NoisyCausal: A Benchmark for Evaluating Causal Reasoning Under Structured Noise

Anonymous ACL submission

Abstract

Causal reasoning in natural language requires identifying relevant variables, understanding their interactions, and reasoning about effects and interventions, often under noisy or ambiguous conditions. While large language models (LLMs) exhibit strong general reasoning abilities, they struggle to disentangle correlation from causation, particularly when observations are partially incorrect or irrelevant information is present. In this work, we introduce *NoisyCausal*, a new benchmark designed to evaluate causal reasoning under structured noise. Each instance is generated from a ground-truth causal graph and contextualized with a natural language scenario by injecting controllable forms of noise, such as irrelevant distractors, value perturbations, confounding, and partial observability. Moreover, we propose a modular reasoning framework that combines LLMs with explicit causal structure to address these challenges. Our method prompts the LLM to extract variables, construct a causal graph from context, and then reformulates the reasoning task as a structured prompt grounded in this graph. Rather than relying on statistical patterns alone, the LLM is guided by symbolic structure, enabling more interpretable and robust inference. Experimental results show that our method significantly outperforms standard prompting and reasoning baselines on *NoisyCausal*. Furthermore, it generalizes well to external benchmarks such as Cladder without task-specific tuning. Our findings highlight the importance of combining causal abstractions with language-driven reasoning to achieve faithful and robust causal understanding in LLMs.

1 Introduction

Causal reasoning is a fundamental component of human cognition and a critical capability for intelligent systems. It enables agents to go beyond surface-level statistical correlations and instead identify underlying mechanisms that govern observations, support counterfactual thinking, and generalize across domains. While recent advances in machine learning have led to impressive progress in language understanding, these systems often fall short when confronted with tasks that require structured reasoning about cause and effect, especially under uncertainty or noise.

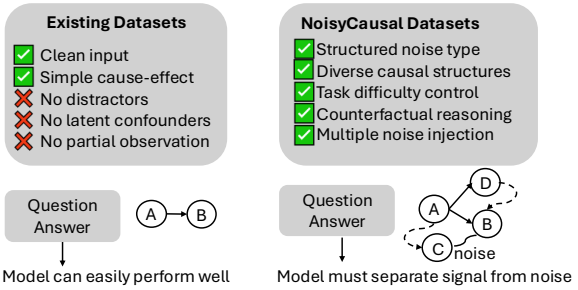

Figure 1: Existing causal reasoning datasets are often constructed with clean, idealized scenarios and direct causal questions, allowing large language models (LLMs) to succeed via pattern matching or shallow heuristics. As shown on the left, these settings typically lack noise, distractors, or confounding variables, which leads to overly optimistic estimates of reasoning ability. In contrast, *NoisyCausal* (right) introduces structured, diverse forms of noise—including value perturbations, irrelevant information, and latent confounders—to simulate more realistic and challenging environments.

Existing benchmarks ((Bondarenko et al., 2022; Jin et al., 2023)) for causal reasoning focus on clean, abstract scenarios where the causal structure is either simple or implicitly assumed. As a result, they fail to reflect the challenges faced in real-world reasoning, where observations are often noisy, incomplete, or confounded by irrelevant information. Moreover, current models ((Lasheras et al., 2025; Luo et al., 2025)) typically treat causal reasoning as a generic prediction problem, without explicitly modeling the dependencies or structure that underlie the task. This limits their interpretability and makes them fragile in the presence of spurious associations or misleading correlations.

To address these limitations, we introduce *Noisy-*

Causal, a new benchmark dataset specifically designed to evaluate causal reasoning under realistic, structured noise. Each instance in NoisyCausal is generated from a well-defined causal graph that encodes the actual dependencies among variables. A corresponding natural language scenario is constructed around this structure, such as a medical, mechanical, or social context. From this background, we generate variable values through a structural causal model (SCM), ensuring that clean samples are consistent with the causal rules. We then apply carefully designed perturbations—such as irrelevant variable injection, value flipping, partial observability, and latent confounding—to simulate noise and ambiguity in a controlled fashion. This allows us to construct samples of varying complexity and difficulty, enabling fine-grained analysis of model robustness, reasoning fidelity, and generalization.

This benchmark enables robust evaluation and motivates the design of a modular framework that can leverage its structured signals for improved reasoning. Complementing the dataset, we propose a causal reasoning framework that leverages large language models (LLMs) in a structured and interpretable way. Rather than relying on end-to-end black-box architectures or symbolic graph propagation, our model decomposes the reasoning process into modular steps. It begins by extracting task-relevant variables from the natural language prompt and constructing a causal graph that encodes directional relationships among them. The LLM is then prompted to answer causal questions with the observed variables and the textualized graph structure. This design allows the system to retain LLMs' flexibility and semantic richness while constraining the reasoning process using explicit causal structure. The causal graph serves as an interpretable intermediate and a way to guide the LLM's attention toward plausible reasoning paths.

