

From Text to Structure: The PACKaGE Benchmark for Bidirectional Causal Reasoning in Paragraphs and Graphs

Juhyeon Kim¹, Dohun Lee¹, Jiwon Jang¹, Joseph Shim¹, Hyunwoo Park¹

¹Graduate School of Data Science, Seoul National University

Correspondence: Hyunwoo Park hyunwoopark@snu.ac.kr

Abstract

Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains, yet they struggle with causal reasoning tasks. We introduce PACKaGE (Paragraph-level Annotated Causal Knowledge and Graph-tExt), a causal benchmark dataset that pairs paragraph-level texts with explicit causal graphs. Unlike prior sentence-level resources, our benchmark dataset (i) operates at paragraph level to capture various and interdependent causal pathways; (ii) offers multi-modal annotations including full graphs and contextualized relations; and (iii) maps extended passages to comprehensive graphs. We detail a four-stage construction pipeline: collecting causality-labeled and fact-centric texts, inducing graphs from span-tagged pairs, condensing nodes, and recomposing path-aligned narratives. Overall, the paired evaluation highlights a persistent discrepancy between linguistic competence and causal understanding in LLMs, pointing to the need for further development.

1 Introduction

Recent studies have established that Large Language Models (LLMs) can perform domain-specific reasoning at an expert level across diverse domains such as medicine, chemistry, and physics (OpenAI et al., 2024; Grattafiori et al., 2024). For example, Singhal et al. (2022) and Nori et al. (2023) show that LLMs can handle clinical reasoning tasks. Lewkowycz et al. (2022) present a model, which solves mathematics and physics reasoning problems.

Despite these advances, LLMs exhibit systematic failures in causal reasoning—relying on superficial heuristics rather than genuine inference (Yamin et al., 2024; Zečević et al., 2023), struggling even with chain-of-thought prompting (Jin et al., 2024, 2023), and failing to calibrate uncertainty in causal judgments (Yin et al., 2023). While

various methods were proposed to convey graph-based and causal knowledge to LLMs (Besta et al., 2024; Edge et al., 2024; Wang et al., 2025), existing datasets remain limited by low graph complexity or weak alignment between text and graphs. This absence prevents systematic evaluation of a critical capability: whether LLMs can bidirectionally translate between causal language and causal structure—extracting graphs from text and generating text that preserves graph semantics. Leveraging graphs—particularly Directed Acyclic Graphs (DAGs)—offers a principled way to enhance causal reasoning in LLMs by representing causal structure (Yamin et al., 2024). However, despite the growing interest in graph-based reasoning paradigms (Besta et al., 2024; Wang et al., 2025), there currently exists no dataset in which textual narratives and their corresponding causal graphs are paired, limiting systematic evaluation and development.

Therefore, we propose a novel benchmark dataset, **PACKaGE** (Paragraph-level Annotated Causal Knowledge and Graph-tExt) designed to bridge the gap between linguistic knowledge and causal reasoning in LLMs. First, it supports multi-hop causal extraction evaluation: models must identify not just pairwise relations but interconnected causal pathways. Second, it provides a testbed for round-trip fidelity, revealing whether information is preserved or degraded through modality conversion. Third, by offering three annotation levels, it allows controlled study of how structural metadata affects LLM generation—where we find, counterintuitively, that adding role labels hurts performance. When the causal relationships expressed in the text are structured as a graph, they form the causal graph shown in Figure 1, in JSON format.

- We construct the first bidirectional causal benchmark comprising paragraph-DAG pairs, where each paragraph is mapped to a DAG that captures various causal pathways.

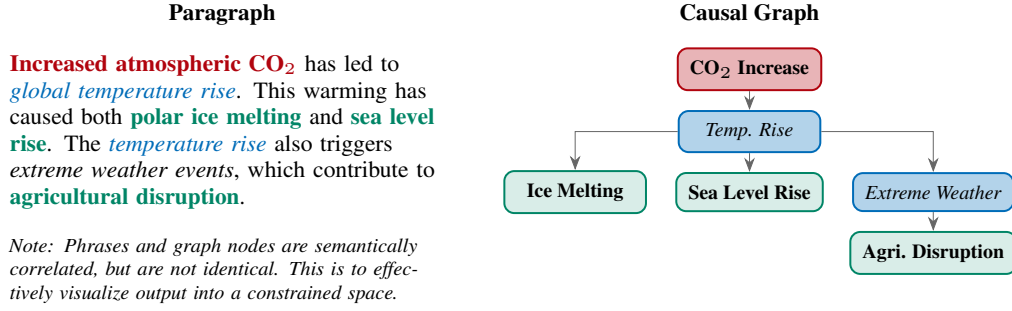


Figure 1: Example of Text–Graph Pairing. The paragraph (left) contains multiple causal claims that map to a causal graph (right). Underlined spans correspond to graph nodes; causal verbs define edges.

- We propose a four-stage pipeline that constructs paragraph-level causal graphs from fact-grounded texts.
- We evaluate LLMs on four bidirectional tasks and report several counterintuitive findings on model scale, tasks, and causal reasoning.

2 Related Works

2.1 Causal Graphs and LLMs

Recent studies emphasized the limitations of LLMs in causal reasoning or explored graph-based methods to inject external causal knowledge. For example, Jin et al. (2024) provided empirical evidence of LLMs’ inability to grasp causal relationships in specific scenarios. Approaches like Graph of Thoughts leveraged structured graph-based reasoning to improve LLMs’ access to relational information (Besta et al., 2024). Edge et al. (2024) proposed improving retrieval effectiveness via graph-based reasoning in RAG systems. Furthermore, rather than training models directly on causality, some research integrates causal discovery outputs into downstream components—such as retrievers in RAG (Wang et al., 2025) or uses external knowledge to guide reasoning (Lee et al., 2025; Kampani et al., 2024; Kazemi et al., 2025; Ban et al., 2025a).

2.2 Causality-Labeled Datasets

Prior work on constructing datasets that capture causal relationships has focused on the extraction of causal structures from text (Yang et al., 2022), increasingly leveraging LLMs (Jiralerspong et al., 2026; Ban et al., 2025b; Wang, 2024). However, the inverse direction—generating text from causal graphs and reconstructing those graphs remains underexplored, despite its importance for causal reasoning and the construction of knowledge graphs (Phatak et al., 2024).

Existing causality datasets span multiple granularity levels but share two structural limitations. At the word and sentence level, the Causal News Corpus (CNC) dataset provides binary causality labels (Tan et al., 2022) and RECESS offers fine-grained span annotations (Tan et al., 2023a). Tan et al. (2023b) unified annotation frameworks, while others target specific domains such as biomedicine (Pyysalo et al., 2007; Bansal et al., 2025), finance (Moreno-Sandoval et al., 2023), and commonsense reasoning (Mostafazadeh et al., 2020; Du et al., 2022). Current benchmarks assess inference capabilities regarding social media (Ding et al., 2025), and temporal relations (Caselli and Vossen, 2017). CauseNet (Heindorf et al., 2020) extracts causal relations using template-based methods. Relation extraction benchmarks (Zhang et al., 2017) and DocRED (Yao et al., 2019) provided architectural insights, despite not targeting causality.

Graph-to-text research (Gardent et al., 2017; Nan et al., 2021; Castro Ferreira et al., 2020; Colas et al., 2021) verbalizes knowledge graphs but does not model causal semantics. Phatak et al. (2024) explore LLM capabilities for narrating causal graphs and Wang et al. (2025) integrate causal graphs into retrieval systems, yet comprehensive evaluation of bidirectional causal graph and text remains limited.

Despite this breadth, all of these resources treat causality extraction as a *unidirectional* text-to-graph task and remain at sentence-level granularity, limiting their capacity for multi-hop causal evaluation (Table 1). As Table 1 shows, no exist-

Dataset	Level	Graph	Paired Text
CauseNet	Word	✓	✗
CNC & RECESS	Sentence	✗	✗
DocRED	Document	✓	✗
PACKaGE (Ours)	Paragraph	✓	✓

Table 1: Comparison with existing causal datasets

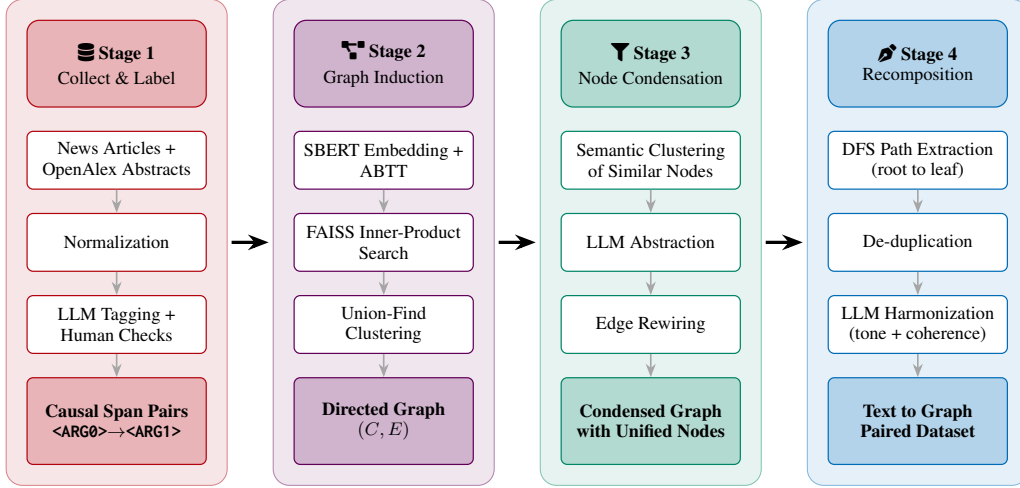


Figure 2: Dataset generation framework depicted in a four-stage pipeline.

ing resource pairs paragraph-level text with causal graphs in both directions. PACKaGE fills this gap by graph-text modeling of whether generated text preserves causal semantics and whether extracted graphs reflect textual causality.

3 PACKaGE

We define a causal relation as a tuple (C, E) , where a cause concept C (represented by noun phrases or clauses) exerts a potential influence on an effect concept E (Heindorf et al., 2020). To capture the complexity of real-world narratives beyond binary pairs, we classify nodes within a causal chain based on their topological roles: source nodes (Root) (Source causes with no incoming edges), Intermediate nodes (mediators acting as both effects and subsequent causes) and Leaf nodes (final effects with no outgoing edges) (Zhang et al., 2024).

The process for constructing the paired dataset spans four sequential stages (Stage 1: Section 3.1, Stage 2: Section 3.2, Stage 3: Section 3.3, Stage 4: Section 3.4) in Figure 2. Detailed depiction, including the prompts are provided in Appendix A and B.

3.1 Collecting and Labeling Causal Fragments

We start from a base corpus that mixes causality-labeled and fact-centric texts without labeling. Existing causality datasets are sentence-level and tend to be low in structural complexity, which limits their ability to support learning and evaluating richer causal reasoning. We acquire (i) sentence-level corpora with explicit causal labels and (ii) the fact-centric corpora, including news articles (CNC,

RECESS) and abstracts of research papers that describe factual events and relations. More specifically, we utilize 5,000 abstracts selected from OpenAlex (Priem et al., 2022), covering the period from 2014 to 2025. These abstracts consist of a balanced distribution across fields: Life Sciences (20%), Computing & Mathematics (20%), Social Sciences (20%), Physics (20%), and Others (20%).

We select abstracts because they are concise and comprehensive, containing core findings with minimal extraneous modifiers such as adjectives or adverbs (Dernoncourt and Lee, 2017; Wadden et al., 2020; Luan et al., 2018). This characteristic mitigates the risk of spurious correlations arising from irrelevant linguistic features (Kim et al., 2022). Also, these academic papers are characterized by objective, value-neutral writing that focuses on factual causation, which helps reduce biased guidance during training (Niiniluoto, 2009; van den Berg and Jeong, 2022).

Before the annotation, we apply a normalization to the texts: punctuation / whitespace unification (hyphen-to-slash collapse, repeated spaces), simple plural-to-singular normalization at the surface level (strip possessive “s”; drop terminal *s/es*), and number masking to a canonical token N. This step reduces spurious lexical variance while preserving semantic content (Kim et al., 2022).

