

576 Technical Appendices and Supplementary Material

577 A Synthetic Data Experiments

578 A.1 Selected Synthetic Prompts

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

586 A.1.1 Topological Experiment - Generating Random Events (E)

587 "generate n random distinct events"

588 A.1.2 Parametric Experiment -Generating a Pair of Causal Events (E)

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

593 A.1.3 Parametric Experiment - Generating a Pair of Anti-Causal Events (E)

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

598 A.1.4 Forward Topological Narrative (N)

599 "Output a short narrative (use one sentence) that expresses the causal link $[E1 \rightarrow E2]$. By causal link,
600 we mean that the sentence should convey that $E1$ directly caused $E2$. In other words, it should be
601 clear from the narrative that $E2$ would not have happened had $E1$ not happened. Ensure that the words
602 $[E1, E2]$ are present in the new sentence and $E1$ appears before $E2$. Only output the new sentence."
603 Repeat for all causal/anti-causal links

604 A.1.5 Reverse Topological Narrative (N)

605 "Output a short narrative (use one sentence) that expresses the causal link $[E1 \rightarrow E2]$. By causal link,
606 we mean that the sentence should convey that $E1$ directly caused $E2$. In other words, it should be
607 clear from the narrative that $E2$ would not have happened had $E1$ not happened. Ensure that the words
608 $[E1, E2]$ are present in the new sentence and $E2$ appears before $E1$. Only output the new sentence."
609 Repeat for all causal/anti-causal links

610 A.1.6 Standard Prompt

611 "Use this narrative N as context. Did A cause B ? Output your answer with $\langle answer \rangle Yes/No \langle$
612 $/answer \rangle$. The cause can be direct or indirect."

613 A.1.7 Chain of Thought Prompt

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

A.1.8 In-Context Prompt

“Use this narrative N as context. Did A cause B ? Output your answer with $\langle answer \rangle Yes/No \langle /answer \rangle$. 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.”

A.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 $\langle answer \rangle Yes/No \langle /answer \rangle$. The cause can be direct or indirect.”

A.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.

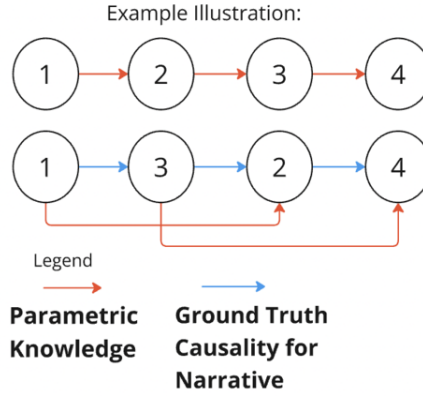


Figure 6: Example illustration (right) is of how G , the ground truth causality, is set up.

A.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

$$\{E_1, E_2, \dots, E_n\} \subset \mathcal{E}$$

uniformly at random without replacement.

2. **Determine motif counts.** (for $n \geq 4$)

$$k_{\max} = \lfloor n/2 \rfloor, \quad k_{\text{tot}} \sim \text{Uniform}(2, k_{\max}),$$

$$k_{\text{col}} \sim \text{Uniform}(1, k_{\text{tot}} - 1), \quad k_{\text{fork}} = k_{\text{tot}} - k_{\text{col}}.$$

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.

644 (b) Select a “child” node c that is neither p_1 nor p_2 and not yet used as a child.
645 (c) Add edges

$$p_1 \rightarrow c \quad \text{and} \quad p_2 \rightarrow c,$$

646 thereby forming a collider at c .
647 4. **Fork creation.** Repeat k_{fork} times:
648 (a) Select a “parent” node p from those not yet used.
649 (b) Select two distinct “child” nodes c_1, c_2 from the remaining unused nodes.
650 (c) Add edges

$$p \rightarrow c_1 \quad \text{and} \quad p \rightarrow c_2,$$

651 forming a fork with shared parent p .
652 5. **Chain-connect remaining nodes.** Let \mathcal{R} be the set of nodes not yet involved in any collider
653 or fork.
654 (a) Order $\mathcal{R} = \{r_1, \dots, r_m\}$ arbitrarily, then add chain edges

$$r_1 \rightarrow r_2, \quad r_2 \rightarrow r_3, \quad \dots, \quad r_{m-1} \rightarrow r_m.$$