Our approach brings several advantages. First, disentangling variable identification, structural modeling, and reasoning improves interpretability and modularity. Second, it supports flexible reasoning over noisy or incomplete inputs, as the LLM can adapt to uncertain observations while respecting causal constraints. Third, it enables counterfactual and interventional reasoning via simple changes to the input graph or variable values, making it suitable for applications that require hypotheticals or "what-if" analysis.

We evaluate our method on the NoisyCausal benchmark and demonstrate that combining causal structure with LLM reasoning leads to improved performance over standard prompting baselines. Our results suggest that structured prompting grounded in causal graphs offers a promising direction for enabling LLMs to reason more faithfully and robustly.

In summary, our contributions are:

- We introduce *NoisyCausal*, a new benchmark for evaluating causal reasoning under structured noise, with explicit causal graphs, natural language scenarios, and controllable perturbations;
- We propose a modular framework that uses causal graphs to structure the reasoning process and delegates the final inference to a large language model, combining interpretability with linguistic flexibility;
- We provide empirical evidence that our method improves performance and robustness across various reasoning tasks compared to baseline LLM prompting approaches.

2 Related Work

Causal Reasoning Benchmarks. Evaluating models on causal reasoning has been an active research area across NLP, vision, and knowledge representation. Early work such as COPA ((Roemmele et al., 2011)) focuses on binary selection between plausible causes and effects, emphasizing shallow commonsense plausibility. Other datasets like Event2Mind ((Rashkin et al., 2018)) and ATOMIC ((Sap et al., 2019)) extend this direction by modeling causal commonsense in social and narrative contexts. However, these benchmarks generally rely on implicit structure, and models are not required to explicitly recover or reason over a formal causal graph.

More recent datasets have attempted to bring structural causal modeling into the evaluation loop. CausalQA ((Bondarenko et al., 2022)) introduces causal reasoning questions over tabular datasets, involving do-calculus and interventions. Cladder ((Jin et al., 2023)) provides an algorithm that can automatically generate causal reasoning questions. These efforts provide valuable insights but are either limited in domain diversity, lack textual grounding, or assume complete observability and clean environments. Most notably, they do not explicitly test how models handle structured noise

or spurious variables that may mislead correlation-based heuristics.

Reasoning with Large Language Models. Large language models (LLMs) such as GPT-3 ((Brown et al., 2020)), PaLM ((Chowdhery et al., 2023)), and GPT-4 (Achiam et al., 2023) have demonstrated remarkable emergent capabilities in reasoning, planning, and few-shot generalization. Techniques such as chain-of-thought prompting ((Wei et al., 2022)), tree-of-thought ((Yao et al., 2023a)), and ReAct ((Yao et al., 2023b)) have improved multi-step reasoning through intermediate generation. LLMs have also been used to assist with formal logic problems ((Pan et al., 2023)), math ((Zhou et al., 2023)), and tool use ((Yuan et al., 2024)). However, causal reasoning remains a uniquely difficult challenge for LLMs. Recent studies ((Jiang et al., 2023; Cheng et al., 2024)) show that LLMs frequently confuse correlation with causation and fail to reason accurately about interventions or counterfactuals, especially in noisy or ambiguous environments. In many cases, models default to learned statistical associations from pretraining corpora, rather than identifying mechanistic dependencies. For example, they may predict that "people who cough take medicine" without understanding that coughing is an effect of illness and not a cause of recovery. These limitations highlight the need for structural guidance in the reasoning process.

Causal Graph Modeling and Structured Inference. Explicit causal modeling has a long history in statistics and AI, with foundational work such as Pearl's structural causal models ((Pearl, 2009)) and the PC/FCI family of causal discovery algorithms ((Malinsky and Danks, 2018)). In machine learning, causal graphs have been integrated into representation learning ((Schölkopf et al., 2021)), counterfactual simulation ((Zuo et al., 2022)), and reinforcement learning ((Dasgupta et al., 2019)). Many recent approaches use graph neural networks (GNNs) to perform inference over learned or given causal graphs, such as CausalGNN ((Wang et al., 2022)) or structure-aware transformers ((Chen et al., 2022)).

3 Dataset Construction – The *NoisyCausal* Benchmark

To rigorously evaluate causal reasoning capabilities in language models, we construct *NoisyCausal*, a synthetic benchmark that couples symbolic causal structure with realistic natural language prompts and structured noise. The dataset generation pipeline is illustrated in Figure 2. It consists of five sequential stages: (1) causal graph sampling, (2) semantic grounding, (3) structural causal model (SCM) sampling, (4) structured noise injection, and (5) natural language question assembly.

Causal Graph Sampling We begin by generating a directed acyclic graph (DAG) $G = (V, E)$, where each node $v_i \in V$ represents an abstract variable and each edge $(v_i \to v_j) \in E$ denotes a direct causal influence. To encourage structural diversity, we vary graph size between 3 and 7 nodes and sample from multiple topological motifs such as chains, forks, colliders, and graphs with multiple converging parents. We use topological sorting to ensure acyclicity and discard any cyclic or disconnected graphs. Each generated graph forms the backbone of an eventual reasoning instance.