3.2 Graph Induction

We construct a directed paragraph-level causal graph from sentence-extracted pairs and then condense nodes for downstream reasoning like Figure 4. From each sentence, we extract candidate causal pairs using span tags $\langle \text{ARG0} \rangle$ for cause,

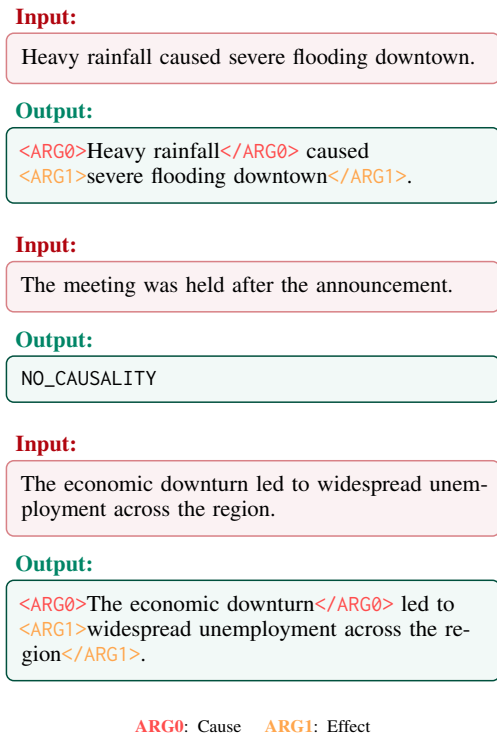


Figure 3: Examples of LLM-based causal span tagging: cause (ARG0) and effect (ARG1) spans, or returns NO_CAUSALITY for non-causal sentences.

<ARG1> for effect) (Tan et al., 2022, 2023a). Each pair yields two node surface strings and a directed edge, with the source sentence as edge provenance.

We unify near-duplicate node strings using dense retrieval and connectivity-based clustering (Gutiérrez et al., 2025). First, we encode every entity with Sentence-BERT and L_2 -normalize the embeddings to obtain cosine similarity (Reimers and Gurevych, 2019). We apply All-but-the-Top (ABTT) that removes the top d^* directions before re-normalizing (Mu et al., 2018). Because embeddings tend to concentrate along a few high-variance, corpus-specific axes, removing these directions yields a more isotropic space in which nearest neighbors reflect semantic proximity. (Mu et al., 2018). We leverage temperature scaling to calibrate similarity contrasts (Guo et al., 2017; Shih et al., 2023). Next, we use FAISS inner-product search to retrieve top- k neighbors per node (Johnson et al., 2021; Douze et al., 2026), then find semantic union to connect pairs exceeding similarity threshold.

3.3 Node Condensation

For components larger than a size threshold, we recompute exact cosine similarities and re-split by graph connectivity to prevent over-merging. Fi-

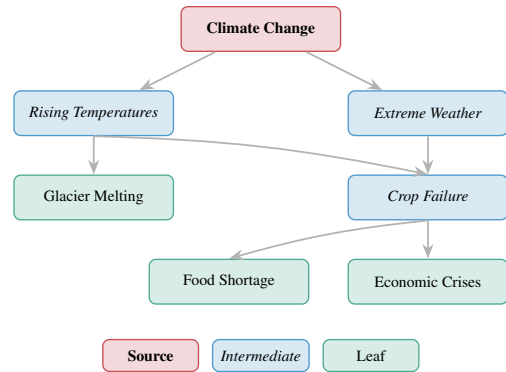


Figure 4: Example causal graph with node role classification. **Source** nodes (in-degree = 0) initiate causal chains, **Intermediate** nodes propagate effects, and **Leaf** nodes (out-degree = 0) represent terminal outcomes.

nally, each component receives a temporary representative—the shortest surface string—used only as a placeholder. For every component, we create an abstract representative label using an LLM that returns a compact set of keywords and a 1–3 sentence summary. LLM-based abstraction improves label compactness and readability for heterogeneous components (Maragheh et al., 2023; Chataut et al., 2024). The resulting keyword label becomes the final node for remapping edges: all edges touching any member of a component are rewired to the component’s abstract representative.

3.4 Path Extraction and Text Recomposition

Given the summarized graph, we identify root nodes with zero in-degree and enumerate all root-to-leaf paths using depth-first search while preventing cycles. We excluded cases where the cycles prevented the construction of a DAG. Along each traversed edge, we collect the source sentence(s) and concatenate them into a path-level text. To control redundancy in the texts, we apply progressively stricter de-duplication strategies, ranging from retaining the original text to removing repeated or similar sentences based on a SequenceMatcher similarity threshold (Nag et al., 2022). Each path-level text is then lightly edited by an LLM to harmonize tone and improve local coherence while preserving causal content.

We support three complementary output formats for downstream consumption: Plain text: the paragraphs as-is, for readability. Role-annotated causality: we attach node-role metadata derived from graph position: *source node* (zero in-degree), *intermediate node* (non-zero in- and out-degree), and *leaf node* (zero out-degree). Role-annotated causal-

Category	Count	% of Total	% of Errors
Nonsensical	295	5.89	66.29
Problematic	66	1.32	14.83
Artifact	61	1.22	13.71
Gibberish	23	0.46	5.17
Total Errors	445	8.89	100.00
Pass	4,566	91.12	—

Table 2: Distribution of error types identified in the pairing-level review ($N = 5,011$).

ity with relation attribution: we attach node–role metadata derived from graph topology and, for each role, we record where the cause originates and where the effect applies: *source node* (zero in-degree), *intermediate node* (non-zero in- and out-degree), and *leaf node* (zero out-degree). We attach a variety of causal roles by type because, when LLMs are trained with such annotations, the presence of tags can boost their performance on causality-related tasks (Wang et al., 2023; Chen et al., 2024; Wang et al., 2024). To increase dataset difficulty, we merged disjoint subgraphs into separate paragraphs for small graphs and applied additional smoothing.

Consistency with semantic causality. To avoid over-reliance on specific verbs (e.g., *cause*, *lead to*), our prompts instruct the LLM to detect semantically implied causality (e.g., nominalizations like “a consequence of”, discourse markers like “as a result”). Conversely, merely temporal (*after / before*), conditional/hypothetical (*if / might*), or associative (*associated with / correlated with*) statements are not tagged unless an asserted causal claim is present, reducing false positives from temporal or correlational language (Girju et al., 2007).

Outputs. We support three complementary output formats. The **Plain** text contains no additional structural information. The **Role**-added text augments the plain representation by annotating nodes with source, intermediate, and leaf, enabling the LLM to infer root-to-leaf paths. The **Role+Relationships** (Role+Rel.) text extends the role-added text by specifying which nodes are directly connected, for the model to capture both paths and relationships accurately.

3.5 Dataset Review

We describe our QC protocol as follows. Multiple authors manually audit the LLM

outputs—including causal/non-causal decisions, cause/effect span boundaries, and directionality—over multiple rounds, with cross-checking (a second author re-verifies a subset) and calibration meetings to resolve disagreements and refine guidelines/prompts. This weak supervision with LLM prompting plus human spot-checks provides a balance between scalability and precision, in line with programmatic labeling and LLM-assisted annotation (Ratner et al., 2017; Gilardi et al., 2023). Detailed case studies are in Section D.

First, we manually audit the LLM-generated labels—including causal/non-causal decisions, cause/effect span boundaries, and directionality—through multiple iterative rounds of cross-checking and calibration. Approximately 12,147 sentences are audited, with reviewer-specific correction rates of 14.3%, 16.3%, and 19.8%. We use error patterns identified during the audits to refine the tagging rules and prompts, and re-apply the updated annotation procedure in subsequent rounds, allowing these improvements to propagate across later tagging rounds. Second, we evaluate the text–graph pairs using a pairing-level review protocol with four error categories: problematic, artifact, nonsensical, and gibberish as shown in Table 2. **Problematic** refers to cases containing violent, inflammatory, or controversial content targeting specific individuals, countries, or groups. **Artifact** denotes cases in which natural semantic interpretation is hindered by ungrammatical expressions, meta-level phrasing, or missing contextual information. **Nonsensical** refers to cases where causal relationships are difficult to establish or where semantic coherence is weak due to excessive reliance on normative statements. **Gibberish** refers to cases in which graph nodes consist of demonstrative pronouns, or remain in clause-level or placeholder form, rendering them insufficiently functional as semantic units.

Among 5,011 reviewed pairs, 445 are identified as erroneous, yielding an error rate of 8.89% and a pass rate of 91.12%. The detected erroneous pairs are excluded at the pairing level. Nonsensical cases are the most frequent, followed by problematic, artifact, and gibberish. The error distribution reveals that quality degradation stems not from superficial grammatical & format issues and node-level representation issues, but rather from the instability of causal structures and the fragility of semantic coherence in LLM-generated outputs. This pattern is consistent with the broader finding that cur-

Metric	PACKaGE (Ours)	CNC	RECESS
Text Len. (chars)	234.61 (266.78)	84.59 (109.54)	119.01 (106.90)
Text Len. (tokens)	37.48 (45.65)	14.93 (19.29)	19.10 (17.84)
Num Sentences	3.73 (2.89)	1 (0.00)	1 (0.00)
Num Nodes	3.61 (1.96)	2.00 (0.00)	2.00 (0.00)
Num Edges	2.77 (2.84)	1.00 (0.00)	1.00 (0.00)
In-Deg / Node	1.29 (0.68)	1.00 (0.00)	1.00 (0.00)
Out-Deg / Node	1.35 (0.76)	1.00 (0.00)	1.00 (0.00)
Density	0.32 (0.16)	0.50 (0.00)	0.50 (0.00)
Source Nodes	1.73 (0.96)	0.00 (0.00)	0.00 (0.00)
Leaf Nodes	1.79 (1.01)	0.00 (0.00)	0.00 (0.00)
Complexity	6.44 (4.65)	0.00 (0.00)	0.00 (0.00)
Chain Motifs	0.18 (0.61)	–	–
Fork Motifs	1.33 (3.57)	–	–
Collider Motifs	1.20 (3.03)	–	–
Bi-fan Motifs	1.15 (3.18)	–	–
Long-chain Motifs	0.05 (0.74)	–	–

Table 3: Descriptive Statistics. Values are means, with standard deviations in parentheses.

rent LLMs lack robust causal reasoning capabilities (Jin et al., 2024), and suggests that paragraph-level causal datasets may serve as a useful benchmark.

4 Experimental Results

4.1 Statistical Summary of the Dataset

We construct a dataset comprising 4,566 graph-text pairs. Table 3 presents the statistics for this dataset, showing the mean and standard deviation per graph-text pair. We define Graph Density as $\frac{m}{n(n-1)}$, where n denotes the number of nodes and m denotes the number of edges. This metric indicates the degree of connectivity between nodes. Complexity is defined as the sum of the number of nodes and edges ($n + m$). CauseNet and DocRED are excluded due to the lack of paired text-graph annotations. While CNC and RECESS provide paired data, they exhibit lower complexity in both text length and graph structure than ours.

We focus on key motifs—recurrent subgraph patterns appearing more frequently than in the networks with identical degree distributions—specifically the chain, fork, collider, bi-fan, and long chain (Wang et al., 2022). These motifs are not merely structural patterns but carry distinct semantic meanings in causal reasoning: chains represent direct causal pathways, forks indicate common causes, colliders capture effect convergence, and bi-fans reflect complex causal dependencies, essential for understanding the causal structure of the graph (Zhang et al., 2021). The definitions of each motif are provided in Section E.

Furthermore, we exclude any instances containing bidirected edges or cycles during the data curation process. All graphs in our dataset strictly adhere to the properties of a DAG. This confirms

that our dataset captures complex causal relationships rather than the simple structures typically found in single isolated sentences.

4.2 Tasks & Metrics

Task 1: Text to Graph (T2G). The objective of this task is to evaluate whether the model can extract causal structures from texts. Given a text paragraph T from our dataset, the model is prompted to generate the corresponding causal graph G_{pred} .

Task 2: Graph to Text (G2T). This task assesses the model’s capability to synthesize descriptions from structured causal information. The model receives a causal graph G (linearized as a sequence of relations) as input and is tasked with generating a descriptive paragraph T_{pred} .

Tasks 3–4: Cyclic Graph–Text Reconstruction (G2T2G / T2G2T). Tasks 3 and 4 combine Task 1 and Task 2 to evaluate whether information is preserved through cyclic transformations between graphs and text. Task 3 performs graph-to-text-to-graph reconstruction ($G \rightarrow T' \rightarrow G_{pred}$), while Task 4 performs the reverse process: text-to-graph-to-text reconstruction ($T \rightarrow G' \rightarrow T_{pred}$).

Metrics for Tasks 1, 3. To account for structural errors (e.g., hallucinations) in LLM outputs, we report node- and edge-level Precision, Recall, and F1 scores instead of fixed-matrix metrics, comparing G_{pred} with G_{gt} (Wu et al., 2024). We report Graph Accuracy, defined as the percentage of predictions exactly matching ground-truth causal triplets.

Metrics for Tasks 2, 4. We evaluate generated text using ROUGE-L and Sentence-BERT cosine similarity (SBERT-cos) (Lin, 2004; Reimers and Gurevych, 2019). While ROUGE-L captures lexical overlap, SBERT-cos measures semantic similarity in embedding space; jointly, these metrics offer a complementary assessment of semantic preservation, even under paraphrasing.