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

$$u \rightarrow r_1.$$

657 B Real-world Causal Graphs

658 B.1 Prompt templates for narrative generation

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

664 Output a short narrative (use one or two sentences) that expresses the causal link
665 $[V_i \rightarrow V_{i+1}]$ and logically follows this narrative:
666 { Narrative for the previous edge $V_{i-1} \rightarrow V_i$ }.
667 Ensure that the combined sentences convey the causal chain $[V_{i-1} \rightarrow V_i \rightarrow$
668 $V_{i+1}]$ and that the words $[V_i, V_{i+1}]$ are present. Only output the newly generated
669 narrative.

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

672 Output a short narrative (use one or two sentences) that expresses the causal link
673 $[V_i \rightarrow V_{i+1}]$ and logically follows this narrative:
674 { Narrative for the previous edge $V_{i+1} \rightarrow V_{i+2}$ }.
675 Ensure that the combined sentences convey the causal chain $[V_i \rightarrow V_{i+1} \rightarrow$
676 $V_{i+2}]$ and that the words $[V_i, V_{i+1}]$ are present. Only output the newly generated
677 narrative.

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

683 Consider the following sentences.
684 { Sentence for edge $V_i \rightarrow V_{i+1}$ }. { Sentence for edge $V_{i+1} \rightarrow V_{i+2}$ }.
685 Do the sentences logically follow each other and express the causal chain $[V_i \rightarrow$
686 $V_{i+1} \rightarrow V_{i+2}]$? Answer with Yes or No.

For verbalizing narratives in the topological order, for a given graph $V_1 \rightarrow V_2 \rightarrow \dots \rightarrow V_N$, we only use sentences such that the above prompt returns *Yes* for every pair of adjacent edges $V_i \rightarrow V_{i+1} \rightarrow V_{i+2}$. 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.

B.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.

B.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 $n \in \{3, \dots, 9\}$, we:

1. Load CauseNet.

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

2. Extract collider and fork motifs.

$$\begin{aligned} \text{Colliders} &= \{(p_1, p_2, c) \mid (p_1, c) \in \mathcal{M}, (p_2, c) \in \mathcal{M}, p_1 \neq p_2\}, \\ \text{Forks} &= \{(r, c_1, c_2) \mid (r, c_1) \in \mathcal{M}, (r, c_2) \in \mathcal{M}, c_1 \neq c_2\}. \end{aligned}$$

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 $n \geq 4$)

$$k_{\text{max}} = \lfloor n/2 \rfloor, \quad k_{\text{tot}} \sim \text{Uniform}(2, k_{\text{max}}),$$

$$k_{\text{col}} \sim \text{Uniform}(1, k_{\text{tot}} - 1), \quad k_{\text{fork}} = k_{\text{tot}} - k_{\text{col}}.$$

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 $\mathcal{G} \subseteq S \times S$.

- Colliders*: for each (p_1, p_2, c) chosen, add $p_1 \rightarrow c$ and $p_2 \rightarrow c$.
- Forks*: for each (r, c_1, c_2) , add $r \rightarrow c_1$ and $r \rightarrow c_2$.
- 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 $u \rightarrow v$ to ensure connectivity.

7. Narrative generation. For each $(u \rightarrow v) \in \mathcal{G}$:

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

721 C Real-World Complex Graph Creation

722 C.1 Additional examples of the generated narratives

723 C.1.1 Semi-synthetic narratives

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

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

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

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

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

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

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

747 C.1.2 Real-world narratives

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

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

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

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

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

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

765 In addition to forced migrations from rising seas, climate change is also increasing
766 extreme weather events causing natural disasters such as cyclonic storms (hurri-
767 canes or typhoons), floods and droughts. This is worsened by extreme weather

768 events caused by climate change. This landmark bill would jump start the economy
769 by creating millions of new clean energy jobs, increase national security by reduc-
770 ing dependence on foreign oil, and preserve the planet by reducing the pollution
771 that causes climate change.