Semantic Grounding Once a graph is sampled, we assign each node a real-world semantic label based on predefined domains such as medicine, education, and economics. For example, a path like $A \to B \to C$ may be interpreted as "infection causes medication intake, which affects recovery." These labels are drawn from a curated vocabulary and adjusted to ensure internal coherence within the scenario. In addition, we assign metadata to each variable, including its type (binary, categorical, continuous), observability (observable or latent), and role (e.g., symptom, cause, mediator, or outcome). The whole graph is then embedded into a naturalistic background story, serving as context for downstream reasoning questions.

SCM Sampling Given a semantically grounded causal graph, we generate a consistent assignment of variable values by sampling from a structural causal model (SCM). Each node is associated with a structural equation f_i that determines its value based on its parents. The form of f_i is sampled from a class of logic-based or probabilistic rules: binary nodes may use logical conjunction/disjunction, categorical nodes may follow lookup tables with probabilistic outputs, and continuous nodes are defined using additive or threshold-based functions. Sampling proceeds in topological order to respect causality. For example, given a scenario where A represents infection,

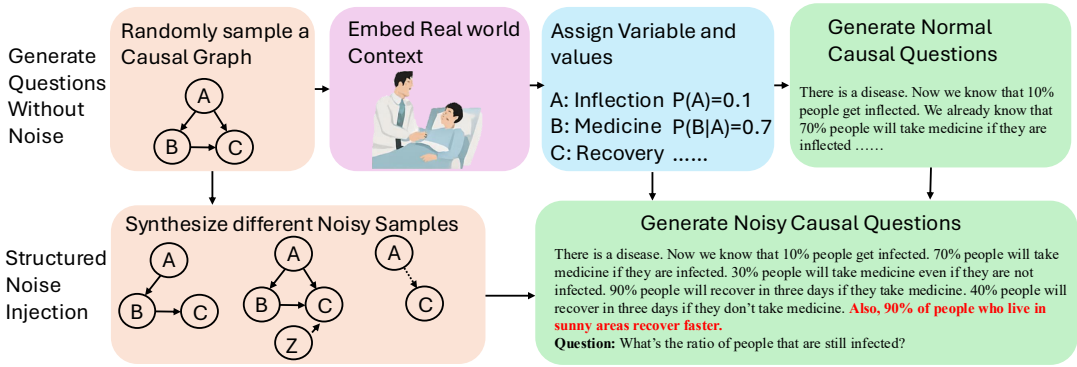

Figure 2: Overview of the *NoisyCausal* dataset construction pipeline. The process begins by randomly sampling a causal graph and embedding it into a realistic real-world scenario (e.g., modeling infection, medicine intake, and recovery). Each variable is assigned a semantic role and a probabilistic function. Clean, noise-free natural language questions are then generated from this grounded causal model. To simulate real-world uncertainty, we inject structured noise—such as irrelevant variables, value perturbations, latent confounders, and missing information—into both the graph and variable observations. This leads to the synthesis of multiple noisy instances, which are then used to formulate diverse, linguistically fluent reasoning questions. The resulting dataset enables rigorous evaluation of causal reasoning robustness in large language models across varying noise conditions.

B represents taking medicine, and C represents recovery, we may specify: $P(A = 1) = 0.1$ (10% infection rate), $P(B = 1|A = 1) = 0.7$, $P(B = 1|A = 0) = 0.3$ (higher medicine uptake among infected), $P(C = 1|A = 1) = 0.4$, $P(C = 1|B = 1) = 0.9$ (recovery probability conditioned on infection or treatment).

Such probabilistic dependencies instantiate clean observational traces that reflect the underlying causal mechanisms.

Structured Noise Injection To model real-world imperfections, we introduce structured noise types that distort variable values, the causal relationships, or the question context. Here, we organize the noise types from local variable-level perturbations to higher-level structural and linguistic disturbances. These include: **Value Perturbation (VP):** Randomly alter variable assignments. For example, flipping the value of C (recovery) from 1 to 0 even if the structural model predicts recovery; **Irrelevant Variable Injection (IV):** Introduce non-causal variables (e.g., "drinks tea") that correlate spuriously with outcomes like recovery, misleading LLMs; **Partial Masking (PM):** Hide values of selected observed variables (e.g., B), simulating missing information scenarios that force the model to reason under uncertainty; **Causal Swap (CS):** Swap values of causally linked nodes (e.g., B and C), which breaks conditional logic (e.g., makes it appear that recovery causes medicine intake); **Latent Confounders (CI):** Simulate hidden variables influencing multiple nodes (e.g., an unobserved variable Z influencing both A and C), introducing dependencies not shown in the observed graph; **Question Perturbation (QP):** Modify or corrupt the question such that its assumptions contradict the proper SCM (e.g., asking about outcomes under counterfactual settings that are logically inconsistent).

Each noise type is applied under a controlled probability distribution and can be composed with others to simulate increasingly challenging inference conditions. This allows for fine-grained control over task difficulty and supports comprehensive robustness evaluation. Clean and noisy variants of each sample are stored in parallel to enable contrastive evaluation and supervision.