4.3 Experimental Setup

In-context learning. We evaluate models using causal motifs as in-context learning such as motifs, leveraging an instruction-following LLM with task-specific prompts (Besta et al., 2024; Kazemi et al., 2025). For T2G, the objective is to produce a formatted JSON graph; for G2T, to verbalize nodes and directed edges; for G2T2G, to maximize round-trip fidelity. We use mixed supervision:

Type	Model	F1	Accuracy
Plain	GPT-5-mini	(0.937, 0.894)	(0.943, 0.895)
	Grok-4-fast	(0.924, 0.874)	(0.925, 0.872)
	Llama-3-8B	(0.591, 0.459)	(0.619, 0.467)
	OLMo-2-7B	(0.533, 0.141)	(0.545, 0.146)
	Qwen-2.5-7B	(0.670, 0.541)	(0.692, 0.539)
Role	GPT-5-mini	(0.829, 0.802)	(0.833, 0.813)
	Grok-4-fast	(0.925, 0.875)	(0.924, 0.871)
	Llama-3-8B	(0.589, 0.447)	(0.614, 0.451)
	OLMo-2-7B	(0.290, 0.068)	(0.296, 0.068)
	Qwen-2.5-7B	(0.305, 0.132)	(0.317, 0.133)
Role +Rel.	GPT-5-mini	(0.823, 0.816)	(0.847, 0.792)
	Grok-4-fast	(0.921, 0.889)	(0.932, 0.880)
	Llama-3-8B	(0.282, 0.125)	(0.298, 0.125)
	OLMo-2-7B	(0.495, 0.190)	(0.515, 0.195)
	Qwen-2.5-7B	(0.450, 0.252)	(0.463, 0.256)

Table 4: Numerical Results for the T2G task. Values in parentheses denote (node score, edge score).

ground-truth graphs for T2G, ground-truth texts for G2T, and teacher-forced round-trips for G2T2G.

Baseline models. We evaluate a range of LLMs across all tasks, including GPT-5, GPT-5-mini (OpenAI, 2025), Grok-4-fast reasoning (xAI, 2025), Llama-3-8B (Grattafiori et al., 2024), OLMo-2-7B (OLMo et al., 2025), Qwen-2.5-7B (Team et al., 2025). We use deterministic decoding settings, such as setting the temperature to zero. All texts are normalized (e.g., casefold, punctuation/whitespace unification) prior to node matching. To isolate causal reasoning ability from format compliance, we enforce JSON-only output via system prompts, apply rule-based post-processing (regex extraction, bracket matching) for non-conforming outputs. Prompts are provided in Section A and the more details of experiments are in Section C.

4.4 Baseline LLM Results

Tables 4 and 5 summarize the performance results for selected models and metrics across T2G and G2T. In Table 4, performance is measured separately for nodes and edges, with each cell reporting the scores in the format (node score, edge score). Due to space constraints, we omit Precision and Recall, some LLMs from the main table and the results of G2T2G and T2G2T; complete tables and ablation studies are provided in Section F.

Extraction and generation as distinct capabilities. We observe a divergent trend in model performance across the two tasks: GPT-5-mini and Grok-4-Fast excelled in T2G but struggled in G2T, in contrast to the performance of Qwen-2.5-7B and Llama-3-8B. This rank inversion indicates that identifying graph structure from text (extraction)

Type	Model	ROUGE-L	SBERT-cos
Plain	GPT-5-mini	0.256	0.766
	Grok-4-fast	0.209	0.738
	Llama-3-8B	0.292	0.787
	OLMo-2-7B	0.209	0.768
	Qwen-2.5-7B	0.335	0.817
Role	GPT-5-mini	0.240	0.669
	Grok-4-fast	0.195	0.618
	Llama-3-8B	0.262	0.675
	OLMo-2-7B	0.195	0.647
	Qwen-2.5-7B	0.291	0.676
Role +Rel.	GPT-5-mini	0.272	0.754
	Grok-4-fast	0.195	0.659
	Llama-3-8B	0.250	0.710
	OLMo-2-7B	0.212	0.688
	Qwen-2.5-7B	0.248	0.706

Table 5: Numerical results for the G2T task.

and verbalizing graphs into text (generation) are dissociated capabilities within current LLMs rather than a unified ability. This dissociation propagates into round-trip tasks. In T2G2T (extraction-first), GPT-5-mini leads by leveraging its strong T2G; in G2T2G (generation-first), OLMo-2-7B outperforms GPT-5-mini, whose weaker generation step produces intermediate text that is difficult to re-extract. To our knowledge, this extraction-generation dissociation has not been documented in prior work, because existing benchmarks evaluate only one direction. The bidirectional design of PACKaGE makes this finding possible, and it suggests that future work on causal LLMs should target extraction and generation as separate objectives.

Structural metadata. Despite providing information designed to facilitate causal verification (Liu et al., 2024), we observe a performance degradation across LLMs when generating texts annotated with causal roles or relations. This pattern suggests that marking nodes as sources, intermediates, or leaves introduces additional challenges for current LLMs, which struggle to integrate structural metadata into fluent natural language. Adding explicit relation information (Role+Rel.) partially mitigates this degradation. This indicates that relational context provides useful cues, helping models align generated text with the underlying causal structure. These results highlight both the promise and the current limitations of LLMs in verbalizing complex causal graphs, and motivate future research on structure-aware and causality-sensitive text generation methods (Xu et al., 2025).

Asymmetric weakness in G2T2G. Bidirectional transformations between text and graphs are crucial (Yamin et al., 2024); however, neither closed-source nor open-source models perform consistently across all tasks. For instance, both GPT-5-mini and Grok-4-Fast Reasoning exhibit performance degradation on the G2T2G task. As shown in Section F, their F1 scores under the plain-text setting are lower than those of open-source models. These results indicate that current LLMs still require further development to achieve robust and consistent performance across the bidirectional causal reasoning.

Scale and causal consistency. A notable asymmetry emerges across task types: closed-source models dominate in T2G, yet relatively smaller open-source models outperform them on G2T, G2T2G, and T2G2T under identical output templates. This performance gap persists even when both models are constrained by identical output templates and formats, suggesting that the disparity is not due to external differences in instruction or formats. These results indicate that beyond a certain model capacity, performance on structured causal transformation tasks may be influenced less by raw causal reasoning ability than by factors such as verbosity control and alignment with evaluation metrics.

For generation tasks, we hypothesize that larger models’ tendency toward semantically rich but stylistically diverse outputs (Hou et al., 2026) introduces lexical variation that penalizes metric-based evaluation. Larger LLMs tend to generate longer outputs, which generally hurts performance on G2T and T2G2T as shown in Figures 10 and 11. T2G is a causal extraction task from text to directed graphs, a well-studied task with abundant supervised data that closed-source models, trained on larger corpora, are more likely to have encountered. In round-trip tasks, this asymmetry compounds: T2G2T inherits high-quality intermediate graphs from T2G, mitigating length bias, whereas G2T2G propagates errors from poor intermediate texts, obscuring any consistent length-performance correlation. This observation underscores the need for improvements in LLMs, not only through scaling model capacity, but also by enhancing robustness to structural constraints in the causal tasks.

5 Discussion

Synthetic data. To construct a dataset grounded in real-world facts, we used fact-centric and scientific sources. Although synthetic causal sentences can be produced by fixing causal verbs (e.g., “cause”) and substituting lexical items, this often breaks factual coherence and introduces hallucinations. To minimize this risk, we construct our dataset from fact-centric and scientific sources. This design choice prioritizes factual grounding and reflects a conservative approach to data quality.

Implicit knowledge in sentence ordering. To mitigate positional bias, we randomize sentence order within paragraphs. We observed that LLMs often confused causal direction when effects appeared before causes in text, suggesting they had learned a linguistic bias where causes precede effects both temporally and textually as detailed in Section D. By randomizing sentence placement, we aimed to reduce reliance on such heuristics and encourage models to reason about causality itself rather than sentence sequence.

In-context learning. Few-shot examples consisting of causal motifs and their descriptions are provided in either graph or text. While this in-context learning aids in graph construction, its impact on text reconstruction is relatively limited. This is because knowing the type of motif primarily contributes to understanding the forms of frequently occurring causal structures, which is fundamentally different from identifying causal relationships in new sentences containing different factual content (Barjašić et al., 2021). Since motifs represent abstract structural categories rather than providing specific causal knowledge between nodes, their utility in terms of text reconstruction is limited. Nevertheless, these motifs still provide useful guidance in the process of graph construction.

Intra-sentential tagging. We use intra-sentential causal extraction as the atomic unit of PACKaGE and construct paragraph-level causal structures compositionally across subsequent stages. In Stage 1, we require both “<ARG0>” and “<ARG1>” to occur within a single sentence and extract one primary causal pair per sentence (Appendix A). We assemble these local causal pairs into inter-sentential structures through Stages 2–4: we merge semantically equivalent entities across sentences using SBERT, abstract related nodes and rewire their edges using an LLM, and extract multi-hop

paths with DFS to form path-aligned paragraphs. Through this composition, we recover causal structures that span multiple source sentences whenever their local relations are connected through shared or semantically merged nodes, including chains, forks, colliders, bi-fans, and longer causal paths. We favor this compositional strategy over direct cross-sentence causal annotation because inter-sentential relations often depend on discourse markers and anaphoric references that are difficult to resolve reliably in context. As discussed in Appendix D, we observed that such context dependence is closely related to unresolved references and coherence failures, which also constitute the dominant error types in our quality-control analysis. We therefore use sentence-level extraction to obtain more controlled local causal units and rely on graph-level composition to build paragraph-level structures while reducing annotation uncertainty from direct cross-sentence tagging.

6 Conclusion

Despite recent advances, LLMs still exhibit limitations in causal reasoning, motivating efforts to incorporate external knowledge and graph structures. We introduce a paired text–graph dataset that enables joint learning of causal language and corresponding causal graphs, extending beyond sentence-level and low-complexity. By linking causal annotations with academic papers across diverse domains, we construct higher-complexity texts and graphs through semantic node abstraction while preserving causal edges. This process resulted in a dataset that not only increases the complexity of both text and graphs but also facilitates bidirectional reasoning between textual and graph representations. We evaluate models’ understanding across these dual modalities, and our experiments show that current LLMs struggle. We hope that our dataset provides a foundation for integrating causal knowledge into LLMs.

Limitations

Evaluating causality in text remains an open problem, as there is no widely accepted metric that can faithfully capture semantic causal relations. Although heuristic identification of causes and effects is possible, such methods do not provide a principled evaluation of causality. We therefore use SBERT cosine similarity as a supplementary measure for text evaluation, while acknowledging its

limitations. This design choice is motivated by the absence of established alternatives and should be understood as a limitation of current evaluation methodologies rather than of our proposed framework. Also, we did not evaluate all tasks using the most computationally intensive reasoning-focused models. Nevertheless, we expect these models to exhibit trends similar to those observed in other tasks. In addition, since prompt engineering and in-context learning are beyond the scope of this work, we do not conduct extensive ablation studies targeting these aspects. We expect a more thorough investigation of prompt design and in-context learning strategies to be pursued in future work.

