Failure Modes of LLMs for Causal Reasoning on Narratives

Khurram Yamin [1] **Shantanu Gupta** [1] **Gaurav Ghosal** [1] **Zach Lipton** [1] **Bryan Wilder** [1]

Abstract

The ability to robustly identify causal relationships is essential for autonomous decision-making and adaptation to novel scenarios. However, accurately inferring causal structure requires integrating both world knowledge and abstract logical reasoning. In this work, we investigate the interaction between these two capabilities through the representative task of causal reasoning over narratives. Through controlled synthetic, semi-synthetic and real-world experiments, we find that state-of-the-art large language models (LLMs) often rely on superficial heuristics—for example, inferring causality from event order or recalling memorized world knowledge without attending to context. Furthermore, we show that simple reformulations of the task can elicit more robust reasoning behavior. Our evaluation spans a range of causal structures, from linear chains to complex graphs involving colliders and forks. These findings uncover systematic patterns in how LLMs perform causal reasoning and lay the groundwork for developing methods that better align LLM behavior with principled causal inference.

1. Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities across diverse tasks by ingesting vast quantities of unsupervised data, which endows them with extensive world knowledge. This success has fueled interest in deploying LLMs as autonomous agents. A fundamental requirement for reliable autonomy, however, is causal reasoning: agents must go beyond mere correlations and anticipate the effects of their actions. It remains unclear whether causal inference emerges naturally from large-scale pretraining.

Causal reasoning poses unique challenges because it demands both domain-specific knowledge and flexible inferential strategies. Unlike mathematical benchmarks, which draw on well-defined solution procedures, causal questions often hinge on contextual details about events and their relationships. Models must resist the temptation to default to memorized associations, especially in unusual or counter-intuitive scenarios. Striking the right balance between recalling factual knowledge and applying logical reasoning is critical for robust causal inference. Previous evaluations have largely treated reasoning and factual recall in isolation: mathematical or coding tasks assess pure logical skill, while knowledge benchmarks measure retrieval. The interaction between these abilities—and the potential clashes that arise when they conflict—has received little attention.

We explore this interaction through causal reasoning over textual narratives that describe an underlying graph structure. Our narratives either follow a simple chain $V_1 \to V_2 \to \cdots \to V_N$, or more complex graphs containing forks and colliders. We examine two core tasks: (a) deciding whether one event influences another, and (b) reconstructing the full causal graph implied by the text. These tasks capture essential primitives of causal reasoning, and we pay special attention to long event sequences and atypical cause-effect pairings that contradict common-sense expectations.

Our experiments reveal two dominant failure modes. First, LLMs exhibit a strong positional bias: they tend to treat earlier events as causes and later events as effects. When narratives are presented in reverse causal order, accuracy falls sharply because the model mis-attributes causal direction. Second, LLMs rely on their parametric "world knowledge" as a shortcut: if pretrained associations conflict with the story's implications, the model ignores the narrative and defaults to its memorized beliefs. Neither chain-of-thought prompting nor in-context learning resolves these issues.

Remarkably, asking the model to extract the entire causal graph from the narrative, and then directly using that graph for reasoning largely overcomes both biases and is robust to increased narrative length. We validate these findings on both synthetic, LLM-generated stories and real-world cause-effect data.

[*]Equal contribution [1]Department of Machine Learning, Carnegie Mellon. Correspondence to: Khurram Yamin <khurram.yamin24@gmail.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

2. Related Works

Causal Reasoning in Large Language Models Jin et al. (2023) developed a benchmark for LLM causal reasoning using causal graphs, finding models can struggle. However, their queries required probability calculations, potentially conflating causal reasoning with arithmetic failures. Tan et al. (2022) showed a neural network's capability to label causal structures in news sentences. Joshi et al. (2024b) chronicled failure modes in textual, non-narrative data (e.g., "Event 1 Causes Event 2"). Our paper expands this by testing LLMs on plausible real and synthetic everyday texts. In contrast, Jin et al. (2024) used only statistical language indicating event correlations as input.

Work by (Gordon et al., 2012; Joshi et al., 2024a; Ho et al., 2023; Zhang et al., 2023; Wang et al., 2023; Ashwani et al., 2024) studies common-sense causal inference, where models might rely on memorized pretrained knowledge, achieving good performance without general causal reasoning. Our work seeks to disentangle this by testing cases where causal relationships contradict common-sense knowledge. This provides a more robust measurement of causal reasoning in unfamiliar and atypical scenarios.

A key distinction of our work is its focus on longer-form narratives. Existing works like (Gordon et al., 2012; Zečević et al., 2023; Ho et al., 2023; Frohberg & Binder, 2022; Li et al., 2023; Gao et al., 2023) primarily examine short-form questions on single causal relationships. In contrast, we study longer, more complex event sequences. Furthermore, unlike domain-specific question banks (e.g., Intuitive Physics in (Zečević et al., 2023)), our narratives cover diverse topics, offering a more realistic and varied examination of LLM causal reasoning. Our work is unique as the first to analyze non-common sense causal reasoning within everyday language narratives, providing a more robust test than prior studies.

Causal Story Generation Kıcıman et al. (2024) shows LLMs' strong causal text generation. Ammanabrolu et al. (2020) used soft causal relations and commonsense inferences for coherent narratives. Tian et al. (2021) employed counterfactual knowledge for realistic hyperbole. Li et al. (2022) found indicating sentence numbers for cause/effect allows generation to better respect causal relations. In regards to our work, we focus on synthetic texts that are explicit and simple. Unlike Ammanabrolu et al. (2020) and Li et al. (2022), we embed explicit causal language, requiring no commonsense inference as our experiments intentionally contradict common sense to test narrative reliance. We also avoid abstract language unlike Tian et al. (2021).

3. Experiments with Synthetic Data

3.1. Setting

Synthetic Narrative Generation In our synthetic experiments, we use three leading LLMs: OpenAI's GPT-4o (OpenAI et al., 2024), Anthropic's Claude 3.5 Sonnet (Anthropic, 2024), and the open source LLama 3.1 8b (Grattafiori et al., 2024). While we focus on GPT-4o in the main text, results from other models are in the Appendix. The purpose of our synthetic setup is to carefully control the conditions under which the LLMs are tested. We first use the LLM to generate events (which are real world phenomena like *rain* or *plants growing*). Then these events are linked together into a chain graph G that acts as the causal ground truth (eg *rain* $\rightarrow$ *plants growing*). The LLM is given G and asked to create a narrative that stays faithful to the causal relationships in G. These narratives are checked by researchers to ensure consistency with their base causal graphs. More specifically, when constructing the dataset, we asked researchers (select authors who were blind to the true underlying graph) to reconstruct the causal chains given just the narratives, and 98 percent of the time (out of 100 random samples), the humans were able to find the unique correct causal ordering. Roughly 2500 narrative samples were generated. (all narratives in supp. files and select narratives in Appendix).

Providing only the narrative as input (and not G), we then ask the LLM to find G', the predicted underlying causal structure expressed by the narrative. In other words, the LLM is asked to output a causal graph that it thinks embodies the relationships in the narrative. Next, a series of causal questions is created by randomly sampling 10 tuples of events from G and asking the LLM whether an event in the tuple causes the other based on the narrative and/or G'.