Natural Language QA Assembly Finally, we transform each data instance into a natural language prompt and a reasoning question. The generated text explicitly or implicitly reflects the assigned probabilities and variable interactions defined in Step 3 for each scenario. For instance, a prompt may state: "Only 10% of people are infected. Among those infected, 70% take medicine. Medicine leads to recovery in 90% of cases. What is the likelihood that someone recovers if infected and takes medicine?" We support several task types, including interventional queries ("If the patient had not taken medicine, what would happen?"), counterfactual reasoning ("Had the infection not occurred, would recovery still be likely?"), More-

over, attributional analysis ("What caused the recovery?"). The background scenario is narrated fluently, and variable names are embedded into the text using handcrafted templates and paraphrasing engines to produce diverse and natural questions. Answers are derived based on the clean SCM and the causal graph, even in the presence of distractors or inconsistent evidence.

4 Causal Graph-Based Reasoning Model

To enable robust causal reasoning under noisy natural language conditions, we propose a framework that integrates lightweight causal graph construction with large language models (LLMs). Instead of relying on symbolic inference or message-passing algorithms, we use causal graphs as intermediate, interpretable structures that organize relevant variables and their directional relationships. The LLM then performs reasoning based on this structured representation.

The process begins with variable extraction, where the LLM identifies key entities (e.g., "fever", "medicine") and latent factors (e.g., "infection") from the background context. Next, we construct a task-specific causal graph using LLM-based prompting: for each variable pair, the model predicts causal directionality, forming a directed acyclic graph. Low-confidence or inconsistent edges are post-processed or deferred.

Once the graph is built, we compile a natural language prompt containing the background, observed variable values, and a textual description of the graph. The LLM uses this structured input to answer the reasoning question, producing answers that reflect both surface context and the underlying causal structure.

This architecture enhances explainability, supports counterfactual queries via prompt edits, and improves robustness under noise. By separating structure extraction from inference, our method encourages causal reasoning over correlation, bridging symbolic structure and LLM flexibility.

5 Experiments

5.1 Datasets

Our primary evaluation dataset is **NoisyCausal**, a synthetic benchmark containing 10,617 question-answer pairs with controlled causal structures and structured noise injections. Each instance includes a natural language background, an observation set (clean or perturbed), a causal reasoning question,

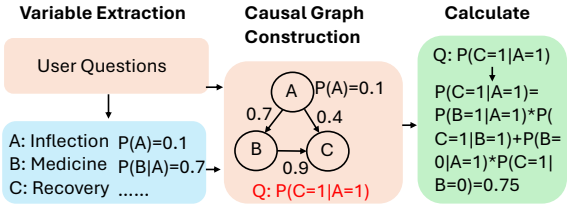

Figure 3: Overview of the graph-guided causal reasoning framework. The process begins by extracting task-relevant variables and constructing a question-specific causal graph via LLM prompting. This graph encodes directional relationships among variables (e.g., "infection" causes "medicine", which causes "recovery"). Observed variable values and graph structure are then converted into a structured natural language prompt, which is passed to an LLM for final inference.

and the ground-truth answer derived from a symbolic SCM. We evaluate models under six primary noise types (Value Perturbation, Irrelevant Variables, Causal Swap, Partial Masking, Confounders, and Question Perturbation) and multi-noise combinations of increasing difficulty. We also report results on the **Cladder** datasets for generalization.

5.2 Main Results

Table 1 presents the performance of various models on the *NoisyCausal* benchmark and the external Cladder dataset. Our graph-guided method achieves the highest accuracy across all conditions, including both clean and noisy settings. In the noise-free case, it reaches 80.7% accuracy, outperforming GPT-4 (62.8%) and Causal CoT (73.4%), indicating that structural guidance improves even standard reasoning. Under structured noise types such as value perturbation (VP), irrelevant variables (IV), and latent confounders (CI), traditional LLMs and CoT-style methods show noticeable degradation, with GPT-3.5 falling to as low as 43.8% on question perturbation (QP). In contrast, our model maintains stable performance across all perturbations (e.g., 77.3% on VP, 74.6% on IV, 73.5% on CI), highlighting its robustness to uncertainty. While advanced prompting strategies like ToT, ReAct, and Reflexion offer moderate gains over vanilla prompting, they are still sensitive to causal inconsistencies such as swaps or confounders. Finally, our model generalizes well to Cladder, achieving 82.3% accuracy without task-specific tuning, outperforming all baselines. These results demonstrate the benefit of combining causal structure with LLMs for robust and interpretable reasoning under noisy conditions.

Table 1: Accuracy (%) of different methods across datasets and noise types. The left block shows performance on our *NoisyCausal* benchmark under different noise types (VP: Value Perturbation, IV: Irrelevant Variable, etc.); the right block shows generalization to Cladder without task-specific tuning.