Ethical considerations

We focused on constructing the dataset in a manner that is as neutral and fact-centric as possible, to improve the ethical quality of the dataset. This is because our goal is to convey to LLMs only culturally and value-neutral causal relationships, thereby avoiding biased guidance. Furthermore, we use data that excludes personal information, such as paper abstracts, thereby eliminating the risk of privacy leakage. We also reflected these principles during the quality control process. Furthermore, due to these characteristics, the dataset can be considered to pose a relatively low risk of potential harmful outcomes in downstream applications. As discussed in Section 5, synthetically generated datasets may introduce illusory or spurious causal relationships; in contrast, our dataset is relatively free from such issues. AI models were used solely as an auxiliary tool for literature search in Section 2.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MND) (No. RS-2026-25551943, Military-Industry-Academia Cooperation Center (Seoul (Yongsan)))

References

- Anthropic. 2024. Claude. <https://www.anthropic.com>.
- Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, and Huanhuan Chen. 2025a. *Llm-driven causal discovery via harmonized prior*. *IEEE*

- Transactions on Knowledge and Data Engineering*, 37(4):1943–1960.
- Taiyu Ban, Lyuzhou Chen, Derui Lyu, Xiangyu Wang, Qinrui Zhu, Qiang Tu, and Huanhuan Chen. 2025b. [Integrating Large Language Model for Improved Causal Discovery](#). *IEEE Transactions on Artificial Intelligence*, 6(11):3030–3042.
- Nency Bansal, R C Sri Dhinesh, Ayush Pathak, and Manikandan Narayanan. 2025. [Beyond associations: A benchmark causal relation extraction dataset \(cred\) of disease-causing genes, its comparative evaluation, interpretation and application](#). *bioRxiv*.
- Irena Barjašić, Hrvoje Štefančić, Vedrana Pribičević, and Vinko Zlatić. 2021. [Causal motifs and existence of endogenous cascades in directed networks with application to company defaults](#). *Preprint*, arXiv:2011.08148.
- Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michał Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. [Graph of thoughts: solving elaborate problems with large language models](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’24/IAAI’24/EAAI’24. AAAI Press.
- 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.
- Thiago Castro Ferreira, Claire Gardent, Nikolai Ilinykh, Chris van der Lee, Simon Mille, Diego Moussallem, and Anastasia Shimorina. 2020. [The 2020 bilingual, bi-directional WebNLG+ shared task: Overview and evaluation results \(WebNLG+ 2020\)](#). In *Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+)*, pages 55–76, Dublin, Ireland (Virtual). Association for Computational Linguistics.
- Sandeep Chataut, Tuyen Do, Bichar Dip Shrestha Gurung, Shiva Aryal, Anup Khanal, Carol Lushbough, and Etienne Gnimpieba. 2024. [Comparative study of domain driven terms extraction using large language models](#). *Preprint*, arXiv:2404.02330.
- Zhenbin Chen, Zhixin Li, Yufei Zeng, Canlong Zhang, and Huifang Ma. 2024. [Gap: A novel generative context-aware prompt-tuning method for relation extraction](#). *Expert Syst. Appl.*, 248(C).
- Anthony Colas, Ali Sadeghian, Yue Wang, and Daisy Zhe Wang. 2021. Eventnarrative: A large-scale event-centric dataset for knowledge graph-to-text generation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. [Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities](#). *Preprint*, arXiv:2507.06261.
- Franck Dernoncourt and Ji Young Lee. 2017. [PubMed 200k RCT: a dataset for sequential sentence classification in medical abstracts](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 308–313, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Xiaohan Ding, Kaike Ping, Buse Çarık, and Eugenia Rho. 2025. [A multi-level benchmark for causal language understanding in social media discourse](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28776–28790, Suzhou, China. Association for Computational Linguistics.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2026. [The faiss library](#). *IEEE Transactions on Big Data*, 12(2):346–361.
- 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.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva N. Mody, Steven Truitt, and Jonathan Larson. 2024. [From local to global: A graph rag approach to query-focused summarization](#). *ArXiv*, abs/2404.16130.
- Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. [The WebNLG challenge: Generating text from RDF data](#). In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. [Chatgpt outperforms crowd workers for text-annotation tasks](#). *Proceedings of the National Academy of Sciences of the United States of America*, 120.
- 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.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML’17*, page 1321–1330. JMLR.org.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From rag to memory: non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning, ICML’25*. JMLR.org.
- Stefan Heindorf, Yan Scholten, Henning Wachsmuth, Axel-Cyrille Ngonga Ngomo, and Martin Potthast. 2020. [Causenet: Towards a causality graph extracted from the web](#). In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20*, page 3023–3030, New York, NY, USA. Association for Computing Machinery.
- Zhaoyi Joey Hou, Bowei Alvin Zhang, Yining Lu, Bhiman Kumar Baghel, Anneliese Brei, Ximing Lu, Meng Jiang, Faeze Brahman, Snigdha Chaturvedi, Haw-Shiuan Chang, Daniel Khashabi, and Xiang Lorraine Li. 2026. [Creativityprism: A cross-domain evaluation framework for large language model creativity](#). *Preprint*, arXiv:2510.20091.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: assessing causal reasoning in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA. Curran Associates Inc.
- Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2024. [Can large language models infer causation from correlation?](#) *Preprint*, arXiv:2306.05836.
- Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. 2026. [Efficient causal graph discovery using large language models](#). *Preprint*, arXiv:2402.01207.
- Jeff Johnson, Matthijs Douze, and Herve Jegou. 2021. [Billion-Scale Similarity Search with GPUs](#). *IEEE Transactions on Big Data*, 7(03):535–547.
- Shiv Kampani, David Hidary, Constantijn van der Poel, Martin Ganahl, and Brenda Miao. 2024. [Llm-initialized differentiable causal discovery](#). *Preprint*, arXiv:2410.21141.
- Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V Le, and Orhan Firat. 2025. [BIG-bench extra hard](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26473–26501, Vienna, Austria. Association for Computational Linguistics.
- Juhyeon Kim, Yesong Choe, and Sanghack Lee. 2022. [SNU-causality lab @ causal news corpus 2022: Detecting causality by data augmentation via part-of-speech tagging](#). In *Proceedings of the 5th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE)*, pages 44–49, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Chanhui Lee, Juhyeon Kim, Yongjun Jeong, Yoon-seok Yeom, Juhyun Lyu, Junghee Kim, Sangmin Lee, Sangjun Han, Hyeokjun Choe, Soyeon Park, Woohyung Lim, Sungbin Lim, and Sanghack Lee. 2025. [On incorporating prior knowledge extracted from large language models into causal discovery](#). *IEEE Access*, 13:194691–194713.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. 2022. Solving quantitative reasoning problems with language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS ’22*, Red Hook, NY, USA. Curran Associates Inc.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Cheng Liu, Wei Xiang, and Bang Wang. 2024. [Identifying while learning for document event causality identification](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3815–3827, Bangkok, Thailand. Association for Computational Linguistics.
- Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. [Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

- Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.
- Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. [Llm-take: Theme-aware keyword extraction using large language models](#). In *2023 IEEE International Conference on Big Data (BigData)*, pages 4318–4324.
- R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. 2002. [Network motifs: Simple building blocks of complex networks](#). *Science*, 298(5594):824–827.
- Antonio Moreno-Sandoval, Jordi Porta-Zamorano, Blanca Carbajo-Coronado, Doaa Samy, Dominique Mariko, and Mahmoud El-Haj. 2023. [The financial document causality detection shared task \(fincausal 2023\)](#). In *2023 IEEE International Conference on Big Data (BigData)*, pages 2855–2860.
- 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.
- Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2018. [All-but-the-top: Simple and effective post-processing for word representations](#). *Preprint*, arXiv:1702.01417.
- M. Sree Ram Kiran Nag, Gorla Srinivas, Kotagiri Venkata Rao, Sairam Vakkalanka, and S. Nagen-dram. 2022. [An efficient procedure for identifying the similarity between french and english languages with sequence matcher technique](#). *Advances in Data Science and Management*.
- Linyong Nan, Dragomir Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xian-gru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Fiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, and 5 others. 2021. [DART: Open-domain structured data record to text generation](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 432–447, Online. Association for Computational Linguistics.
- Ilkka Niiniluoto. 2009. Facts and values—a useful distinction. *Pragmatist perspectives*, pages 109–133.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. [Capabilities of gpt-4 on medical challenge problems](#). *Preprint*, arXiv:2303.13375.
- Ohio Supercomputer Center. 1987. [Ohio supercomputer center](#).
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groen-eveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 24 others. 2025. [2 olmo 2 furious](#). *Preprint*, arXiv:2501.00656.
- OpenAI. 2025. GPT-5 and GPT-5 Mini. <https://openai.com>.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Atharva Phatak, Vijay K. Mago, Ameeta Agrawal, Aravind Inbasekaran, and Philippe J. Giabbanelli. 2024. [Narrating causal graphs with large language models](#). *Preprint*, arXiv:2403.07118.
- Jason Priem, Heather Piwowar, and Richard Orr. 2022. [Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts](#). *Preprint*, arXiv:2205.01833.
- 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 Bioinformatics*, 8:50 – 50.
- Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. [Snorkel: rapid training data creation with weak supervision](#). *Proc. VLDB Endow.*, 11(3):269–282.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Andy Shih, Dorsa Sadigh, and Stefano Ermon. 2023. Long horizon temperature scaling. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Agüera y Arcas, Dale Webster, and 11 others. 2022. [Large language models encode clinical knowledge](#). *Preprint*, arXiv:2212.13138.

- Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See-Kiong Ng. 2023a. **RECESS: Resource for extracting cause, effect, and signal spans**. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–82, Nusa Dua, Bali. Association for Computational Linguistics.
- Fiona Anting Tan, Ali Hürriyetoğlu, Tommaso Caselli, Nelleke Oostdijk, Tadashi Nomoto, Hansi Hettiarachchi, Iqra Ameer, Onur Uca, Farhana Ferdousi Liza, and Tiancheng Hu. 2022. **The causal news corpus: Annotating causal relations in event sentences from news**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2298–2310, Marseille, France. European Language Resources Association.
- Fiona Anting Tan, Xinyu Zuo, and See-Kiong Ng. 2023b. **Unicausal: Unified benchmark and repository for causal text mining**. In *Big Data Analytics and Knowledge Discovery - 25th International Conference, DaWaK 2023, Penang, Malaysia, August 28-30, 2023, Proceedings*, volume 14148 of *Lecture Notes in Computer Science*, pages 248–262. Springer.
- Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. **Qwen2.5 technical report**. *Preprint*, arXiv:2412.15115.
- Axel van den Berg and Tay Jeong. 2022. Cutting off the branch on which we are sitting? on postpositivism, value neutrality, and the “bias paradox”. *Society*, 59(6):631–647.
- David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. **Fact or fiction: Verifying scientific claims**. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.
- Bin Wang, LvHang Cheng, JinFang Sheng, ZhengAng Hou, and YaoXing Chang. 2022. Graph convolutional networks fusing motif-structure information. *Scientific Reports*, 12(1):10735.
- Kai Wang, Yanping Chen, Kunjian Wen, Chao Wei, Bo Dong, Qinghua Zheng, and Yongbin Qin. 2023. **Cue prompt adapting model for relation extraction**. *Connection Science*, 35(1):2161478.
- Nengbo Wang, Xiaotian Han, Jagdip Singh, Jing Ma, and Vipin Chaudhary. 2025. **CausalRAG: Integrating causal graphs into retrieval-augmented generation**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22680–22693, Vienna, Austria. Association for Computational Linguistics.
- Zeyu Wang. 2024. **CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models**. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.
- Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024. **Document-level causal relation extraction with knowledge-guided binary question answering**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.
- Xixi Wu, Yifei Shen, Caihua Shan, Kaitao Song, Siwei Wang, Bohang Zhang, Jiarui Feng, Hong Cheng, Wei Chen, Yun Xiong, and Dongsheng Li. 2024. Can graph learning improve planning in llm-based agents? In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS ’24*, Red Hook, NY, USA. Curran Associates Inc.
- xAI. 2025. Grok-4: Fast reasoning language model. <https://x.ai>.
- Haotian Xu, Yuning You, and Tengfei Ma. 2025. **When structure doesn’t help: LLMs do not read text-attributed graphs as effectively as we expected**. In *The Fourth Learning on Graphs Conference*.
- Khurram Yamin, Shantanu Gupta, Gaurav R. Ghosal, Zachary C. Lipton, and Bryan Wilder. 2024. **Failure modes of llms for causal reasoning on narratives**. *Preprint*, arXiv:2410.23884.
- 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.
- Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. **DocRED: A large-scale document-level relation extraction dataset**. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.
- Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. **Do large language models know what they don’t know?** In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.
- Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. **A survey on causal discovery: Theory and practice**. *Int. J. Approx. Reasoning*, 151(C):101–129.
- Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. **Causal parrots: Large language models may talk causality but are not causal**. *Preprint*, arXiv:2308.13067.

Xiaoge Zhang, Xiao-Lin Wang, Fenglei Fan, Yiu-Ming Cheung, and Indranil Bose. 2024. [Enhancing the performance of neural networks through causal discovery and integration of domain knowledge](#). *Preprint*, arXiv:2311.17303.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. [Position-aware attention and supervised data improve slot filling](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Chee-Kong Lee. 2021. Motif-based graph self-supervised learning for molecular property prediction. In *Proceedings of the 35th International Conference on Neural Information Processing Systems, NIPS '21*, Red Hook, NY, USA. Curran Associates Inc.

A Prompts

This section details the prompts employed for constructing the dataset and conducting experiments for the newly designed task.

A.1 Prompts for Making the Dataset

First, to tag causal relationships using the LLM for sentences where the presence of causality was initially undetermined, we used the following prompt:

System prompt
You are a precise causal expert.
Follow the instructions exactly. Output must be either:
1) NO_CAUSALITY
2) The ORIGINAL sentence with exactly one pair of tags inserted:
<ARG0>...</ARG0> for the cause, <ARG1>...</ARG1> for the effect.
Do NOT change any other characters or order.
PLAIN TEXT ONLY (no quotes, no code blocks).

Instructions:
- If NO causal relationship is expressed, output exactly: NO_CAUSALITY
- If a causal relationship IS expressed, output the ORIGINAL sentence but insert:
- <ARG0>...</ARG0> around the minimal cause span
- <ARG1>...</ARG1> around the minimal effect span
- Output ONE primary cause-effect pair only (choose the most central relation if multiple exist).
- Do NOT change any other characters, spacing, or word order besides inserting the tags.
- PLAIN TEXT ONLY (no quotes, no code blocks).