Prompting Strategies We evaluate five prompting styles for causal reasoning where the names in italics represent those used in the legends of figures: **Standard QA Prompting** (*Standard*), where the model is simply asked to identify the causal relation between two narrative events; **Chain-of-Thought** (*CoT*), which instructs the model to articulate step-by-step reasoning before answering; **In-Context Learning** (*In-Context*), which precedes the query with illustrative question–answer examples; **Explicit Causal Graph Extraction** (*Graph*), which asks the model to generate an entire causal graph G' over all events and assesses whether the ordering of the target pair is correct; **Narrative-Augmented Graph Extraction** (*Narr-Graph*), which first elicits G' and then supplies both G' and the original narrative for joint reasoning about the causal pair. Exact prompts Appendix B.

3.2. Impact of Event Ordering

Our experiments show that LLMs rely on the ordering in which the events are verbalized in a narrative when deter-

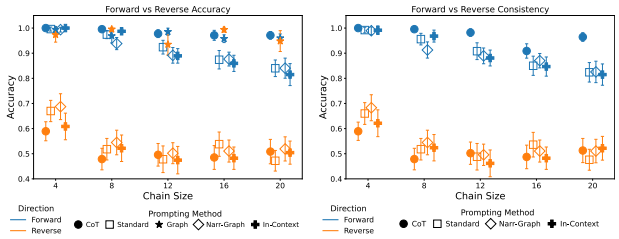


Figure 1: GPT-4o Test of the LLM's ability to reason on narratives written in the Forward and Reverse topological orientations. Chain size is the number of nodes in ground truth G. Prompts described in 3.1. Accuracy measures LLM answer agreement with G, and consistency measures agreement with G'. The points in the graph are represented with a slight horizontal stagger around the relevant chain sizes for visual ease. We show a 95% CI.

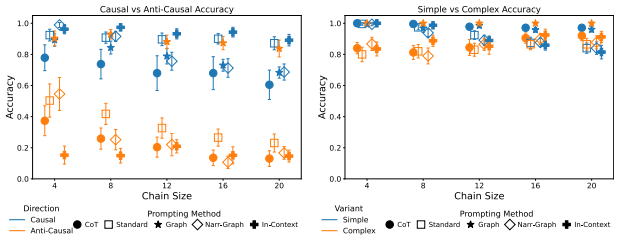


Figure 2: (Left) GPT-4o test of the LLM's ability to reason on narratives that agree with parametric knowledge (Causal) and disagree with parametric knowledge (Anti-Causal). (Right) GPT-4o test of the LLM's ability to reason on narratives generated from Complex graphs as opposed to Simple chain graphs. 95 % CI is shown.

mining causal relationships. To investigate this, we started with randomly generated events that were used to make a ground truth graph G. During the creation of the narrative, we specified that the LLM either places the events in (1) the order that matches the topological causal ordering of the graph (e.g., if event A (directly or indirectly) causes B, then event A is mentioned before B in the narrative), or (2) a way that runs opposite to the causal ordering (event B would be mentioned before event A in the narrative even though A causes B). We refer to these as *Forward* and *Reverse* orderings, respectively. For example, the following is a GPT-4o generated *Reverse* topological narrative for the causal chain: *Art exhibition*$\rightarrow$ *Wine tasting* $\rightarrow$ *Charity fundraiser*:" The *charity fundraiser* was made possible because of the successful *wine tasting event* that attracted numerous generous patrons. The *wine tasting* was organized as a result of the *art exhibition* drawing in a sophisticated audience interested in cultural experiences. "

LLMs Rely on Event Ordering Across Prompting Strategies As shown in Figure 1 (left), in the *Forward* direction, standard QA, CoT, and In-Context prompts all perform very well. This is in contrast to the *Reverse* orientation when we look at the performance of the standard QA, COT, and In-Context prompts. From this plot, we can see that naive COT and In-Context prompting do not seem to significantly boost accuracy under our conditions. Perhaps more interestingly, we find that the way the LLM answers questions using the narratives is not always consistent with the causal graph G' that the LLM builds when asked to predict the underlying graph structure (see consistency plot in right side of Figure 1, where consistency measures agreement between the answers of the LLM and G'). In the *Reverse* orientation, answers given by the extracted causal graph G' and the previously discussed prompting strategies seem to differ greatly.

Explicit Causal Graph Extraction Avoids Shortcuts This led us to test the accuracy of only using the extracted graph G' to answer causal questions. In this case, once G' is extracted by the LLM, it is not given to the LLM again to answer questions (but rather used directly). We found that this strategy did significantly better in the *Reverse* direction than the other prompting strategies. Surprisingly, using G' in the *Reverse* direction narratives to answer causal questions did as well as using G' in the *Forward* direction narratives. Next, we tried prompting using the narrative and G' (the LLM is given G' in this case in the prompt). This technique could be thought of as a type of CoT prompting strategy. However, in the *Reverse* direction narratives, the increase in accuracy achieved by only using G' completely dissipates. We conjecture that the process of building the extracted Causal Graph G' forces the LLM to engage in long term reasoning , but when the narrative is again provided - the LLM defaults back to the shortcut.

3.3. Impact of Parametric Knowledge (In)consistency

Experimental Setup We also find that LLMs tend to rely on parametric knowledge when it is present, and can fail when narratives are inconsistent with the LLM's parametric knowledge. To test this, we elicit the LLM's pre-existing parametric knowledge when generating the event chains. We prompt the LLM to pick a series of events such that each event has some relation to the subsequent event – either the event is *Causal* to the next event (e.g., disease causes shorter lives) or the event is *Anti-Causal* (e.g., disease causes longer lives). After the ground truth graph is created, we generate the narrative in the *Forward* topological orientation to avoid confounding failure modes. The full process (along with illustration) explaining how the graphs are created is in Appendix B.2. As a textual example, assume that we know a parametric anti-causal link exists from *stressful job* to *increased happiness*, and from *lack of sleep* to *improved cognitive function*. We can then construct the causal chain

Stressful Job → Lack of Sleep → Increased Happiness → Improved Cognitive Function. From this causal chain, we create the narrative: "The constant demands of a *stressful job* led to her experiencing chronic *lack of sleep*. Surprisingly, she found that the *lack of sleep* heightened her sense of euphoria, making her unusually cheerful at work. *Increased happiness* from this cheerfulness seemed to improve her *cognitive function*." If the LLM is asked if a stressful job leads to increased happiness, the parametric knowledge shortcut indicates no – however, the shortcut fails as the narrative indicates that a (indirect) causal link does exist.

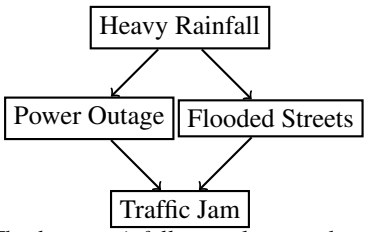
Models Exploit Parametric Knowledge We find that, in synthetic experiments, the LLM finds the correct causal relation generally only when that relation agrees with its parametric knowledge. This is exemplified in the plot in Figure 2 (left) where we see good performance on narratives that agree with parametric knowledge (*Causal* parametric knowledge) and poor performance on narratives that disagree with parametric knowledge (*Anti-Causal* parametric knowledge). We also notice an interesting phenomenon for the *Anti-Causal* case where using just the extracted graph provides massive improvements over any prompting strategy that involves using the narrative to answer questions. It seems that the narrative may only serve to distract the LLM when parametric knowledge conflicts with the narrative.