Model	**NoisyCausal**							**External Dataset**
	W/O Noise	VP	IV	CS	PM	CI	QP	Cladder
GPT-3.5((Brown et al., 2020))	57.9	54.2	49.6	44.7	52.0	47.1	43.8	52.2
GPT-4((Achiam et al., 2023))	62.8	60.4	56.3	50.1	59.8	54.0	48.5	62.0
LLaMa((Touvron et al., 2023))	52.4	50.1	45.7	41.3	48.9	42.6	40.7	44.0
Alpaca((Taori et al., 2023))	53.6	51.4	46.6	42.0	50.3	43.8	41.2	44.7
CoT((Wei et al., 2022))	65.5	63.5	58.9	55.4	60.8	56.2	53.3	–
ToT((Yao et al., 2023a))	68.3	65.1	61.7	57.2	63.4	59.5	54.6	–
ReAct((Yao et al., 2023b))	64.5	64.2	60.4	56.1	62.5	57.0	52.7	–
Reflexion((Shinn et al., 2023))	67.8	65.7	62.3	58.4	64.1	60.2	56.9	–
Causal CoT((Jin et al., 2023))	73.4	70.4	66.9	63.1	68.8	65.2	60.7	70.4
Ours (Graph-Guided)	**80.7**	**77.3**	**74.6**	**71.8**	**76.2**	**73.5**	**69.9**	**82.3**

Table 2: Ablation study results on the NoisyCausal benchmark (Accuracy %). Each variant removes or alters a specific component of our full model to evaluate its contribution. The final block shows the performance drop as more types of noise are combined.

Ablation Setting	**Accuracy (%)**	**vs. Full Model**
Full Model (Ours)	**80.68**	–
Graph Structure and Variable Semantics		
No Graph	65.32	-15.36
Random Graph	60.87	-19.81
Shuffled Variable Names	63.41	-17.27
Prompt Design and Language Robustness		
Edge-Only Prompt	74.34	-6.34
Natural Prompt	78.92	-1.76
Question Variation	80.12	-0.56
Cumulative Noise Combinations		
1 Noise Type	73.47	-7.21
2 Noise Types	67.32	-13.36
3 Noise Types	63.45	-17.23
4 Noise Types	60.51	-20.17
All 6 Noise Types	57.98	-22.70

5.3 Ablation Studies

To better understand the effectiveness and robustness of our proposed framework, we conduct a comprehensive set of ablation studies. These experiments are designed to isolate and evaluate the contributions of individual components within the system, including the causal graph, prompt design, and resilience under increasingly noisy conditions. As shown in Table 2, our analysis is structured into three categories.

Graph Structure and Variable Semantics. We first examine how the presence and correctness of causal structure influence reasoning. Removing the causal graph entirely (*No Graph*) results in significant performance degradation, indicating that the model falls back on shallow statistical heuristics. Replacing the ground-truth graph with a structurally equivalent but randomly rewired version (*Random Graph*) leads to even lower accuracy, confirming that not only the existence of structure but its correctness is crucial. We also test the impact of variable naming by shuffling variable names across the graph (*Shuffled Variable Names*), which disrupts semantic alignment with the question and causes further degradation. These results suggest that structural topology and linguistic consistency contribute substantially to model performance.

Prompt Design and Language Robustness. We evaluate the effect of different ways of presenting the graph on the model. Switching from natural language descriptions to a structured prompt (full model) format improves robustness slightly. Providing the graph as an edge list alone (*Edge-Only Prompt*) hurts performance, likely due to the lack of reasoning context. We also test *Question Variation* by paraphrasing the same prompt in different syntactic forms. The model maintains stable performance, suggesting it generalizes well to surface-level linguistic changes.

Cumulative Noise Combinations. To assess the model's robustness under compounding uncertainty, we introduce a new ablation protocol where noise types are randomly sampled and composed. Instead of isolating individual noise categories (e.g., Value Perturbation or Confounder Injection), we incrementally combine multiple types to simulate more realistic and challenging environments. We observe a graceful decline in performance as the

number of noise types increases: from 73.47% accuracy with one noise type to 57.98% when all six types are combined.

Summary Together, these ablation studies highlight several insights: (1) the presence and correctness of the causal graph structure play a central role in reasoning quality, (2) LLMs benefit from natural language contextualization of structure, not just formal edge lists, (3) robustness varies significantly across different types of noise, with distractors and confounders posing the most serious challenges, and (4) variable and graph extraction modules must be accurate and well-calibrated, as small mistakes can propagate downstream. These findings further support the design philosophy of using causal structure to constrain and guide LLM-based reasoning under uncertainty.

5.4 Graph Perturbation Sensitivity

In ablation studies, we noticed that the correctness of the causal graph structure plays a central role in reasoning quality. Hence, we want to explore the graph perturbation sensitivity further. Here, we perform a controlled structural perturbation analysis. Starting from the ground-truth graph, we introduce common types of errors: edge deletion (ED), where key connections are removed; false edge injection (FE), where spurious edges are added; and edge direction reversal (DR), which flips the cause-and-effect direction of existing edges. As shown in Figure 4, we evaluate three types of structural noise: Edge Deletion, False Edge Injection, and Direction Reversal, each applied with increasing numbers of errors (1–4). we find that while the model is relatively robust to minor edge deletion, it is significantly more sensitive to misleading or reversed causal links. In particular, direction reversal leads to the steepest performance drop, indicating that incorrect directional flow severely disrupts reasoning chains. These results emphasize the importance of accurate graph construction and motivate future research in validating or refining LLM-generated causal graphs before inference.