Causality detection policy:
- Cover explicit causal markers such as "cause(s)/caused", "lead(s) to/led to", "result(s) in/resulted in", "because (of)", "due to",

"owing to", "as a result of", "therefore/thus/hence".
- ALSO consider semantically implied causality even without those markers (nominalizations, appositives, causal nouns like "reason/result/consequence/impact/effect", paraphrases, or cross-lingual cues).
- Do NOT tag if it is merely temporal/conditional/associative without asserted causality:
temporal: "after/before/when/during";
conditional/hypothetical: "if/might/could/may";
associative: "associated with/correlated with/linked to".
- If the sentence already contains <ARG0> or <ARG1> tags, do not create nested/duplicate tags; return NO_CAUSALITY in that case.

Input: The protest was invoked due to low wages of taxi drivers.

Output: <ARG1>The protest was invoked</ARG1> due to <ARG0>low wages of taxi drivers</ARG0>.

Input: Heavy rainfall caused severe flooding downtown.

Output: <ARG0>Heavy rainfall</ARG0> caused <ARG1>severe flooding downtown</ARG1>.

Input: The residents of the streets set the house on fire in protest of the poor delivery of service.

Output: <ARG1>The residents of the streets set the house on fire</ARG1> in <ARG0>protest of the poor delivery of service</ARG0>.

Input: There were five people who died in the first blast and over 50 were injured, several critically.

Output: <ARG1>There were five people who died</ARG1> in <ARG0>the first blast</ARG0> and over 50 were injured, several critically.

For entities identified as similar based on the FAISS results, the following prompt was utilized to condense them using the LLM:

You are a precise scientific summarizer.
Return:
Line1: ultra-short title (<=3 words, no punctuation)
Line2: one concise sentence capturing the shared meaning.

SURFACE FORMS:
{text_block}

The following prompt was applied for polishing scientific text paragraphs to read naturally and cohesively while strictly preserving causal meaning and sentence structure.

You are a careful scientific editor.
Smooth the paragraph so it reads naturally and cohesively.

- Preserve causal meaning precisely.
- Do NOT compress multiple sentences into one; keep sentence boundaries.
- Remove redundancy if any; vary diction lightly.
- Pay strict attention to correct article usage and capitalization.
- Contextual Addition: If sentences appear unrelated or lack context, add 1-3 explanatory sentences to bridge the gap. However, do NOT add sentences if they are unnecessary.

Return only the improved paragraph.
{paragraph}

A.2 Annotation Convention

Our prompt follows the following tagging rules:

- <ARG0>: Marks the cause
- <ARG1>: Marks the effect
- NO_CAUSALITY: Indicates no causal relationship exists

A.3 In-context Learning

The system prompt used across our tasks is shown below, with only the task specification modified accordingly. The example corresponds to G2T task.

SYSTEM RULES:

- Task: Graph-to-Text. Write a paragraph from the graph.
- Always return outputs in the exact format requested by the task.
- For graphs, output STRICT JSON with keys: nodes, edges. No comments, no trailing commas.
- Node ids must come from the input text; do not invent new entities.
- If uncertain, omit rather than hallucinate.
- Be concise and consistent.

We design separate instruction prompts for T2G and G2T settings, and alternate between these prompts for the cyclic tasks (G2T2G and T2G2T).

T2G_PROMPT = """You are a precise information extraction model.
From the INPUT TEXT, extract a directed causal graph as strict JSON.

Rules:

- Return ONLY valid JSON, no prose.
- JSON schema:

```

{{
  "nodes": [{{"id": "..."}}, ...],
  "edges": [{{"source": "...", "target": "..."}}, ...]
}}

```

- Keep node ids as short natural phrases found in the text.

- Do NOT invent nodes/edges that are not supported.

INPUT TEXT:
{context}

G2T_PROMPT = """You are a scientific writer.
Given a directed causal graph (JSON), write a short paragraph (3-6 sentences) that naturally expresses the causal relationships. Keep facts faithful.
Return ONLY the paragraph, no extra commentary.

GRAPH JSON:
{graph_json}
"""

The few-shot examples used for in-context learning are summarized below, where each example is annotated with its corresponding motif type. For Tasks 1 and 3, which require text reconstruction, we provide text-based examples, whereas for Tasks 2 and 4, which involve graph reconstruction, we supply examples in JSON graph format.

```

{"input": "EXAMPLE_TYPE: long_chain\nnGRAPH_JSON: {\"nodes\": [\"Fog\", \"Moisture\", \"Wind\", \"Weathering\", \"Cementation\"], \"edges\": [[\"Fog\", \"Moisture\"], [\"Moisture\", \"Wind\"], [\"Wind\", \"Weathering\"], [\"Weathering\", \"Cementation\"]]}\", \"output\": \"Wind triggered weathering. Moreover, moisture led to wind. Additionally, weathering brought about cementation. Separately, fog caused moisture.\"}

{"input": "EXAMPLE_TYPE: fork\nnGRAPH_JSON: {\"nodes\": [\"Body\", \"Heat\", \"Current\", \"Perturbation\", \"Instability\"], \"edges\": [[\"Body\", \"Heat\"], [\"Heat\", \"Current\"], [\"Current\", \"Perturbation\"], [\"Perturbation\", \"Instability\"]]}\", \"output\": \"Current led to Perturbation. Also, Heat caused Current. In addition, Perturbation triggered Instability. Separately, Body resulted in Heat.\"}

{"input": "EXAMPLE_TYPE: collider\nnGRAPH_JSON: {\"nodes\": [{\"id\": \"Acid aminobutyric\"}, {\"id\": \"Acid arachidonic\"}, {\"id\": \"Brain gaba\"}, {\"id\": \"Defense eicosanoid\"}, {\"id\": \"Gaba\"}, {\"id\": \"Gaba metabolic\"}, {\"id\": \"pathophysiology of anxiety disorder\"}], \"edges\": [{\"source\": \"Acid aminobutyric\", \"target\": \"pathophysiology of anxiety disorder\"}, {\"source\": \"Acid arachidonic\", \"target\": \"Defense eicosanoid\"}, {\"source\": \"Gaba\", \"target\": \"pathophysiology of anxiety disorder\"}, {\"source\": \"Gaba metabolic\", \"target\": \"Brain gaba\"}]}\", \"output\": \"Aminobutyric acid brought about the pathophysiology of anxiety disorder. Arachidonic acid gave rise to an eicosanoid involved in defense. GABA caused the

```



```
is found guilty\"}, {\"source\": \"Died attempting\", \"target\": \"he is guilty\"}, {\"source\": \"attempting to cause death of two policemen and death of inspector\", \"target\": \"he is guilty\"}]\"}
```

B Dataset Details

In this section, we describe how the PACKaGE dataset was constructed and detail the hyperparameters used.

B.1 Quality Control

Each stage in the pipeline incorporates validation to ensure data quality. In the summarization stage, we verify that the summary captures essential meaning without information loss. During the refinement stage, validation should focus on whether sentence boundaries are correctly maintained and whether the underlying causal semantics remain intact. Finally, in the tagging stage, the tagging process is validated to guarantee that tags are properly paired and minimally span the intended cause-effect relations.

B.2 Dataset Setting Details

Graph induction and node condensation. We construct a directed paragraph-level causal graph from sentence-extracted pairs and then condense nodes for downstream reasoning. From each sentence, we extract candidate causal pairs using span tags (<ARG0> for cause, <ARG1> for effect) (Tan et al., 2022, 2023a). Each pair yields two node surface strings and a directed edge, along with the source sentence as edge provenance.

We unify near-duplicate or closely related node strings using dense retrieval and connectivity-based clustering (Gutiérrez et al., 2025). First, we encode every entity with Sentence-BERT¹ and L_2 -normalize the embeddings to compute cosine similarity (Reimers and Gurevych, 2019). We apply All-but-the-Top (ABTT), a principal-component suppression step that projects out the top d directions before re-normalizing (Mu et al., 2018). Since these directions carry corpus-level rather than semantic variation, their removal makes nearest-neighbor structure reflect semantic proximity. We set $d^* = 2$, which suppresses these dominant components while avoiding excessive removal of semantic information. We perform temperature scaling (set = 1) by dividing vectors to sharpen or smooth similarity contrasts (Guo et al., 2017;

¹Hugging Face sentence-transformers/all-MiniLM-L6-v2

Shih et al., 2023). Next, we run FAISS inner-product search to retrieve the top-50 neighbors per node, which scales to large inventories while maintaining high recall (Johnson et al., 2021; Douze et al., 2026). Thresholded union-find connect pairs whose similarity exceeds 0.84 and form components with union-find. Refinement of large components for components larger than a size threshold, recompute exact cosine similarities and split by graph connectivity under 40, preventing over-merging. For each component, pick the shortest surface string as a temporary representative (only as a placeholder for the next step). The hyperparameters were selected based on ablation results, with the final dataset constructed using the configuration that the authors judged to yield the most suitable textual outputs.

Hyperparameters. Unless noted, we use GPT-5 for annotating causality and all-MiniLM-L6-v2 as the SBERT encoder, ABTT dimension $d^* = 2$, FAISS top- $k = 50$, similarity threshold $\tau = 0.84$, refinement trigger at component size ≥ 40 , and *consecutive* de-duplication for path narratives. Cluster abstraction is driven by an LLM with a TF-IDF fallback. All string normalizations (punctuation/whitespace unification, surface plural-to-singular rules, number masking) are applied prior to embedding and clustering.

License. OpenAlex data is openly available under the Creative Commons Attribution 4.0 license. The Causal News Corpus (CNC) dataset used in this work is publicly released under the CC0 1.0 Universal (CC0-1.0) license.

C Experimental Details

This section describes the hyperparameters and settings used for task execution by LLMs, as well as the models employed in our experiments.

C.1 Evaluation Formulas

For Tasks 1 and 3, we compute the following metrics for nodes and edges:

Precision:

$$\text{Precision} = \frac{|\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted}|} \quad (1)$$

Recall:

$$\text{Recall} = \frac{|\text{Predicted} \cap \text{Ground Truth}|}{|\text{Ground Truth}|} \quad (2)$$

F1 Score:

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Accuracy:

$$\text{Accuracy} = \frac{|\text{Predicted} \cap \text{Ground Truth}|}{|\text{Predicted} \cup \text{Ground Truth}|} \quad (4)$$

These metrics are computed separately for nodes and edges, providing a comprehensive assessment of the model’s performance in capturing causal structure.

C.2 Implementation Notes and Tips

Output determinism. For T2G, enforce JSON-only output via system prompt and constrained decoding; reject generations that contain trailing prose. For G2T, keeping node surface forms intact eases round-trip matching.

Few-shot selection. Prefer short, unambiguous pairs that cover varied phrasing (*cause/lead to/result in, because/due to/as a result of*) and nominalized patterns (*consequence of, impact on*) to improve robustness.

In-context learning. We train with instruction-style (prompt, target) pairs: sentencetograph for T2G, graphtotext for G2T, and chained supervision for G2T2G. We mix in rejection sampling to drop malformed JSON in T2G and hallucinated nodes in G2T.

Baseline models. The baseline models used in our experiments include GPT-5, GPT-5-mini (OpenAI, 2025), Grok-4-fast Reasoning (xAI, 2025), Llama-3-8B (Grattafiori et al., 2024), OLMo-2-7B (OLMo et al., 2025), Qwen-2.5-7B (Team et al., 2025), Claude Sonnet 4.5, Haiku 4.5 (Anthropic, 2024), and Gemini 2.5 flash (Comanici et al., 2025). However, in the case of Gemma3-4B, the model failed to adequately understand the task—likely due to its relatively small parameter size—and consequently produced unreliable outputs. As a result, we do not report its results separately. It can be viewed as indicative of the challenging nature of our dataset. For reproducibility, we set the temperature to zero and enforced deterministic execution on CUDA whenever possible, using a fixed random seed of 42. When asynchronous execution was enabled, the experiments required approximately 18 hours on average to complete. Experiments were primarily conducted on one NVIDIA H100 GPU

and four NVIDIA RTX 3090 GPUs hosted on the OSC cluster (Ohio Supercomputer Center, 1987). All texts are normalized (e.g., casefold, punctuation/whitespace unification, light plural-to-singular, number masking) prior to node matching.

Format robustness. To ensure that our evaluation measures causal reasoning rather than mere format compliance, we controlled for formatting artifacts (e.g., JSON parsing failures or the inclusion of unnecessary prose) through three strategies. First, the system prompt explicitly specified the required output format and repeatedly enforced a JSON-only constraint. Second, when the output did not strictly follow the required format, we applied rule-based post-processing (e.g., regex extraction and bracket matching) to recover the substantive graph or text content before performing the evaluation. Third, during node matching, we complemented exact matching with casefold normalization and cosine-similarity-based soft matching so that semantically equivalent expressions (e.g., “temperature rise” vs. “rising temperatures”) were treated as valid matches. Therefore, the performance differences reported in this paper reflect differences in causal reasoning capability rather than differences in format compliance.