3.4. Impact of Narrative Complexity

Narrative Length In conditions where the LLM exhibits failure modes (*Reverse* and *Anti-Causal* orientations), the performance also tends to decay as the size of the narrative and the number of events in the narrative increases. As we can see in Figures 1 and 2 (Left), it seems that the longer the narrative is, the more the LLM relies on shortcuts instead of performing reasoning. However, the extracted graph G' can often maintain a consistently high level of accuracy.

Causal Graph Complexity As the bulk of our work has focused on detecting the simplest failure modes possible, we studied narratives with an underlying chain graph structure. However, the presence of more complex causal structures in the narrative could exacerbate the existing failure modes or trigger novel failures. To study this, we create causal graphs utilizing two common causal structures: *Forks* (one node has a causal relationship to multiple other nodes) and *Colliders* (multiple nodes have a causal relationship to the same node). We generate narratives (the complete algorithm is described in Appendix B.3) such that each underlying causal graph contains at least one of these structures, and may randomly contain multiple such structures based on the size of the narrative. An example is shown in Figure 3.

As can be seen in Figure 2 (right side), we find that while the LLM generally performs worse at reasoning about the complex narratives than simple narratives (with underlying chain graphs), the gap is very starkly less than can be seen in the other failure modes. This finding can be supported by (Dettki et al., 2025) which finds that GPT-4o reasons similarly to humans on a single sentence that describes one collider relation. Our work extends their work by using a long-form narrative based on a causal graph with potentially multiple colliders and forks instead of only one collider.



Narrative: The *heavy rainfall* not only caused a *power outage* in several neighborhoods but also led to *flooded streets*. The aftermath of the *power outage* (disabling traffic lights) and the *flooded roads* (blocking street access) caused a *traffic jam*.

Figure 3: Causal graph with story showing a fork (first sentence) and a collider (second sentence).

4. Semi-Synthetic and Real-World Narratives

We extend our analysis to narratives involving real-world causal graphs from *CauseNet* (Heindorf et al., 2020), a large-scale knowledge graph of causal relationships between real-world concepts (extracted from Wikipedia and ClueWeb12 (Callan, 2012)). We perform semi-synthetic (CauseNet Graph with LLM generated sentences) and real-world experiments (Graph and sentences from CauseNet) using *GPT-4o* and Llama-3.1 8B. Analysis in Appendix A.

5. Discussion

Our work takes initial strides towards examining the success and failure of LLMs to reason causally on narratives that express causal events. Firstly, we find that LLMs rely heavily on **topological ordering**. Secondly, we find that LLMs rely on their **parametric knowledge** as a shortcut to infer causal relations. Finally, we examine the role of **causal complexity**, finding that LLM accuracy degrades as the narrative length increases and slightly worsens when narratives contain structures like colliders/forks. We also show explicit causal graph generation elicits reliable reasoning.

5.1. Limitations and Future Works

One limitation of our work is that we did not test for certain forms of reasoning such as counterfactual cases. Our analysis also has implications for algorithmic interventions to improve causal reasoning such as targeted fine-tuning.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ammanabrolu, P., Cheung, W., Broniec, W., and Riedl, M. O. Automated storytelling via causal, commonsense plot ordering, 2020. URL https://arxiv.org/abs/2009.00829.

Anthropic. Claude 3.5 sonnet. 2024. Available at: https://www.anthropic.com/news/claude-3-5-sonnet.

Ashwani, S., Hegde, K., Mannuru, N. R., Jindal, M., Sengar, D. S., Kathala, K. C. R., Banga, D., Jain, V., and Chadha, A. Cause and effect: Can large language models truly understand causality?, 2024. URL https://arxiv.org/abs/2402.18139.

Callan, J. The lemur project and its clueweb12 dataset. In *Invited talk at the SIGIR 2012 Workshop on Open-Source Information Retrieval*, 2012.

Dettki, H. M., Lake, B. M., Wu, C. M., and Rehder, B. Do large language models reason causally like us? even better?, 2025. URL https://arxiv.org/abs/2502.10215.

Frohberg, J. and Binder, F. Crass: A novel data set and benchmark to test counterfactual reasoning of large language models, 2022. URL https://arxiv.org/abs/2112.11941.

Gao, J., Ding, X., Qin, B., and Liu, T. Is chatgpt a good causal reasoner? a comprehensive evaluation, 2023. URL https://arxiv.org/abs/2305.07375.

Gordon, A., Kozareva, Z., and Roemmele, M. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In Agirre, E., Bos, J., Diab, M., Manandhar, S., Marton, Y., and Yuret, D. (eds.), **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pp. 394–398, Montréal, Canada, 7-8 June 2012. Association for Computational Linguistics. URL https://aclanthology.org/S12-1052/.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C. C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G. L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I. A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K. V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M. K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P. S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R. S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S. S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X. E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z. D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A., Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A.,

Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B. D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.-H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.-T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G. M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.-E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.-B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K. H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M. L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M. J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N. P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S. J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S. C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V. S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V. T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., and Ma, Z. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Heindorf, S., Scholten, Y., Wachsmuth, H., Ngonga Ngomo, A.-C., and Potthast, M. Causenet: Towards a causality graph extracted from the web. In *ACM international conference on Information & Knowledge Management*, 2020.

Ho, M., Sharma, A., Chang, J., Saxon, M., Levy, S., Lu, Y., and Wang, W. Y. Wikiwhy: Answering and explaining cause-and-effect questions. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=vaxnu-Utr4l.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., LYU, Z., Blin, K., Adauto, F. G., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=e2wtjx0Yqu.

Jin, Z., Liu, J., LYU, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M. T., and Schölkopf, B. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vqIH0ObdqL.

Joshi, A., Ahmad, A., and Modi, A. Cold: Causal reasoning in closed daily activities, 2024a. URL https://arxiv.org/abs/2411.19500.

Joshi, N., Saparov, A., Wang, Y., and He, H. Llms are prone to fallacies in causal inference, 2024b. URL https://arxiv.org/abs/2406.12158.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Li, B., Martin, L. J., and Callison-Burch, C. CIS^2: A simplified commonsense inference evaluation for story prose, 2022. URL https://arxiv.org/abs/2202.07880.

Li, J., Yu, L., and Ettinger, A. Counterfactual reasoning: Testing language models' understanding of hypothetical scenarios, 2023. URL https://arxiv.org/abs/2305.16572.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Tan, F. A., Hürriyetoğlu, A., Caselli, T., Oostdijk, N., Nomoto, T., Hettiarachchi, H., Ameer, I., Uca, O., Liza, F. F., and Hu, T. The causal news corpus: Annotating causal relations in event sentences from news. In *Language Resources and Evaluation Conference*. European Language Resources Association, 2022. URL https://aclanthology.org/2022.lrec-1.246.

Tian, Y., krishna Sridhar, A., and Peng, N. Hypogen: Hyperbole generation with commonsense and counterfactual knowledge, 2021. URL https://arxiv.org/abs/2109.05097.

Wang, Z., Do, Q. V., Zhang, H., Zhang, J., Wang, W., Fang, T., Song, Y., Wong, G., and See, S. COLA: Contextualized commonsense causal reasoning from the causal inference perspective. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5253–5271, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.288. URL https://aclanthology.org/2023.acl-long.288/.