5.5 Structure Discovery Reliability

A key component of our framework is the ability to automatically construct causal graphs from natural language using a prompted large language model (LLM). While our downstream performance suggests that these structures are often useful, we seek to evaluate the accuracy and reliability of LLM-extracted graphs directly. To this end, we select 1000 NoisyCausal examples with known ground-truth graphs and prompt the LLM to extract relevant variables and causal edges. We compare the predicted graphs to the ground truth using edge-level Precision, Recall, and F1 score.

Table 3: Structure discovery performance across models, noise levels, and prompt designs. Reported: Edge-level Precision / Recall / F1 (%).

Condition	Precision	Recall	F1 Score
GPT-4 (clean, structured prompt)	87.1	83.5	85.2
GPT-4 (value perturbed)	81.3	78.9	80.1
GPT-4 (confounding + masking)	76.2	72.1	74.1
GPT-4 (clean, natural prompt)	82.5	77.4	79.8
LLaMA-2 (clean, structured prompt)	65.4	60.7	62.9
LLaMA-2 (value perturbed)	60.4	57.6	58.9
LLaMA-2 (confounding + masking)	52.3	49.6	50.9

We evaluate performance across three axes: (1) different noise levels (clean vs. perturbed vs. masked/confounded), (2) prompt styles (structured vs. natural language), and (3) model types (GPT-4 vs. LLaMA). Results in Table 3 show that GPT-4 achieves high accuracy in clean settings (F1 > 85) and remains robust under moderate noise, while smaller models like LLaMA degrade significantly. Structured prompts yield better extraction quality than open-ended instructions. These findings suggest that LLMs can recover causal structure with reasonable fidelity, but remain sensitive to input noise and prompt formulation—highlighting the need for better prompt calibration, graph validation, or hybrid symbolic support.

5.6 Error Typology Analysis

To gain deeper insights into the types of failures exhibited by our model, we conduct a qualitative and quantitative analysis of model errors across the NoisyCausal benchmark. Specifically, we manually inspect a representative subset of incorrect predictions and categorize them into distinct error types based on their underlying cause. This analysis allows us to better understand how different forms of reasoning breakdown occur and how they relate to specific challenges introduced by structured noise. We define the following error categories: **E1. Variable Identification Failure**: The model fails to recognize or track a key variable mentioned in the prompt (e.g., overlooks a latent confounder or misidentifies the target outcome); **E2. Causal Path Misinterpretation**: The model identifies variables but incorrectly infers the causal direction or reasoning chain (e.g., infers recovery

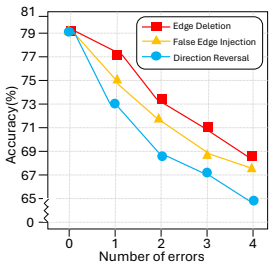

Figure 4: Impact of causal graph perturbations on reasoning accuracy.

Figure 5: Distribution of error types across a sample of 100 incorrect predictions.

→ medicine instead of medicine → recovery); **E3. Numerical Reasoning Error**: The model identifies the structure correctly but fails to compute or compare probabilities accurately, often due to surface-level heuristics; **E4. Distractor Overfitting**: The model is misled by irrelevant or spurious variables introduced as structured noise, such as "people who drink tea recover faster"; **E5. Question Misalignment**: The model misunderstands the logical intent of the question, especially in counterfactual or interventional settings.

We randomly sample 100 error cases from the validation set across all six noise types and assign each error to one of the five categories. The results are summarized in Figure 5

The most common failure mode involves incorrect causal path interpretation (E2), especially in the presence of swapped edges or latent confounding. This highlights the need for stronger graph validation or constraint mechanisms. A substantial portion of errors are also due to distractor overfitting (E4), suggesting that LLMs still struggle to distinguish spurious associations even when guided by structure. Interestingly, numerical reasoning errors (E3) are not limited to statistical estimation, but also occur when conflicting cues (e.g., base rates vs conditional probabilities) are present.

These findings underscore the multifaceted nature of causal reasoning errors and suggest that future improvements should not only target graph construction accuracy, but also incorporate better reasoning supervision, distractor suppression, and probabilistic calibration.

5.7 Belief-Inconsistent Perturbation

We design a new experiment to evaluate model robustness when exposed to misleading but structurally irrelevant contextual beliefs. Specifically, we introduce **Belief-Inconsistent Perturbation (BIP)**, where the prompt includes incorrect or commonly held misconceptions that contradict the underlying causal graph, while the actual variable relationships and numerical values remain unchanged. For example, the scenario may state "Many people believe that taking medicine increases the chance of infection," even though the true causal direction is `Infection` → `Medicine`. We inject such statements into 1000 clean Noisy-Causal instances and compare performance with the original versions. As shown in Table 4, models relying on surface-level reasoning (e.g., GPT-3.5, GPT-4) show notable accuracy drops (up to 6%), suggesting vulnerability to belief-level cues. In contrast, our graph-guided model degrades minimally, demonstrating that structural grounding improves resistance to contextual bias.