D Case Study

In this section, we present a case study of errors observed during quality control of LLM-based causal labeling, as well as errors encountered while performing the proposed tasks.

D.1 Causal Annotation

To analyze errors made by LLMs during causal labeling, we identify several representative failure cases observed during annotation. We discuss four representative cases below.

D.1.1 Case 1: Purpose-means relations

LLMs often misclassified purpose-means relations as causal relations. For example:

- “*The former state penitentiary in Alton was used during the Civil War to hold up to 12,000 Confederate prisoners of war.*”
- “*Kaibō-kan were designed for a similar role in the Imperial Japanese Navy.*”

In these cases, the expressions describe functional purposes (“used to hold,” “designed for”) rather

than causal mechanisms where one event brings about another.

D.1.2 Case 2: Sequential relations

Temporal or naming sequences were incorrectly tagged as causal relationships, particularly in “named after” constructions:

- *“The club plays its home matches at the Asim Ferhatović Hase Stadium, named after legendary club striker Asim Ferhatović.”*
- *“The town is named for Serpentine Creek, a tributary of the Loddon River.”*
- *“The school was so called after Cyrene, the birthplace of Aristippus.”*

These examples represent naming conventions or temporal ordering rather than causal processes.

D.1.3 Case 3: Semantically narrow or contextually inappropriate tagging

In some instances, causal annotations were either too narrow in scope or semantically inconsistent with the broader context:

- *“‘Mi Chico Latino’ received mixed reviews from music critics, who noted it was a contribution to the Latin pop phenomenon at the time, while others criticized Halliwell’s Spanish pronunciation.”*

Here, “it was a contribution to the Latin pop phenomenon” was tagged as a causal relation, but this framing conflates descriptive attribution with causation and lacks clear causal semantics.

D.1.4 Case 4: Anaphoric references

LLMs often tagged pronominal or anaphoric references that, although causally plausible within individual sentences, become ambiguous when integrated into our graph-construction pipeline:

- *“It is thematically divided into two sides: the rock music of Saturday Nights and the more country-influenced Sunday Mornings.”*

Although “country-influenced” suggests a causal modifier within the sentence, the pronominal reference (“it”) lacks clear antecedent grounding in our node-merging framework. To maintain graph consistency and prevent spurious node connections, we exclude such cases from the final dataset. These error patterns highlight areas where LLMs require further refinement to distinguish genuine causation from superficially similar linguistic patterns.

D.2 Model-Specific Failures in Prediction

We analyze representative examples where local models (Llama, OLMo, and Qwen) exhibit distinct failure patterns during graph-to-text generation. Two cases are presented: one involving basic causal structure and another involving hierarchical causality.

D.2.1 Case 1: Independent causal edges

The ground truth graph contains four nodes and two independent causal edges:

- Abdul security → Wounded
- Abhimanyu killed → Uttaraa widowed

The gold text correctly verbalizes these as two separate events by explicitly stating that:

*“Abdul’s security left people wounded.
Abhimanyu’s death left Uttaraa widowed.
Each statement refers to a distinct event.”*

Each of the three models exhibits a distinct type of misrepresentation of the text. Here, we list each of them, classifying them as semantic misinterpretation, incorrect causal chaining, and partial success.

Semantic misinterpretation (Llama). Llama generates: *“Abhimanyu’s actions led to Uttaraa being widowed.”* The model misinterprets the node label Abhimanyu killed as describing Abhimanyu performing an action, rather than Abhimanyu being killed. This reveals a failure in parsing passive constructions embedded in node labels, where the model defaults to an active-voice interpretation.

Incorrect causal chaining (OLMo). OLMo produces: *“Abdul’s security measures led to Wounded being injured, which in turn caused Uttaraa to become a widow.”* The model hallucinates a causal chain connecting two independent subgraphs. This suggests an over-reliance on narrative coherence, where the model attempts to construct a unified story even when the graph structure explicitly indicates disconnected components.

Partial success with residual errors (Qwen). Qwen outputs: *“Abdul’s security measures led to someone being wounded, while Abhimanyu’s actions resulted in Uttara becoming widowed.”* While correctly preserving the independence of the two relations (using “while” as a separator), Qwen shares Llama’s misinterpretation of the passive construction in Abhimanyu killed.

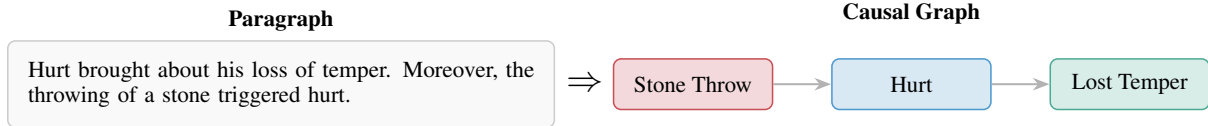


Figure 5: Chain: Text-graph pair showing sequential causation where each variable mediates the effect to the next.

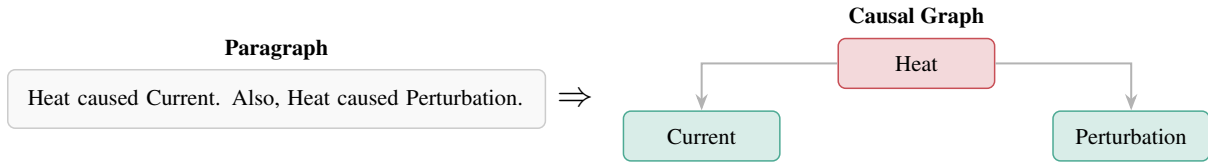


Figure 6: Fork: Text-graph pair where Heat acts as a common cause of Current and Perturbation, creating potential spurious correlation.

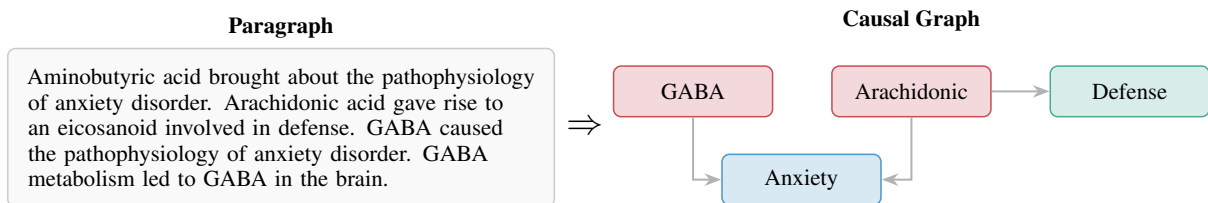


Figure 7: Collider: Text-graph pair where GABA and Arachidonic acid independently cause Anxiety, forming a collider structure.

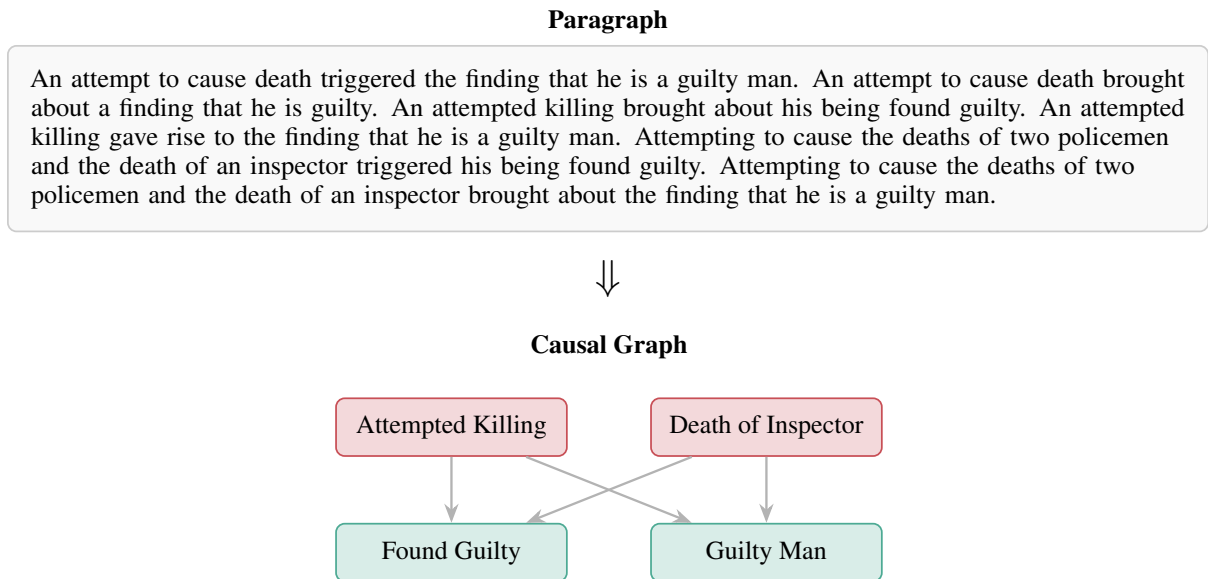


Figure 8: Bi-fan: Text-graph pair with four-node structure where two source nodes (Attempted Killing, Death of Inspector) independently influence two target nodes (Found Guilty, Guilty Man).

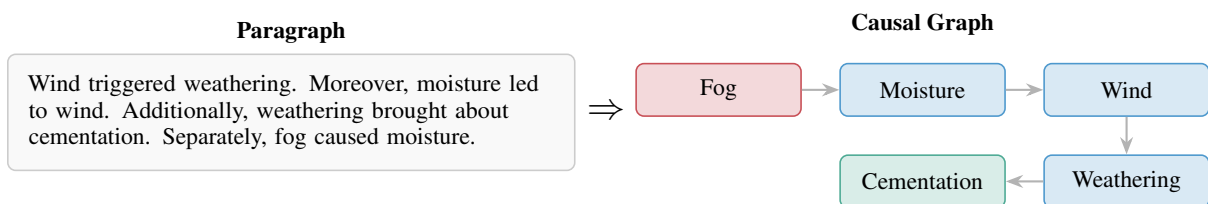


Figure 9: Long Chain: Text-graph pair with 5+ nodes representing complex multi-hop causal reasoning.

Discussion. This case highlights two systematic challenges: (i) node label parsing, where models struggle with implicit grammatical structures, and (ii) graph topology preservation, where models may impose spurious connections to achieve narrative fluency. Future work may benefit from explicit graph-structure conditioning or auxiliary supervision on grammatical role identification within node labels.

D.2.2 Case 2: Hierarchical and event-level causality

We further analyze a complex example involving hierarchical and event-level causality, where the three models exhibit clearly differentiated failure patterns despite being evaluated on the same input.

The ground-truth graph encodes a hierarchical causal structure in which both an abstract collective action and multiple concrete diplomatic events independently lead to governmental outcomes. Specifically, the gold annotations distinguish between African boycott as a high-level cause and two event-level nodes describing an African diplomat’s decision to boycott the celebration, with and without an India-specific context. The gold text correctly verbalizes this structure by explicitly separating abstract causes from concrete events, stating that:

“An African boycott triggered the government to take the issue lightly. An African diplomat’s decision to boycott the celebration in protest of renewed discrimination against them resulted in governmental outcomes. Each statement refers to a distinct causal event.”

Three models each exhibited qualitatively different misrepresentations of this structure. Here, we list each of them, classifying the types into abstraction-induced collapse, partial structural coverage, and event fragmentation.

Abstraction-induced causal collapse (Qwen). Qwen largely preserves the high-level causal relations but collapses multiple event-level causes into a single abstract node. While the gold graph distinguishes between African boycott and two concrete diplomatic boycott events, Qwen omits these fine-grained nodes and directly connects the abstract cause to all downstream outcomes. This results in a superficially coherent graph that sacrifices causal hierarchy and event-level granularity.

Partial event preservation with incomplete coverage (Llama). Llama recovers several event-level nodes and correctly links them to governmental outcomes, indicating sensitivity to explicit event descriptions. However, it fails to preserve all valid causal paths, omitting some effects when multiple independent causes lead to the same outcome. As a result, the predicted graph captures only a subset of the intended causal structure.

Event fragmentation and spurious structural expansion (OLMo). OLMo deviates most strongly from the gold graph by fragmenting coherent event-level nodes into actors, actions, and contextual entities. This over-decomposition introduces spurious intermediate nodes and syntactically driven edges, obscuring causal attribution and degrading interpretability despite producing a structurally dense graph.

Discussion. Across models, failures reflect systematic inductive biases rather than isolated errors: (i) Qwen favors abstraction at the cost of causal granularity, (ii) Llama preserves events but exhibits incomplete causal coverage, and (iii) OLMo over-fragments events, introducing spurious structure. Together, these patterns suggest that current LLM-based approaches struggle to simultaneously preserve event atomicity, causal hierarchy, and edge completeness, highlighting the need for graph-aware constraints during causal graph generation.

D.3 Four Types of Errors

We present representative examples of the four error types identified during our dataset review process; the corresponding statistics for each category are reported in Section 3.5.