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal, 2023. URL https://arxiv.org/abs/2308.13067.

Zhang, C., Bauer, S., Bennett, P., Gao, J., Gong, W., Hilmkil, A., Jennings, J., Ma, C., Minka, T., Pawlowski, N., et al. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*, 2023.

Technical Appendices and Supplementary Material

A. Semi-Synthetic and Real-World Experiments

In this section, we extend our analysis to narratives involving real-world causal graphs from *CauseNet* (Heindorf et al., 2020), a large-scale knowledge graph of (claimed) causal relationships between real-world concepts. We perform experiments using the *GPT-4o* (OpenAI et al., 2024) and Llama-3.1 8B models for our experiments. We concentrate our analysis on the same factors (positional biases and parametric knowledge consistency) as explored in the semi-synthetic settings.

A.1. Experimental Setting

The *CauseNet* dataset can be represented as a collection of D tuples $\{(C_i, E_i, \mathbf{S}_i\}_{i=1}^{D}$, where C_i denotes the cause (e.g., fatigue), E_i denotes the effect (e.g., accidents), and $\mathbf{S}_i$ is a set of sentences (extracted from Wikipedia and ClueWeb12 (Callan, 2012)) that entail a causal relationship from C_i to E_i. We retrieve causal chain graphs $V_1 \to V_2 \to \ldots \to V_N$ of various lengths, where each causal relation $V_i \to V_{i+1}$ is from *CauseNet* and verbalize these chains as narratives in the following ways:

Semi-synthetic narratives. In this setting, we use real causal graphs from *CauseNet* but synthetically verbalize them via the LLM. In particular, we prompt the LLM to generate sentences for each edge ($V_i \to V_{i+1}$) in the causal graph, while ensuring the sensibility of the entire narrative. For eammple, the following is a narrative for the chain *fatigue* $\to$ *accidents* $\to$ *injury*:

> *Fatigue* can cloud judgment and slow reaction times, leading to an increase in *accidents* on the road. As a result, these *accidents* often lead to serious *injury* for those involved, highlighting the dangerous consequences of driving while fatigued.

Real-world narratives. For the real-world narratives, the sentence for each edge is chosen from the *CauseNet* dataset. To ensure that the narrative as a whole remains coherent, we prompt the LLM to ensure that the sentences for every pair of adjacent edges logically follow each other. For example, the following is the narrative for the causal chain *fatigue* $\to$ *accidents* $\to$ *injury*:

> Workers work long hours in mines and factories where *fatigue* and a lack of concentration can easily cause *accidents*. These *accidents* are the leading cause of *injury* in this country for people ages 1-34.

Additional examples of semi-synthetic and real-world narratives are presented in Appendix D.1 (the entire set of narratives used for our experiments is available in the linked code).

Prompting Strategies For simplicity, we limit the prompting techniques used to (see Appendix D.2 for the prompt templates): **Standard QA Prompting**, **Chain-of-Thought** and **Explicit Causal Graph Extraction**. We evaluate the accuracy for each pair of nodes (V_i, V_j) for the three prompting strategies on the semi-synthetic and real-world narratives.

A.2. Impact of Event Ordering and Chain Length

As described in the previous section, we verbalize each causal chain graph $V_1 \to V_2 \to \ldots \to V_N$ from *CauseNet* into a narrative in the forward and reverse topological order. In both the semi-synthetic (Fig. 4 left) and real-world narratives (Fig. 4 right), the *Forward Graph* strategy performs the best, with its accuracy remaining stable even as the chain length increases. We observe that *Forward Standard and CoT* outperforms *Reverse Standard and CoT*, with the *Reverse* accuracy declining substantially as the chain size gets large. We also see that in this regime, extracting the causal graph makes inference in the *Reverse* orientation competitive with inference in *Forward*.

A.3. Effect of Parametric Knowledge Consistency

Experiment Setup Next, we analyze the extent to which the LLM relies on its parametric knowledge to answer causal reasoning queries as opposed to the causal structure expressed in the narrative. For every pair of nodes (V_i, V_j) in the chain graphs, we elicit the parametric knowledge of the LLM by asking the LLM whether a causal effect between the two nodes

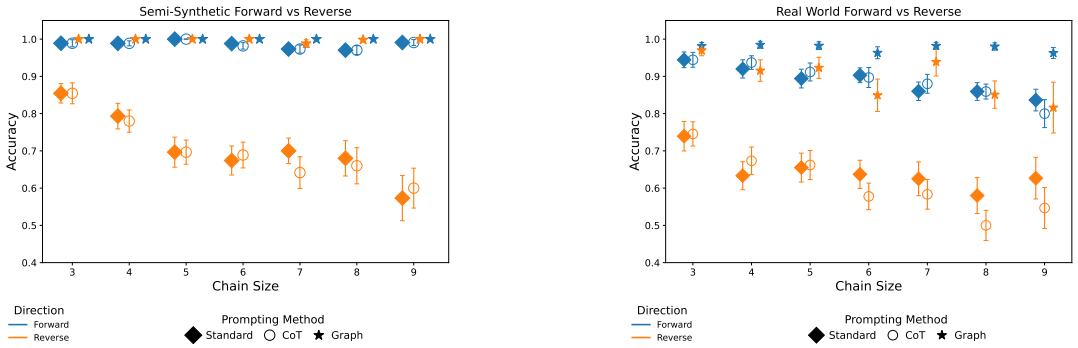


Figure 4: The accuracy of various prompting strategies (error bars denote 95% CIs). We observe that the accuracy is lower in the reverse direction (and tends to decay as the chains get longer).

	Standard	CoT	Graph
Semi-synthetic			
Without Conflict	99.8	99.6	99.9
With Conflict	67.2	73.1	98.7
Real-world			
Without Conflict	90.9	89.2	97.9
With Conflict	52.1	57.6	93.2

Table 1: The average accuracy across different narratives with the three prompting strategies partitioned by whether the cause-effect pairs conflict with the LLM's parametric knowledge (we omit the 95% CIs as they are smaller than 0.3).

would be atypical (see Appendix C.2 for the exact prompts utilized). Through these prompts, we identify cause and effect chains which contradict the model's parametric knowledge. For example, in a chain graph from our dataset, there is a path from *streambank erosion* to *higher prices*, but this contradicts the LLM's parametric knowledge since this causal effect may not typically exist in the real-world. In total, we find that roughly 5 percent of the relations in CauseNet violate the LLM's pretraining knowledge. We sampled narratives from CauseNet until we got 100 (of chain sizes between 3 and 9) narratives that contain relations that violate the LLM's pre-training knowledge and 100 that are consistent. These narratives are constructed in the *Forward* topological ordering to avoid confounding failure modes.

LLM Performance Suffers on Atypical Causal Relations We evaluate the three prompting strategies separately on the subsets of cause-and-effect pairs that are in agreement and in conflict with the parametric knowledge (see Table 1). We observe that when there is no conflict (i.e., the parametric knowledge agrees with the causality expressed in the narrative), the accuracies with and without CoT are greater than 90%. However, when the parametric knowledge conflicts with the narrative's causality, the accuracy is significantly lower, even with CoT. This suggests that when asked to reason about cause and effect in a narrative, the LLM seems to rely heavily on its parametric knowledge and is unable to grasp the specific causal chains expressed in the narrative itself (despite the causal chains as a whole being realistic).