Table 4: Performance under Belief-Inconsistent Perturbation (BIP). Accuracy (%) on original and belief-perturbed questions.

Model	Original	+ BIP
GPT-3.5	57.9	51.2
GPT-4	62.8	56.7
Causal CoT	73.4	69.1
Ours (Graph-Guided)	**80.7**	**79.5**

6 Conclusion

We introduced NoisyCausal, a benchmark for evaluating LLMs' causal reasoning under structured noise. Each instance is grounded in a causal graph and perturbed with distractors, confounders, or inconsistent information to assess robustness and reasoning fidelity. To address this challenge, we proposed a modular graph-guided framework that extracts variables and constructs causal graphs from natural language, reformulating the task into structured prompts for more interpretable and noise-resilient inference. Experiments show our method consistently outperforms standard prompting and generalizes well to external datasets like Cladder. Ablation studies highlight the importance of accurate structure and prompt design. Looking forward, NoisyCausal can support future research in causally aware LLMs, including extensions to real-world and multimodal tasks, and improvements in efficiency via lightweight reasoning agents. Our findings demonstrate the value of combining symbolic structure with language-based reasoning for robust and faithful AI.

7 Limitations

While our framework shows strong performance under noisy conditions, several limitations remain. First, it relies on LLM-based variable extraction and graph construction, which may introduce cascading errors. Future work could incorporate confidence estimation or correction mechanisms to improve stability. Second, the current modular design lacks end-to-end optimization. Joint training of variable extractors and graph builders, or lightweight finetuning strategies, may enhance robustness in low-resource settings. Third, although *NoisyCausal* includes diverse synthetic noise, real-world generalization remains untested. Future benchmarks with real or multimodal data (e.g., tables, images) are needed to evaluate broader applicability. Finally, our method assumes a recoverable structure from the input. Open-domain or discourse-level tasks may require stronger integration with causal discovery or contrastive learning to infer latent structure.

AI Disclosure We used ChatGPT solely for grammar correction and language polishing. All research content, literature analysis, and writing were conducted independently by the authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. Causalqa: A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. 2022. Structure-aware transformer for graph representation learning. In *International conference on machine learning*, pages 3469–3489. PMLR.

Furui Cheng, Vilém Zouhar, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Interactive analysis of llms using meaningful counterfactuals. *arXiv preprint arXiv:2405.00708*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. 2019. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*.

Haitao Jiang, Lin Ge, Yuhe Gao, Jianian Wang, and Rui Song. 2023. Large language model for causal decision making. *arXiv preprint arXiv:2312.17122*.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Uriel Lasheras, Elioenai Alves, and Vladia Pinheiro. 2025. Interventional and counterfactual causal reasoning for llm-based ai agents: A dataset and evaluation in portuguese. *Procesamiento del Lenguaje Natural*, 74:363–382.

Hang Luo, Jian Zhang, and Chujun Li. 2025. Causal graphs meet thoughts: Enhancing complex reasoning in graph-augmented llms. *arXiv preprint arXiv:2501.14892*.

Daniel Malinsky and David Danks. 2018. Causal discovery algorithms: A practical guide. *Philosophy Compass*, 13(1):e12470.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. Logic-lm: Empowering large language models with symbolic solvers for faithful logical reasoning. *arXiv preprint arXiv:2305.12295*.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A Smith, and Yejin Choi. 2018. Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, Adam Sadilek, Srinivasan Venkatramanan, and Madhav Marathe. 2022. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 12191–12199.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Ren Kan, Dongsheng Li, and Deqing Yang. 2024. Easytool: Enhancing llm-based agents with concise tool instruction. *arXiv preprint arXiv:2401.06201*.

Aojun Zhou, Ke Wang, Zimu Lu, Weikang Shi, Sichun Luo, Zipeng Qin, Shaoqing Lu, Anya Jia, Linqi Song, Mingjie Zhan, and 1 others. 2023. Solving challenging math word problems using gpt-4 code interpreter with code-based self-verification. *arXiv preprint arXiv:2308.07921*.

Aoqi Zuo, Susan Wei, Tongliang Liu, Bo Han, Kun Zhang, and Mingming Gong. 2022. Counterfactual fairness with partially known causal graph. *Advances in Neural Information Processing Systems*, 35:1238–1252.

A Dataset Composition and Annotation.

Each *NoisyCausal* instance includes: (1) a natural language background and reasoning question, (2) a corresponding causal graph G (optionally provided to models), (3) both clean and noisy variable assignments, (4) the correct answer computed from the SCM, and (5) metadata describing the applied noise types and severity. These components allow *NoisyCausal* to support robust benchmarking of causal reasoning across structured noise conditions, enabling diagnostic insights into model behavior under uncertainty.

B Example of different noise injection

To supplement the formal definitions of structured noise in the main text, Table 5 presents concrete examples illustrating each perturbation type used in the NoisyCausal benchmark. For clarity, all examples are based on a common question template, with only one modification applied per instance. The modified elements are highlighted in red to isolate the injected noise.