Problematic. Problematic refers to cases containing violent, inflammatory, or controversial content targeting specific individuals, countries, or groups. Some instances were derived from news corpora and contained crime-related content in which causal relationships are relatively unambiguous. Because such instances risk identifying or implicating specific individuals or groups, they were removed from the dataset.

“Ahmed was convicted of the murder of a police officer. The statement concerns a criminal case. ... ”

Artifact. Artifact denotes cases in which natural semantic interpretation is hindered by ungrammatical expressions, meta-level phrasing, or missing contextual information. During the LLM-based paraphrasing stage, certain instances exhibited grammatical inconsistencies or irregular causal structures that were distorted from the original meaning. These instances were discarded as they no longer faithfully represent the intended causal structure.

“Saved similar caused Draw plan. Saved similar led to Internal plan. Saved situation brought about Draw plan. Saved situation gave rise to Internal plan. That they can be saved from the same situation caused Internal plan. That they can be saved from the same situation led to Drew plan. Saved similar triggered Drew plan. Saved situation resulted in Drew plan. That they can be saved from the same situation resulted in Draw plan.”

Nonsense. Nonsensical refers to cases where causal relationships are difficult to establish or where semantic coherence is weak due to excessive reliance on normative statements. We observed cases where LLM refinement produced passages in which sentences were joined illogically or without sufficient contextual grounding, resulting in text that lacked narrative coherence. Such instances were excluded accordingly.

“Remote sensing produced assessment data. Removal caused Mount. Removal led to Amass.”

Gibberish. Gibberish refers to cases in which graph nodes consist of demonstrative pronouns, or remain in clause-level or placeholder form, rendering them insufficiently functional as semantic units. Because our pipeline extracts sentences containing causal relationships, certain extracted instances—while causally valid in their original context—relied on anaphoric expressions (e.g., due to pronouns or incomplete references) or omitted surrounding context, rendering the causal relationship ambiguous. These instances were also removed.

“Getrss good resulted in the party leader not condemning the violence by a worker from RSS. Getrss good caused the leader to condemn. Getrss possible triggered

the party leader not condemning the violence by a worker from RSS. Getrss possible brought about the leader condemnation. Getrss recommended resulted in the party leader not condemning the violence by a worker from RSS. Getrss recommended caused the leader to condemn. Getrss good led to the condemnation of RSS. Getrss possible gave rise to the condemnation of RSS. Getrss recommended led to the condemnation of RSS.”

E Examples of Text-to-Graph Motifs

This section provides representative text-graph pairs from our dataset. Regarding the fundamental structural motifs, given three vertices X , Y , and Z in a graph: chain is defined as $X \rightarrow Y \rightarrow Z$. Fork represents a common cause structure, defined as $X \leftarrow Y \rightarrow Z$. Collider represents a common effect structure, defined as $X \rightarrow Y \leftarrow Z$ (Zanga et al., 2022). We categorize chains based on the number of nodes: a chain consists of 3 to 4 nodes, whereas a long chain comprises 5 or more nodes. A bi-fan is defined as a structure with four nodes where two distinct common causes influence two separate outcomes (Milo et al., 2002). Examples of these configurations and explanation are provided in Section E. Each figure below presents a natural language paragraph alongside its corresponding causal graph, which demonstrates how our integrated pipeline transforms textual paragraphs into structured causal graphs.

The examples illustrate varying levels of structural complexity ranging from the simplest three-node chains (Figure 5) to more sophisticated long chains with five or more nodes (Figure 9). These diverse structures enable comprehensive evaluation of model performance across different causal configurations.

F Additional Experimental Results

This section describes the complete set of results used in Section 4, including the ablation studies. Tables 7 and 6 report the full results across all evaluated models and metrics. The tables report corresponding results obtained under zero-shot and few-shot settings to minimize the influence of in-context learning.

Overall, the trends observed here are largely consistent with those discussed in Section 4. A notable difference, however, is that once model size

Task	Model	Setting	ROUGE-L	SBERT-cos
G2T — zero-shot				
	GPT-5-mini	Plain	0.250	0.770
	Gemini 2.5 flash	Plain	0.219	0.747
	Claude Haiku 4.5	Plain	0.192	0.736
	Grok-4-fast	Plain	0.199	0.732
	Claude Haiku 4.5	Role	0.180	0.612
	Grok-4-fast	Role	0.186	0.611
	GPT-5-mini	Role+Rel	0.243	0.733
	Claude Haiku 4.5	Role+Rel	0.186	0.649
	Grok-4-fast	Role+Rel	0.190	0.652
G2T				
	GPT-5*	Plain	0.2766	0.8006
	GPT-5-mini	Plain	0.2587	0.7657
	Claude Haiku 4.5	Plain	0.2053	0.7461
	Grok-4-fast	Plain	0.2090	0.7384
	Gemini 2.5 flash	Plain	0.2769	0.7832
	Llama-3-8B	Plain	0.2923	0.7868
	OLMo-2-7B	Plain	0.2094	0.7681
	Qwen-2.5-7B	Plain	0.3345	0.8172
	GPT-5-mini	Role	0.2401	0.6690
	Claude Haiku 4.5	Role	0.1917	0.6169
	Grok-4-fast	Role	0.1950	0.6181
	Llama-3-8B	Role	0.2619	0.6750
	OLMo-2-7B	Role	0.1953	0.6473
	Qwen-2.5-7B	Role	0.2909	0.6763
	GPT-5-mini	Role+Rel	0.2717	0.7537
	Claude Haiku 4.5	Role+Rel	0.1926	0.6505
	Grok-4-fast	Role+Rel	0.1949	0.6590
	Llama-3-8B	Role+Rel	0.2503	0.7099
	OLMo-2-7B	Role+Rel	0.2115	0.6877
	Qwen-2.5-7B	Role+Rel	0.2481	0.7058
T2G2T — zero-shot				
	GPT-5	Plain	0.304	0.829
	GPT-5-mini	Plain	0.381	0.837
	Claude-Sonnet-4.5	Plain	0.041	0.077
	Claude-Haiku-4.5	Plain	0.235	0.754
	Grok-4-fast	Plain	0.224	0.730
	Llama-3-8B	Plain	0.331	0.755
	Qwen-2.5-7B	Plain	0.372	0.802
	OLMo-2-7B	Plain	0.191	0.727
T2G2T				
	GPT-5	Plain	0.304	0.829
	Claude-Sonnet-4.5	Plain	0.041	0.077
	GPT-5-mini	Plain	0.380	0.835
	Claude-Haiku-4.5	Plain	0.235	0.756
	Grok-4-fast	Plain	0.225	0.734
	Llama-3-8B	Plain	0.331	0.755
	Qwen-2.5-7B	Plain	0.372	0.802
	OLMo-2-7B	Plain	0.191	0.727
	GPT-5-mini	Role	0.320	0.694
	Claude-Haiku-4.5	Role	0.218	0.621
	Grok-4-fast	Role	0.209	0.608
	Llama-3-8B	Role	0.308	0.642
	Qwen-2.5-7B	Role	0.322	0.658
	OLMo-2-7B	Role	0.185	0.619
	GPT-5-mini	Role+Rel	0.288	0.741
	Claude-Haiku-4.5	Role+Rel	0.204	0.642
	Grok-4-fast	Role+Rel	0.201	0.637
	Llama-3-8B	Role+Rel	0.268	0.673
	Qwen-2.5-7B	Role+Rel	0.267	0.698
	OLMo-2-7B	Role+Rel	0.205	0.672

Table 6: Text similarity metrics grouped by shared metric format (ROUGE-L and SBERT cosine similarity). * indicates that the corresponding experiment was conducted on a 20% sample of the dataset.

exceeds a certain threshold, performance differences become relatively small. For example, we observe no substantial performance gap between

GPT-5 and GPT-5-mini. In contrast, Claude Sonnet and Claude Haiku consistently exhibit lower performance compared to other models.

Task	Model	Setting	Precision	Recall	F1	Acc.
T2G — zero-shot						
	Qwen-2.5-7B	Plain	(0.574, 0.478)	(0.587, 0.467)	(0.575, 0.470)	(0.587, 0.467)
	OLMo-2-7B	Plain	(0.493, 0.381)	(0.504, 0.384)	(0.495, 0.381)	(0.504, 0.384)
T2G						
	GPT-5-mini	Plain	(0.934, 0.894)	(0.943, 0.895)	(0.937, 0.894)	(0.943, 0.895)
	Claude Sonnet 4.5	Plain	(0.123, 0.119)	(0.123, 0.119)	(0.123, 0.119)	(0.123, 0.119)
	Grok-4-fast	Plain	(0.927, 0.881)	(0.925, 0.872)	(0.924, 0.874)	(0.925, 0.872)
	Llama-3-8B	Plain	(0.575, 0.458)	(0.619, 0.467)	(0.591, 0.459)	(0.619, 0.467)
	OLMo-2-7B	Plain	(0.532, 0.142)	(0.545, 0.146)	(0.533, 0.141)	(0.545, 0.146)
	Qwen-2.5-7B	Plain	(0.663, 0.549)	(0.692, 0.539)	(0.670, 0.541)	(0.692, 0.539)
	GPT-5	Plain	(0.900, 0.824)	(0.901, 0.824)	(0.900, 0.823)	(0.901, 0.824)
	GPT-5-mini	Role	(0.821, 0.717)	(0.833, 0.813)	(0.829, 0.802)	(0.833, 0.813)
	Grok-4-fast	Role	(0.926, 0.882)	(0.924, 0.871)	(0.925, 0.875)	(0.924, 0.871)
	Claude Haiku 4.5	Role	(0.228, 0.198)	(0.207, 0.186)	(0.214, 0.200)	(0.207, 0.186)
	Llama-3-8B	Role	(0.576, 0.455)	(0.614, 0.451)	(0.589, 0.447)	(0.614, 0.451)
	OLMo-2-7B	Role	(0.291, 0.069)	(0.296, 0.068)	(0.290, 0.068)	(0.296, 0.068)
	Qwen-2.5-7B	Role	(0.300, 0.133)	(0.317, 0.133)	(0.305, 0.132)	(0.317, 0.133)
	GPT-5-mini	Role+Rel	(0.815, 0.704)	(0.847, 0.792)	(0.823, 0.816)	(0.847, 0.792)
	Grok-4-fast	Role+Rel	(0.918, 0.864)	(0.932, 0.880)	(0.921, 0.889)	(0.932, 0.880)
	Claude Haiku 4.5	Role+Rel	(0.242, 0.182)	(0.191, 0.168)	(0.227, 0.214)	(0.191, 0.168)
	Llama-3-8B	Role+Rel	(0.275, 0.127)	(0.298, 0.125)	(0.282, 0.125)	(0.298, 0.125)
	OLMo-2-7B	Role+Rel	(0.495, 0.190)	(0.515, 0.195)	(0.495, 0.190)	(0.515, 0.195)
	Qwen-2.5-7B	Role+Rel	(0.446, 0.255)	(0.463, 0.256)	(0.450, 0.252)	(0.463, 0.256)
G2T2G — zero-shot						
	Qwen-2.5-7B	Plain	(0.266, 0.171)	(0.276, 0.171)	(0.269, 0.170)	(0.276, 0.171)
G2T2G						
	GPT-5	Plain	(0.228, 0.092)	(0.249, 0.100)	(0.233, 0.094)	(0.249, 0.100)
	GPT-5-mini	Plain	(0.450, 0.367)	(0.513, 0.400)	(0.467, 0.375)	(0.513, 0.400)
	Grok-4-fast	Plain	(0.080, 0.025)	(0.131, 0.035)	(0.094, 0.027)	(0.131, 0.035)
	OLMo-2-7B	Plain	(0.601, 0.496)	(0.611, 0.504)	(0.602, 0.496)	(0.611, 0.504)
	Llama-3-8B	Plain	(0.228, 0.121)	(0.292, 0.140)	(0.247, 0.124)	(0.292, 0.140)
	Qwen-2.5-7B	Plain	(0.266, 0.171)	(0.276, 0.171)	(0.268, 0.170)	(0.276, 0.171)
	GPT-5-mini	Role	(0.446, 0.360)	(0.507, 0.391)	(0.462, 0.367)	(0.507, 0.391)
	Grok-4-fast	Role	(0.093, 0.042)	(0.170, 0.067)	(0.113, 0.048)	(0.170, 0.067)
	OLMo-2-7B	Role	(0.601, 0.496)	(0.611, 0.504)	(0.602, 0.496)	(0.611, 0.504)
	Llama-3-8B	Role	(0.228, 0.121)	(0.292, 0.140)	(0.247, 0.124)	(0.292, 0.140)
	Qwen-2.5-7B	Role	(0.266, 0.171)	(0.276, 0.171)	(0.268, 0.170)	(0.276, 0.171)
	GPT-5-mini	Role+Rel	(0.441, 0.358)	(0.504, 0.392)	(0.458, 0.366)	(0.504, 0.392)
	Grok-4-fast	Role+Rel	(0.090, 0.039)	(0.165, 0.063)	(0.109, 0.044)	(0.165, 0.063)
	OLMo-2-7B	Role+Rel	(0.601, 0.496)	(0.611, 0.504)	(0.602, 0.496)	(0.611, 0.504)
	Llama-3-8B	Role+Rel	(0.228, 0.121)	(0.292, 0.140)	(0.247, 0.124)	(0.292, 0.140)
	Qwen-2.5-7B	Role+Rel	(0.266, 0.171)	(0.276, 0.171)	(0.268, 0.170)	(0.276, 0.171)

Table 7: Graph-structure metrics grouped by shared metric format. Values in parentheses denote (node score, edge score).