772 C.2 Prompt templates for assessing causal reasoning

773 We use the following template for the Direct prompting strategy:

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

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

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

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

785 Consider the following hypothetical narrative.
786 {narrative}
787 According to the hypothetical narrative, construct a causal chain graph using
788 the following nodes: { nodes in random order }. Ensure that the graph con-
789 tains all the given nodes and only output a single chain graph of the form
790 <graph>node1 → node2 → node3 </graph>. Only output the graph between
791 the <graph></graph>tags.

792 C.3 Necessary Compute

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

799 D Additional Results - Synthetic Data

800 D.1 Forward vs Reverse Experiments Anthropic and LLama

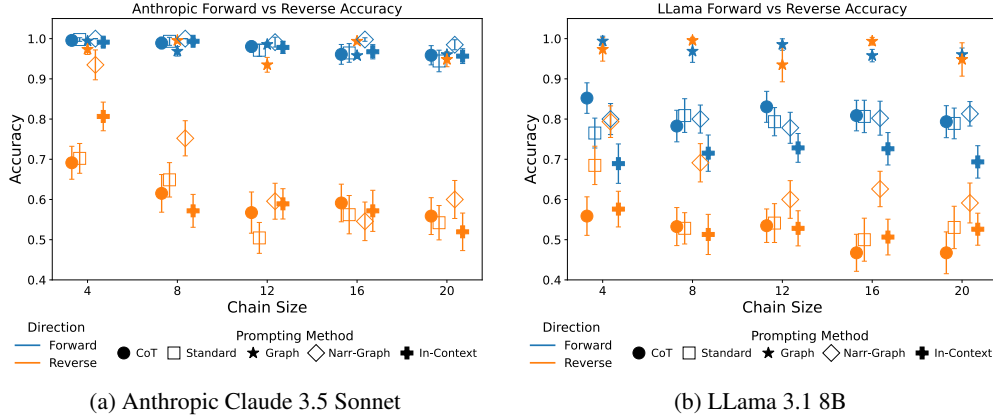


Figure 7: (a) Anthropic Claude 3.5 Sonnet and (b) LLama 3.1 8B Test of the LLM's ability to reason on narratives written in the Forward and Reverse topological orientations. Chain size is the number of nodes in ground truth G . The "Graph" prompting method uses only the extracted graph G' to reason, "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.

801 D.2 Causal Vs Anti-Causal Experiments Anthropic and LLama

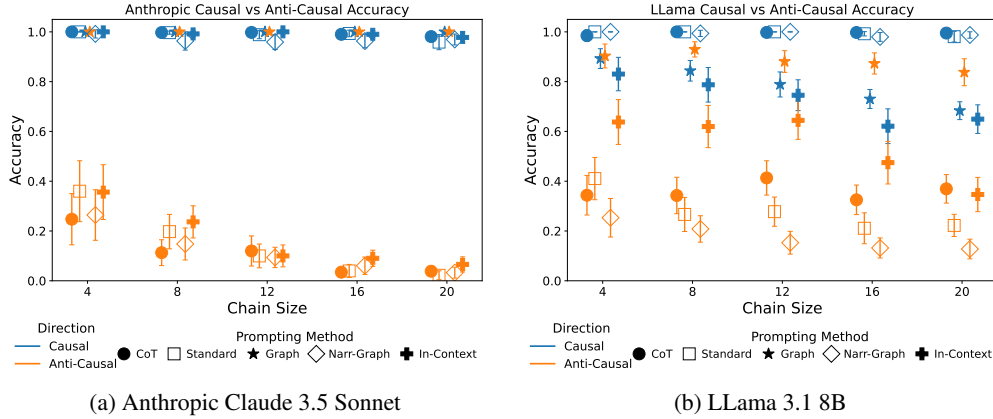


Figure 8: (a) Anthropic Claude 3.5 Sonnet and (b) LLama 3.1 8B Test of the LLM's ability to reason on narratives that agree with parametric knowledge (Causal) and disagree with parametric knowledge (Anti-Causal). 95 % CI is shown.

