataset - CausalTalk

- **Source:** Reddit posts (2020–2024), 43 public health subreddits (COVID-19, vaccines, lockdowns).
- **Size:** 239k submissions + 19M comments; 10,120 annotated posts.
- **Annotation Tasks:**
  - Binary causal classification
  - Explicit vs. Implicit causality
  - Cause–effect span extraction
  - Causal gist generation
- **Annotations:**
  - Gold: 1,320 posts (expert consensus)
  - Silver: 8,800 posts (GPT-4o + human verification)

Methodology - *Annotation & Benchmark Setup*

Annotation Pipeline

- **Gold Annotations (n = 1,320 posts):**
  - Five annotators trained in causal linguistics + public health.
  - Independent labeling for 4 tasks → disagreements resolved via sixth adjudicator.
  - Outputs: causal classification, explicit/implicit type, span extraction, causal gists.
- **Silver Annotations (n = 8,800 posts):**
  - Generated by GPT-4o with RBIC.
  - Human annotators verified & refined every output; and Dimensions assessed:
    1. **Causality Accuracy** (binary correctness)
    2. **Type Accuracy** (explicit vs. implicit)
    3. **Span Relevance** (5-point Likert)
    4. **Gist Conciseness & Coherence** (5-point Likert)
  - **High inter-annotator agreement** (κ ≈ 0.78–0.89 across tasks).

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

Benchmark Setup

- **Task Coverage: (1)** Binary causal classification; (2) Explicit vs. implicit detection; (3) Cause–effect span extraction; (4) Causal gist generation
- **Models (Tasks 1–3):**
  - BERT-base, RoBERTa-base, XLNet-base, DeBERTa-v3, SpanBERT
- **Models (Task 4):**
  - Fine-tuned: T5, FLAN-T5, GPT-2, BART
  - Instruction-tuned LLMs: LLaMA-3.2, Gemini 2.0 Flash, DeepSeek-V3, Claude 3.5 Haiku

Main Results & Insights - *Benchmark Findings*

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

Table1: 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.

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 2: Performance on Task 2 (Explicit vs. Implicit Causality Classification). Results are mean ± standard deviation over five random seeds on the respective heldout test sets (20% of each dataset).

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	$0.423_{0.018}$	$0.221_{0.016}$	$0.366_{0.015}$	$0.475_{0.016}$
$\Delta_{\text{Gemini}^\dagger-\text{FLAN-T5}^*}$		↓ 0.002	↑ 0.082	↑ 0.067	↑ 0.060

Table 3: Performance on Task 3 (Cause–Effect SpanExtraction) between gold and silver standard datasets.Each model is evaluated using both token-level and spanlevel metrics.

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
$\Delta_{\text{Span}^\ddagger-\text{Span}^*}$		↑ 0.07	↑ 0.07	↑ 0.07

Table 4: Performance of causal gist generation on the silver-standard dataset. The upper section includessupervised fine-tuned models (SFT), while the lower section shows zero-shot and few-shot prompting results frominstruction-tuned LLMs.

1. **Implicit Causality Remains Challenging**
   - Models frequently misclassify implicit cases that lack explicit markers (e.g., because, therefore).Error analysis shows difficulty in leveraging discourse context and world knowledge.
2. **Span Extraction is Sensitive to Boundaries**
   - Models identify correct cause/effect tokens but often fail on span boundary precision.Over-extended spans absorb irrelevant clauses; under-extended spans miss essential modifiers.
3. **Silver Data Helps Generalization**
   - Silver-trained models generalize well even when evaluated on gold test sets (cross-evaluation).Suggests verified silver data is a scalable strategy for building causal benchmarks.

Main Results & Insights - *Benchmark Findings*

Contribution:

Introduced **CausalTalk**, the first multi-level benchmark for causal language understanding in informal social media discourse.

Covers four interconnected tasks: binary classification, explicit vs. implicit detection, span extraction, and causal gist generation.

Combines gold-standard expert annotations with large-scale silver-standard annotations (GPT-4o + human verification), achieving both quality and scalability.

Key Findings:

Silver-standard data, once human-verified, is highly reliable and improves robustness, especially for **implicit causality**.

Benchmark results show state-of-the-art models still **struggle with implicit and gist-based causality**, highlighting an open challenge in NLP.

Demonstrates the importance of aligning NLP resources with **human cognitive strategies** (fuzzy-trace theory, gist reasoning).

Future Work

Dataset Expansion:

Incorporate **multilingual forums** (e.g., r/AskEurope, r/mexico) to capture culturally diverse causal expressions.

Extend beyond Reddit to other platforms for **cross-platform causal discourse analysis**.

Annotation Enhancements:

Recruit more **demographically diverse annotators** to reduce bias.

Develop **context-aware annotation tools** allowing annotators to consider surrounding conversation, not just single posts.