Explicit Causal Graph Extraction Avoids Shortcuts Interestingly, when using extracted graph for performing causal reasoning, the performance is very high, both with and without conflicts. This is likely because when asked to extract the graph from the narrative, the LLM pays more attention to the entire narrative as opposed to when directly queried on a cause-and-effect pair (where the LLM defaults to its parametric knowledge). These results show that even when the LLM constructs a reasonably good causal chain graph, the LLM does not leverage this graph when queried directly about the causal effects in the narrative (even with CoT), further highlighting the advantage of extracting the causal graph directly.

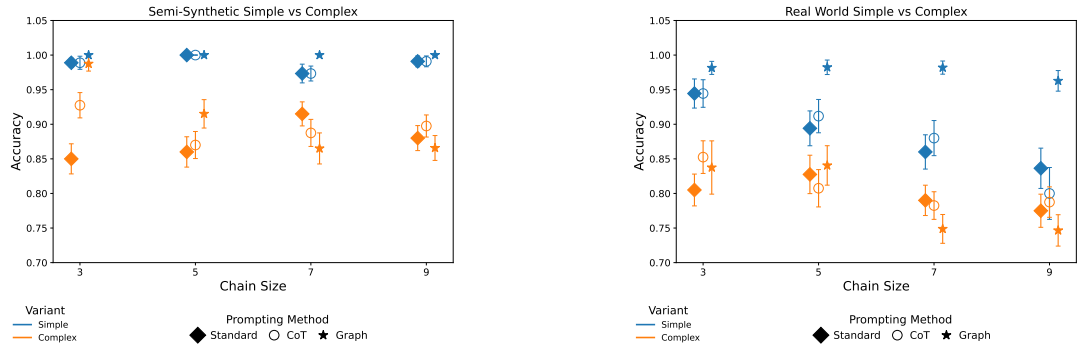


Figure 5: GPT-4o accuracy on narratives generated from Complex graphs as opposed to Simple chain graphs for semi-synthetic narratives (left) and real-world narratives (right). 95 % CI is shown.

A.4. Narrative Complexity

We can see from Figure 4 that LLM performance degrades with narrative length, especially when a failure mode is present. We furthermore experimented with complex narratives with causal graphs containing forks and colliders (full graph and narrative creation algorithm in Appendix C.3). We can see in Figure 5, that in both the semi-synthetic and real-world settings that complex narratives (with colliders and forks) perform worse than simple narratives that have a causal chain graph as the ground truth. This gap ,while clear and noticeable, isn't as stark as failure from parametric knowledge conflict (Table 1) or topological ordering (Figure 4). We do furthermore note that this is one area where extracting an explicit causal graph does not seem to significantly improve performance.

B. Synthetic Data Experiments

B.1. Selected Synthetic Prompts

We use an LLM to generate the events *E*. From the events, we create a ground truth causal graph *G* which is used to structure and inform the narrative sequence and causality. *N* is the corresponding narrative created by the LLM from *G*. To evaluate the LLM's performance, we extract a causal graph, *G'*, from the narrative *N* as produced by the LLM, and compare it with the ground truth causal graph *G*. In this context, *n* refers to the number of events to generate, while *A* and *B* represent pairs of events queried for causal relationships. The task then becomes assessing whether event *A* causes event *B* . All prompts, data processing steps, and results are included in the attached code.

B.1.1. Topological Experiment - Generating Random Events (*E*)

"generate *n* random distinct events"

B.1.2. Parametric Experiment -Generating a Pair of Causal Events (*E*)

"generate a pair of events that cause each other. generate an event that causes another event, for example Cancer $\rightarrow$ Death or Obesity $\rightarrow$ Bad Heart Health. Make sure the event generated is not already in E "
This is repeated as many times as is necessary

B.1.3. Parametric Experiment - Generating a Pair of Anti-Causal Events (*E*)

"generate a pair of events that are anticausal (an event causing the opposite of the normal effect), for example the first event could be cancer and the second event could be a longer life because in reality, cancer causes a shorter life. Make sure the events generated are not already in E."
This is repeated as many times as is necessary

B.1.4. Forward Topological Narrative (*N*)

"Output a short narrative (use one sentence) that expresses the causal link [E1 $\rightarrow$ E2]. By causal link, we mean that the sentence should convey that E1 directly caused E2. In other words, it should be clear from the narrative that E2 would not have happened had E1 not happened. Ensure that the words [E1, E2] are present in the new sentence and E1 appears before E2. Only output the new sentence."
Repeat for all causal/anti-causal links

B.1.5. Reverse Topological Narrative (*N*)

"Output a short narrative (use one sentence) that expresses the causal link [E1 $\rightarrow$ E2]. By causal link, we mean that the sentence should convey that E1 directly caused E2. In other words, it should be clear from the narrative that E2 would not have happened had E1 not happened. Ensure that the words [E1, E2] are present in the new sentence and E2 appears before E1. Only output the new sentence."
Repeat for all causal/anti-causal links

B.1.6. Standard Prompt

"Use this narrative *N* as context. Did *A* cause *B*? Output your answer with $< answer > Yes/No < /answer >$. The cause can be direct or indirect."

B.1.7. Chain of Thought Prompt

"Use this narrative *N* as context. Did *A* cause *B*? Do step by step reasoning. Then output your answer with $< answer > Yes/No < /answer >$. The cause can be direct or indirect."

B.1.8. In-Context Prompt

"Use this narrative *N* as context. Did *A* cause *B*? Output your answer with $< answer > Yes/No < /answer >$. The cause can be direct or indirect. An example narrative would be: Rains leads to plants growing. This then causes increased oxygen

in the atmosphere. A potential question would be does rain cause increased oxygen in the atmosphere? The answer would be Yes. Another example narrative would be: Increased oxygen in the atmosphere is because of plants growing. Plants grow because rain provides them essential nutrients. A potential question would be does rain cause increased oxygen in the atmosphere? The answer would be Yes. Another example narrative would be: Rain leads plants to grow. Plants growing causes less oxygen in the atmosphere. A potential question would be does rain cause less oxygen in the atmosphere? The answer would be Yes."

B.1.9. NARRATIVE + GRAPH PROMPT

"Use this narrative *N* and this causal ordering *G'* ((such that each item is a cause of every item after it, for example the first list item is a cause of the third, fourth, fifth items etc)) as context. Did *A* cause *B*? Output your answer with $< answer > Yes/No < /answer >$. The cause can be direct or indirect."

B.2. Parametric Graph Experiment

Let's call the graph of parametric knowledge P. We then take the odd indexed events (1st, 3rd etc) from P and place them in the first half of the causal ground truth graph G and the even indexed events (2nd, 4th etc) from P in the second half of G. This process is shown in Figure 6.

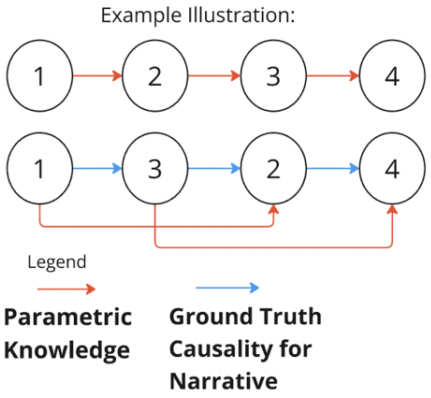


Figure 6: Example illustration (right) is of how G, the ground truth causality, is set up.