Under the zero-shot setting, performance degradation is observed in some cases, while in others the difference remains marginal. Consistent with the findings in Section 4, tasks that require graph construction benefit substantially from in-context

learning, as few-shot examples provide explicit guidance on graph structure. When such guidance is removed in the zero-shot setting, performance drops more noticeably. In contrast, tasks that focus on text reconstruction exhibit only minor perfor-

Task	Model	Output Length Mean (Std)
T2G	OLMo-2-7B	256.24 (158.81)
	Qwen-2.5-7B	276.29 (152.14)
	LLaMA-3-8B	282.17 (175.18)
	Grok-4-fast	310.70 (273.89)
	Haiku-4.5	316.78 (239.85)
	GPT-5-mini	398.38 (369.45)
G2T	Qwen-2.5-7B	153.35 (99.35)
	LLaMA-3-8B	207.93 (104.47)
	GPT-5	266.13 (94.23)
	GPT-5-mini	375.54 (101.92)
	Gemini-2.5-Flash	416.74 (115.22)
	OLMo-2-7B	487.59 (176.51)
	Grok-4-fast	504.45 (95.14)
	Haiku-4.5	530.71 (116.76)
T2G2T	Qwen-2.5-7B	144.11 (66.79)
	LLaMA-3-8B	226.57 (108.86)
	GPT-5-mini	329.15 (117.44)
	Haiku-4.5	463.69 (135.47)
	Grok-4-fast	497.21 (98.54)
	OLMo-2-7B	551.89 (144.01)
G2T2G	Sonnet-4.5	252.23 (76.97)
	Qwen-2.5-7B	288.67 (139.64)
	OLMo-2-7B	298.33 (184.94)
	GPT-5	334.30 (128.42)
	GPT-5-mini	555.74 (412.15)
	Grok-4-fast	707.77 (251.28)

Table 8: Output length statistics across all tasks (Mean (Std)). Gold text mean is 227.41 and gold graph mean is 340.65.

mance degradation, suggesting that these tasks rely less on structural cues provided by few-shot examples.

We hypothesized that larger LLMs tend to generate longer outputs, and that this verbosity would negatively correlate with task performance. Our empirical results partially support this hypothesis, with task-dependent variation as shown in Table 8 and Figure 10 and 11. For the T2G task, we observe the opposite trend: longer outputs are associated with higher accuracy, suggesting that larger, more capable models not only generate more verbose graph representations but also do so more accurately. Larger LLMs tend to generate longer outputs, which generally hurts performance on G2T and T2G2T as shown in Table 8. T2G is a causal extraction task that maps sentences to directed graphs, a well-studied NLP problem for which abundant supervised datasets exist. We conjecture that closed-source models, trained on larger corpora, are more likely to have been exposed to such causal extraction data, giving them a systematic advantage on this particular task type.

In contrast, G2T shows a clear negative correlation between output length and ROUGE-L score. Models producing shorter, more concise outputs – such as Qwen-2.5-7B and LLaMA-3-8B – achieve higher ROUGE-L scores, while larger models that tend toward verbose generation, such as Grok-4-fast and Haiku-4.5, score lower. This aligns with our hypothesis that over-generation hurts performance on tasks where fidelity to a reference text is measured. T2G2T exhibits a broadly similar negative trend, with one notable exception: GPT-5-mini produces the longest output among the models yet achieves the highest ROUGE-L score. We attribute this exception to the nature of the preceding step: T2G2T is conditioned on a T2G output, and T2G is a task that LLMs already perform well. As a result, the intermediate graph representation is of sufficiently high quality that capable models like GPT-5-mini can leverage it effectively, allowing generation quality to compensate for length bias. G2T2G, on the other hand, does not conform to either trend. OLMo-2-7B stands out as a clear outlier, achieving the highest accuracy despite producing relatively short outputs, while Grok-4-fast generates the longest outputs yet performs worst. Unlike T2G2T, G2T2G is conditioned on a G2T output, and G2T is a task where LLMs generally struggle. The poor quality of the intermediate text representation introduces compounding errors across the round-trip conversion. Consequently, output length becomes a poor predictor of final performance and obscures any systematic trend. Taken together, our findings suggest that while the length-bias hypothesis holds for text-output tasks in general, the quality of preceding steps acts as a critical moderator. Specifically, when the upstream task is well-handled (T2G), capable models can overcome length bias. However, when the upstream task is poorly handled (G2T), the resulting noise dominates, causing length-performance correlations to break down entirely.

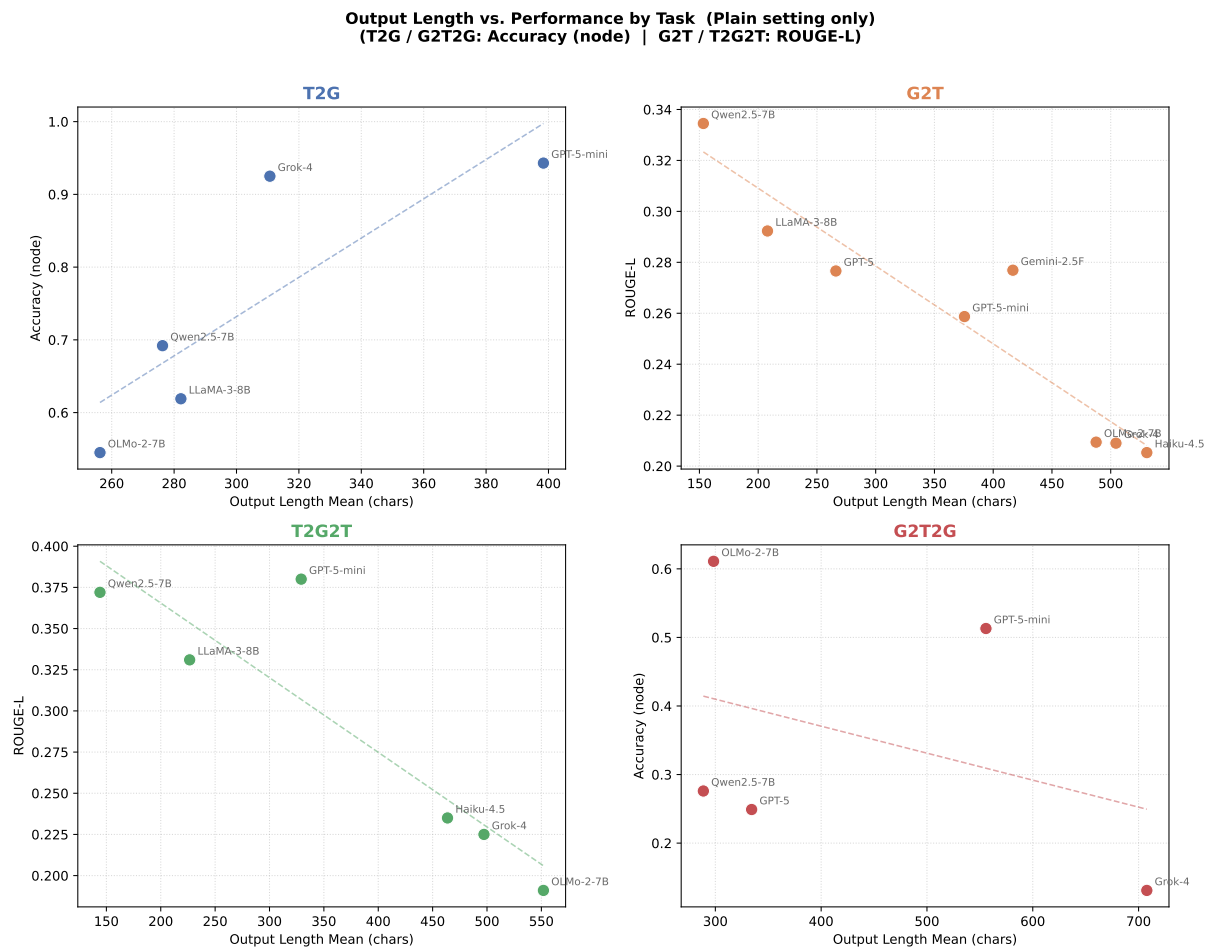


Figure 10: Correlation between output length (X-axis) and task performance (Accuracy, ROUGE-L) (Y-axis) across tasks.

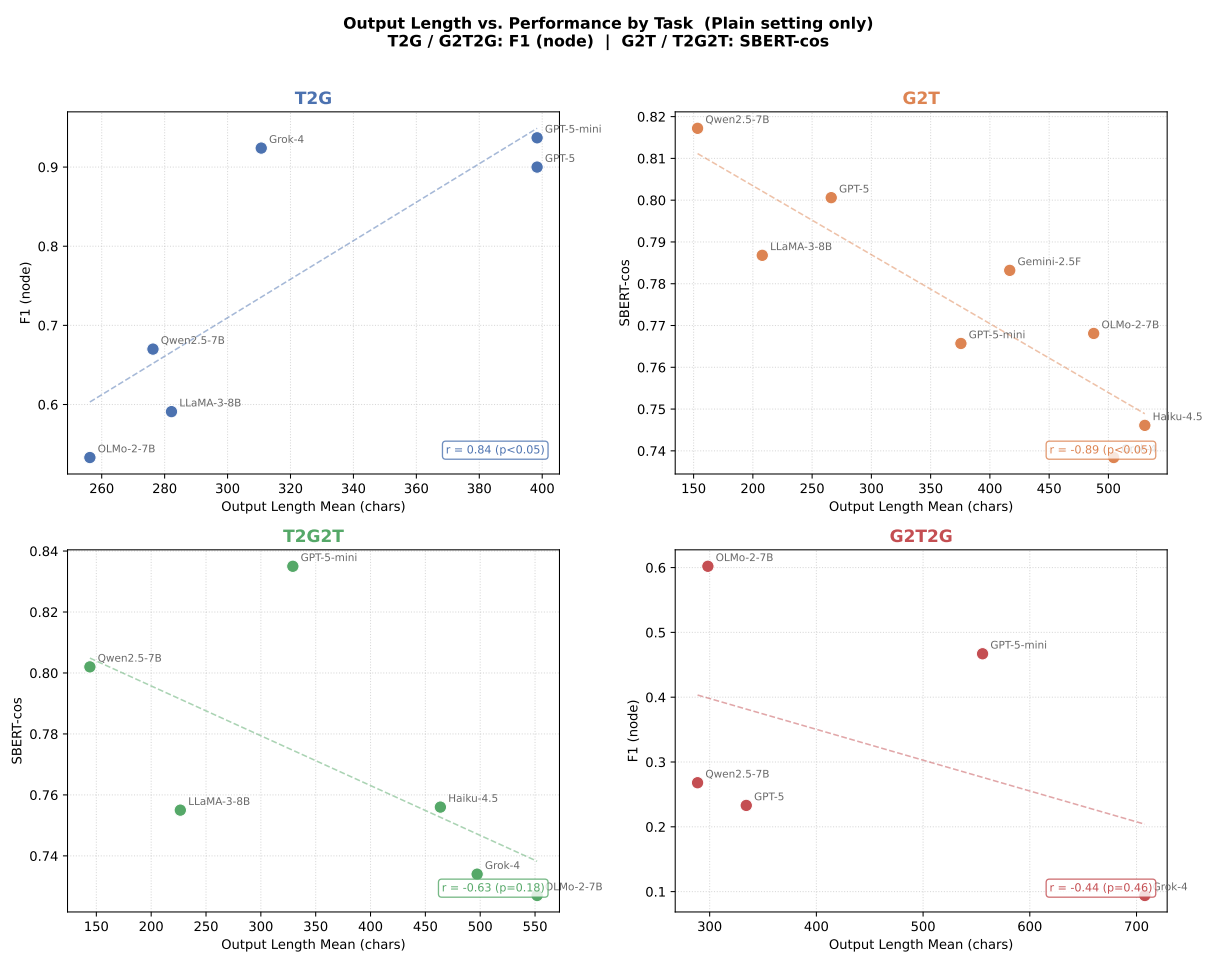


Figure 11: Correlation between output length (X-axis) and task performance (F1, SBERT-cos) (Y-axis) across tasks.