802 **D.3 Complex vs Simple Graphs Anthropic and LLama**

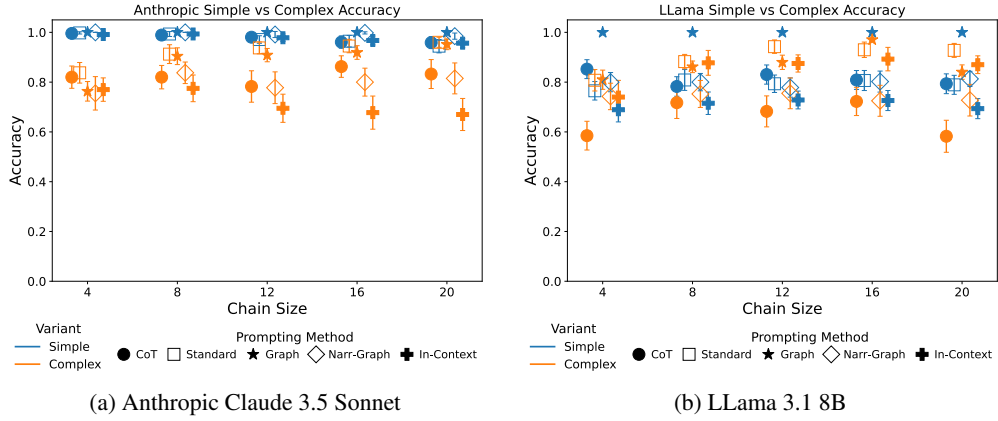


Figure 9: (a) Anthropic Claude 3.5 Sonnet and (b) LLama 3.1 8B Test of the LLM’s ability to reason on narratives generated from Complex graphs as opposed to Simple chain graphs. 95 % CI is shown.

803 E Additional results - Semi-Synthetic and Real World Data

804 E.1 Forward vs Reverse LLama

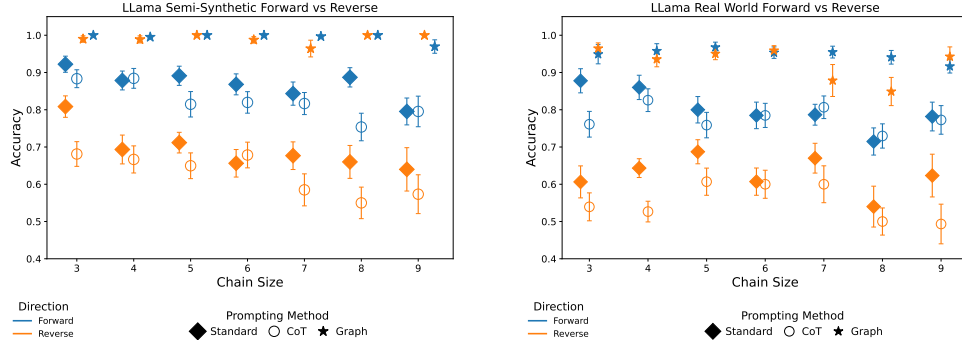


Figure 10: (LLama 3.1 8B) The accuracy of various prompting strategies (error bars denote 95% CIs) in the Semi-Synthetic and Real-World Regimes using CauseNet. We observe that the accuracy is lower in the reverse direction .

805 E.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).

806 E.3 Simple vs Complex LLama

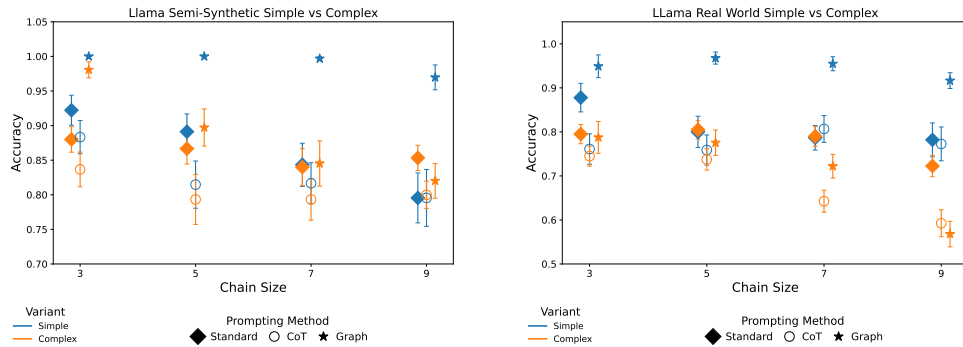


Figure 11: (LLama 3.1 8B) accuracy on narratives generated from Complex graphs as opposed to Simple chain graphs for semi-synthetic narratives (left) and real-world narratives (right). 95 % CI is shown.