B.3. Complex Graph Creation

To generate a ground-truth causal graph *G* with rich structure (colliders, forks, and a spanning chain), for each choice of size n we perform the following algorithm:

1. **Node sampling.** Draw n distinct events
$$\{E_1, E_2, \ldots, E_n\} \subset \mathcal{E}$$
uniformly at random without replacement.

2. **Determine motif counts.** (for $n \geq 4$)
$$k_{\max} = \lfloor n/2 \rfloor, \quad k_{\text{tot}} \sim \text{Uniform}\big(2, k_{\max}\big),$$
$$k_{\text{col}} \sim \text{Uniform}\big(1, k_{\text{tot}} - 1\big), \qquad k_{\text{fork}} = k_{\text{tot}} - k_{\text{col}}.$$

3. **Collider creation.** Repeat k_{col} times:
   (a) Select two distinct "parent" nodes p_1, p_2 from those not yet used in any motif.
   (b) Select a "child" node c that is neither p_1 nor p_2 and not yet used as a child.

(c) Add edges

$$p_1 \;\to\; c \quad \text{and} \quad p_2 \;\to\; c\,,$$

thereby forming a collider at c.

4. **Fork creation.** Repeat k_{fork} times:
   (a) Select a "parent" node p from those not yet used.
   (b) Select two distinct "child" nodes c_1, c_2 from the remaining unused nodes.
   (c) Add edges

$$p \;\to\; c_1 \quad \text{and} \quad p \;\to\; c_2\,,$$

forming a fork with shared parent p.

5. **Chain-connect remaining nodes.** Let $\mathcal{R}$ be the set of nodes not yet involved in any collider or fork.
   (a) Order $\mathcal{R} = \{r_1, \ldots, r_m\}$ arbitrarily, then add chain edges

$$r_1 \to r_2, \;\; r_2 \to r_3, \;\; \ldots, \;\; r_{m-1} \to r_m.$$

   (b) To ensure the entire graph is connected, choose one node u from among the previously used nodes (if any) and add

$$u \;\to\; r_1\,.$$

C. Real-world Causal Graphs

C.1. Prompt templates for narrative generation

Recall that we have a ground truth causal chain graph of the form $V_1 \to V_2 \to \ldots \to V_N$ from *CauseNet* that we need to verbalize into a coherent narrative. For the semi-synthetic narratives, we use the LLM (GPT-4o) to do so one edge at a time, while ensuring that the newly verbalized edge logically follows the previous one. The following is the prompt template for generating the narratives in the topological order of the graph:

> Output a short narrative (use one or two sentences) that expresses the causal link [$V_i \to V_{i+1}$] and logically follows this narrative:
>
> { Narrative for the previous edge $V_{i-1} \to V_i$}.
>
> Ensure that the combined sentences convey the causal chain [$V_{i-1} \to V_i \to V_{i+1}$] and that the words [$V_i$, V_{i+1}] are present. Only output the newly generated narrative.

Similarly, we generate narratives in the reverse topological order of the graph by verbalizing edges in the reverse direction with the following prompt template:

> Output a short narrative (use one or two sentences) that expresses the causal link [$V_i \to V_{i+1}$] and logically follows this narrative:
>
> { Narrative for the previous edge $V_{i+1} \to V_{i+2}$}.
>
> Ensure that the combined sentences convey the causal chain [$V_i \to V_{i+1} \to V_{i+2}$] and that the words [$V_i$, V_{i+1}] are present. Only output the newly generated narrative.

For generating real-world narratives, for each edge $V_i \to V_j$, we use the set of sentences from *CauseNet*. Each edge in *CauseNet* is linked to multiple sentences from various sources. Picking a sentence for each edge at random and concatenating them does not always lead to sensible narratives. To improve the quality of narratives, we use the following prompt to concatenate sentences for adjacent edges:

> Consider the following sentences.
>
> { Sentence for edge $V_i \to V_{i+1}$ }. { Sentence for edge $V_{i+1} \to V_{i+2}$ }.
>
> Do the sentences logically follow each other and express the causal chain [$V_i \to V_{i+1} \to V_{i+2}$]? Answer with Yes or No.

For verbalizing narratives in the topological order, for a given graph $V_1 \to V_2 \to \ldots \to V_N$, we only use sentences such that the above prompt returns *Yes* for every pair of adjacent edges $V_i \to V_{i+1} \to V_{i+2}$. This ensures that the narrative as a whole remains coherent and conveys the entire causal chain graph. We use a similar prompting strategy to verbalize narratives in the reverse topological order.

C.2. Eliciting Parametric Knowledge

We ask the LLM "Does V_i typically have a causal (indirect or direct) effect on V_j?" and "Would it be atypical if V_i had a (indirect or direct) causal effect on V_j?". If the LLM answers "No" and "Yes" to those respective questions, we would consider a causal relationship between V_i and V_j to contradict the LLM's prior knowledge that it learned from its pretraining corpora.

C.3. Semi-Synthetic and Real-World Complex Graph Algorithm

Let $\mathcal{M} = \{(u, v)\}$ be the set of real-world causal edges from CauseNet. For each target size $n \in \{3, \ldots, 9\}$, we:

1. **Load CauseNet.**

$$\mathcal{M} \;=\; \big\{(u, v) \mid u \to v \text{ in CauseNet}\big\}.$$

2. **Extract collider and fork motifs.**

$$\mathsf{Colliders} = \{(p_1, p_2, c) \mid (p_1, c) \in \mathcal{M},\ (p_2, c) \in \mathcal{M},\ p_1 \neq p_2\},$$
$$\mathsf{Forks} = \{(r, c_1, c_2) \mid (r, c_1) \in \mathcal{M},\ (r, c_2) \in \mathcal{M},\ c_1 \neq c_2\}.$$

3. **Determine motif counts.**

$$\text{If } n = 3, \quad (k_{\text{col}}, k_{\text{fork}}) = \begin{cases} (1, 0) & \text{w.p. } 0.5, \\ (0, 1) & \text{w.p. } 0.5. \end{cases}$$

(for $n \geq 4$)

$$k_{\max} \;=\; \lfloor n/2 \rfloor, \quad k_{\text{tot}} \sim \text{Uniform}\big(2,\, k_{\max}\big),$$
$$k_{\text{col}} \sim \text{Uniform}\big(1,\, k_{\text{tot}} - 1\big), \qquad k_{\text{fork}} = k_{\text{tot}} - k_{\text{col}}.$$

4. **Select motifs.**
   - Sample k_{col} distinct triples from Colliders.
   - Sample k_{fork} distinct triples from Forks.

   Let S be the union of all nodes appearing in these sampled triples.

5. **Pad or trim to size n.**
   - If $|S| > n$, uniformly subsample n nodes from S.
   - If $|S| < n$, add random "seed" nodes (not already in S) until $|S| = n$.

6. **Build ground-truth edges $\mathcal{G} \subseteq S \times S$.**
   (a) *Colliders:* for each (p_1, p_2, c) chosen, add $p_1 \to c$ and $p_2 \to c$.
   (b) *Forks:* for each (r, c_1, c_2), add $r \to c_1$ and $r \to c_2$.
   (c) *Chains:* for any remaining $(u, v) \in S \times S$ with $(u, v) \in \mathcal{M}$ and neither u nor v used in the above, add $u \to v$ to ensure connectivity.

7. **Narrative generation.** For each $(u \to v) \in \mathcal{G}$:
   For the semi-synthetic case - prompt the LLM to generate a sentence linking u to v using the forward topological ordering prompt.
   For the real-world case: Find a causal sentence linking u and v in the Cause-Net database

D. Real-World Complex Graph Creation

D.1. Additional examples of the generated narratives

D.1.1. SEMI-SYNTHETIC NARRATIVES

Below, we present some examples of semi-synthetic narratives in the forward and reverse directions.

The narrative in the forward direction for the chain *higher prices* → *reduced demand* → *lower prices*:

> As *higher prices* swept through the market, consumers began to tighten their budgets, leading to a noticeable *reduction in demand* for many goods. As a result of the *reduced demand*, suppliers were forced to *lower prices* in order to attract buyers back to the market.

The narrative in the reverse order for the causal chain *bankruptcy* → *bad credit* → *rejection* → *anger*:

> The sting of rejection ignited a fire within her, transforming her hurt into a seething anger that demanded to be felt. Her bad credit had led to the rejection she never saw coming, and now that sting of rejection ignited a fire within her, transforming her hurt into a seething anger that demanded to be felt. Her bankruptcy had left her with bad credit, a shadow that loomed over her every application, and now that sting of rejection ignited a fire within her, transforming her hurt into a seething anger that demanded to be felt.

The narrative in the reverse order for the causal chain *pollution* → *climate change* → *extreme weather events* → *natural disasters*:

> As extreme weather events become more frequent and severe, they increasingly lead to devastating natural disasters that disrupt communities and ecosystems alike. Climate change is driving the rise in extreme weather events, which in turn are causing unprecedented natural disasters that threaten the stability of communities and the health of ecosystems. Pollution is a major contributor to climate change, which is driving the rise in extreme weather events that threaten the stability of communities and the health of ecosystems.

D.1.2. REAL-WORLD NARRATIVES

Below, we present some examples of real-world narratives in the forward and reverse directions.

The narrative in the forward direction for the chain *higher prices* → *reduced demand* → *lower prices*:

> *Higher prices* generally lead to reduced demand. *Lower prices*, caused by *reduced demand* and increased competition for soybeans and corn, largely contributed to the overall bulk export decline.

The narrative in the reverse order for the causal chain *bankruptcy* → *bad credit* → *rejection* → *anger*:

> Embittered by an abusive upbringing, seething with resentment, irritated by others' failure to fulfill his or her superior sense of entitlement, and fuelled by anger resulting from rejection, the serial bully displays an obsessive, compulsive and self-gratifying urge to displace their uncontrolled aggression onto others whilst exhibiting an apparent lack of insight into their behavior and its effect on people around them. Bad credit normally leads to rejection but now with bad credit secured loan, you can avail the loan of your choice. For example, if you are applying for a loan, the lender may reject your application on the basis of bad credit caused by bankruptcy.

The narrative in the reverse order for the causal chain *pollution* → *climate change* → *extreme weather events* → *natural disasters*:

> In addition to forced migrations from rising seas, climate change is also increasing extreme weather events causing natural disasters such as cyclonic storms (hurricanes or typhoons), floods and droughts. This is worsened by extreme weather events caused by climate change. This landmark bill would jump start the economy by creating millions of new clean energy jobs, increase national security by reducing dependence on foreign oil, and preserve the planet by reducing the pollution that causes climate change.

D.2. Prompt templates for assessing causal reasoning

We use the following template for the Direct prompting strategy:

> Consider the following hypothetical narrative.
>
> {narrative}
>
> According to the hypothetical narrative, does {cause} have a (direct or indirect) causal effect on {effect}? Answer in Yes/No.

We use the following template for the Chain-of-Though (CoT) prompting strategy:

> Consider the following hypothetical narrative.
>
> {narrative}
>
> According to the hypothetical narrative, does {cause} have a (direct or indirect) causal effect on {effect}? Think step-by-step and end your answer with <answer>Yes/No</answer>.

We use the following template to extract a chain graph from the narrative:

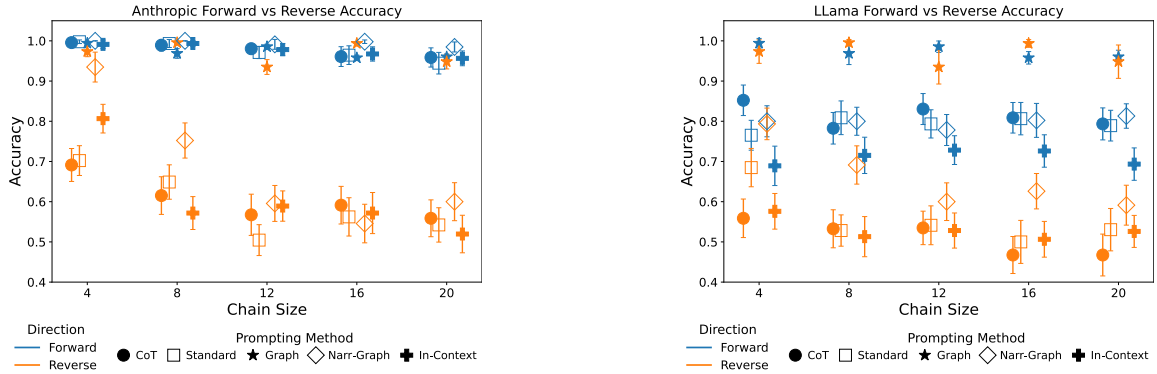
> Consider the following hypothetical narrative.
>
> {narrative}
>
> According to the hypothetical narrative, construct a causal chain graph using the following nodes: { nodes in random order }. Ensure that the graph contains all the given nodes and only output a single chain graph of the form <graph>node1 $\rightarrow$ node2 $\rightarrow$ node3 </graph>. Only output the graph between the <graph></graph>tags.

D.3. Necessary Compute

No pretraining was done so no GPUs were needed. We used cloud based API calls to pre-trained models like ChatGPT, Anthropic and Llama. We estimate that for the synthetic portion, our API calls to ChatGPT, Anthropic and LLama took 10 hours each. For the semi-synthetic and real-world portion, we had roughly 10 hours of API calls for ChatGPT and Llama each. So in total, roughly 50 hours of API usage. As the majority of the computational burden fell on cloud based API calls, no significant CPU resources are required either.

E. Additional Results - Synthetic Data

E.1. Forward vs Reverse Experiments Anthropic and LLama

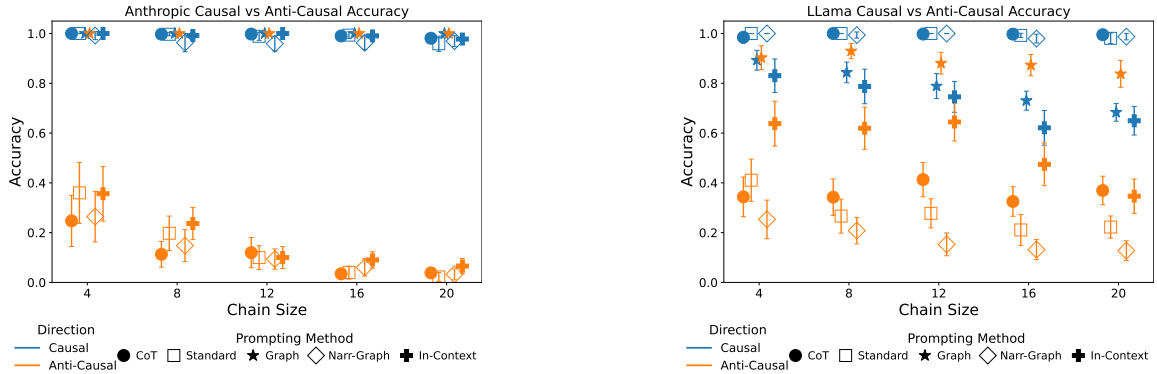


(a) Anthropic Claude 3.5 Sonnet

(b) LLama 3.1 8B

Figure 7: (a) Anthropic Claude 3.5 Sonnet and (b) LLama 3.1 8B Test of the LLM's ability to reason on narratives written in the Forward and Reverse topological orientations. Chain size is the number of nodes in ground truth G. The "Graph" prompting method uses only the extracted graph G' to reason, "Narr-Graph" uses both the narrative and extracted graph, and "Standard, CoT, In-Context" all use only the narrative. Accuracy measures LLM answer agreement with G. The points in the graph are represented with a slight horizontal stagger around the relevant chain sizes (4,8,12 etc) for ease of visual understanding. We show a 95% CI.

E.2. Causal Vs Anti-Causal Experiments Anthropic and LLama

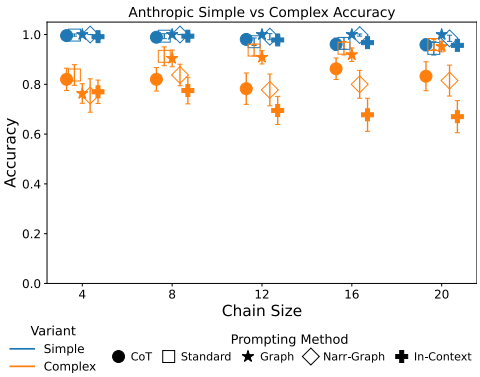


(a) Anthropic Claude 3.5 Sonnet

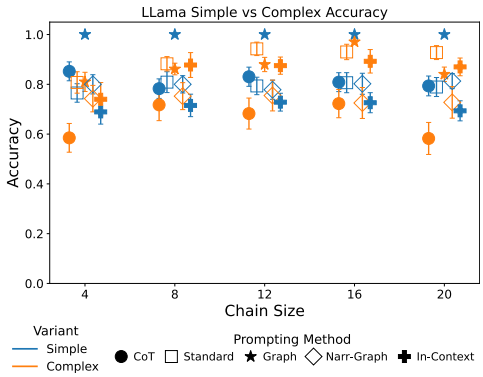
(b) LLama 3.1 8B

Figure 8: (a) Anthropic Claude 3.5 Sonnet and (b) LLama 3.1 8B Test of the LLM's ability to reason on narratives that agree with parametric knowledge (Causal) and disagree with parametric knowledge (Anti-Causal). 95 % CI is shown.

E.3. Complex vs Simple Graphs Anthropic and LLama



(a) Anthropic Claude 3.5 Sonnet



(b) LLama 3.1 8B

Figure 9: (a) Anthropic Claude 3.5 Sonnet and (b) LLama 3.1 8B Test of the LLM's ability to reason on narratives generated from Complex graphs as opposed to Simple chain graphs. 95 % CI is shown.

F. Additional results - Semi-Synthetic and Real World Data

F.1. Forward vs Reverse LLama

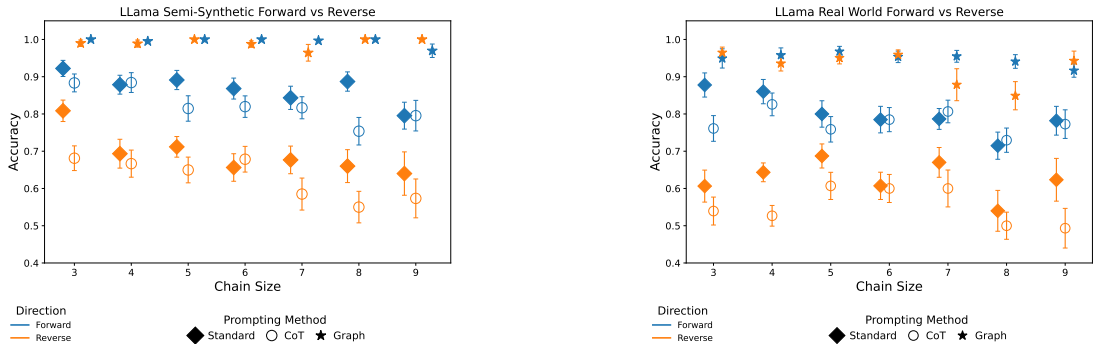


Figure 10: (LLama 3.1 8B) The accuracy of various prompting strategies (error bars denote 95% CIs) in the Semi-Synthetic and Real-World Regimes using CauseNet. We observe that the accuracy is lower in the reverse direction .

F.2. Parametric Experiment LLama

	Standard	CoT	Graph
Semi-synthetic			
Without Conflict	88.4	83.7	99.5
With Conflict	61.4	57.9	98.2
Real-world			
Without Conflict	81.6	79.2	95.1
With Conflict	48.8	49.9	93.2

Table 2: (LLama 3.1 8B) The average accuracy across different narratives with the three prompting strategies partitioned by whether the cause-effect pairs conflict with the LLM's parametric knowledge (we omit the 95% CIs as they are smaller than 0.3).

F.3. Simple vs Complex LLama

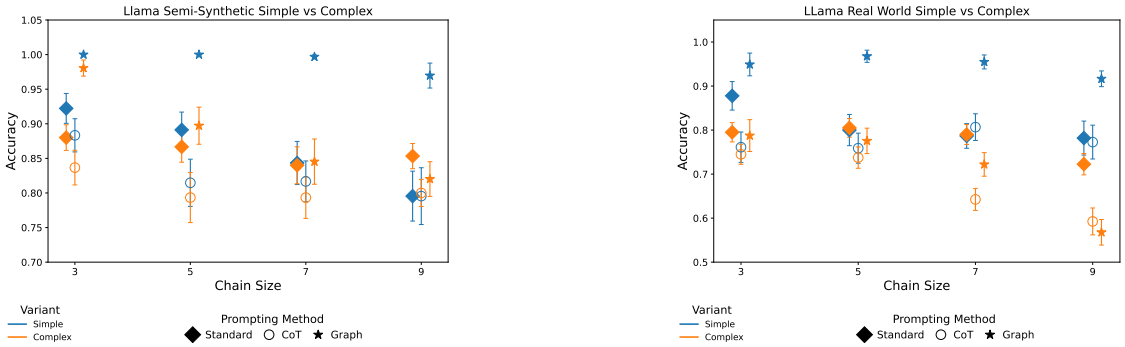


Figure 11: (LLama 3.1 8B) accuracy on narratives generated from Complex graphs as opposed to Simple chain graphs for semi-synthetic narratives (left) and real-world narratives (right). 95 % CI is shown.