Beyond showcasing what is perturbed, we also annotate which part of the causal reasoning pipeline is likely to be affected—such as conditional dependencies, causal directions, variable observability, or question alignment. These annotations help clarify how each type of noise challenges specific reasoning skills, including structure tracking, probabilistic reasoning, confounder adjustment, and counterfactual inference.

This appendix serves as a reference to understand better how different noise patterns are instantiated in our dataset and how they may induce distinct failure modes in language-based causal reasoning models.

C Faliure Cases With Graph Perturbation

Table 6 illustrates representative failure cases in causal reasoning arising from incorrect or incomplete causal graph structures. While the numerical calculations may appear sound in both examples, the underlying assumptions encoded in the graph are flawed—leading to misinterpretation of the causal mechanisms.

In the first case, the model incorrectly assumes that taking medicine (X) causes infection status (Z), inverting the actual causal direction. Although the resulting infection rate appears numerically correct, this structure violates causal semantics and will fail to generalize to interventions or counterfactual queries. In the second case, the model omits a critical confounder (S)—student ability—which influences both the treatment (tutoring, T) and the outcome (exam pass, E). Without adjusting for this hidden variable, the estimated treatment effect conflates correlation with causation, yielding biased results.

These examples demonstrate that correct causal reasoning depends not only on correct math, but on correctly structured causal assumptions. They further underscore the importance of evaluating both the output and the underlying causal model in language-based reasoning systems.

Noise Type	Modified Question (red = perturbed)	What is Perturbed	Potential Impact
VP: Value Perturbation	There is a disease. Now we know that 10% people get infected. 70% people will take medicine if they are infected. 30% people will take medicine even if they are not infected. 80% people will recover in three days if they take medicine. 40% people will recover in three days if they don't take medicine. Question: What's the ratio of people that are still infected?	Numerical value of $P(C \mid B)$ changed from 90% to 80%	Underestimates true effect of treatment, weakening causal inference.
IV: Irrelevant Variable Injection	There is a disease. Now we know that 10% people get infected. 70% people will take medicine if they are infected. 30% people will take medicine even if they are not infected. 90% people will recover in three days if they take medicine. 40% people will recover in three days if they don't take medicine. Also, 90% of people who live in sunny areas recover faster. Question: What's the ratio of people that are still infected?	Injection of irrelevant but highly correlated variable	Model may confuse correlation with causation, leading to spurious inference.
CS: Causal Swap	There is a disease. Now we know that 10% people get infected. People who recover are more likely to have taken medicine. 70% people will take medicine if they are infected. 30% people will take medicine even if they are not infected. 90% people will recover in three days if they take medicine. 40% people will recover in three days if they don't take medicine. Question: What's the ratio of people that are still infected?	Causal direction reversed ($C \rightarrow B$ instead of $B \rightarrow C$)	Misleads the model to reverse causal flow, impacting reasoning fidelity.
PM: Partial Masking	There is a disease. Now we know that 10% people get infected. (Missing) 70% people will take medicine if they are infected. 30% people will take medicine even if they are not infected. 90% people will recover in three days if they take medicine. 40% people will recover in three days if they don't take medicine. Question: What's the ratio of people that are still infected?	Removal of conditional dependency $P(B \mid A)$	Model must reason under uncertainty or fail gracefully.
CI: Confounder Injection	There is a disease. Now we know that 10% people get infected. 70% people will take medicine if they are infected. 30% people will take medicine even if they are not infected. 90% people will recover in three days if they take medicine. 40% people will recover in three days if they don't take medicine. Also, people with strong immune systems tend to both recover more quickly and are less likely to take medicine. Question: What's the ratio of people that are still infected?	Latent confounder Z added ($Z \rightarrow B$, $Z \rightarrow C$)	Introduces backdoor path, potentially biases causal reasoning.
QP: Question Perturbation	There is a disease. Now we know that 10% people get infected. 70% people will take medicine if they are infected. 30% people will take medicine even if they are not infected. 90% people will recover in three days if they take medicine. 40% people will recover in three days if they don't take medicine. Question: If people did not get infected but still took medicine, will they definitely recover?	Question contradicts causal structure	Forces model to answer counterfactual inconsistent with given SCM.

Table 5: Illustration of six structured noise types in NoisyCausal. Modified questions include targeted perturbations (highlighted in red) that affect different components of causal reasoning, such as structure, semantics, or inference assumptions.

Table 6: Failure Cases in Causal Reasoning Due to Incorrect Graph Structures

Error Type	**Original Question and Reasoning Error**	**Incorrect Causal Graph**	**Correct Causal Graph**
Graph Perturbation (X $\rightarrow$ Z)	10% of people are infected. 70% of infected take medicine; 30% of uninfected also take medicine. 90% recover in 3 days if treated; 40% recover otherwise. **Question:** What is the infection rate? **Incorrect reasoning:** 1. $P(X=1) = 0.7 \cdot 0.1 + 0.3 \cdot 0.9 = 0.34$ 2. $P(Z=1\|X=1) = \frac{0.7 \cdot 0.1}{0.34} \approx 0.2059$ 3. $P(Z=1\|X=0) = \frac{0.3 \cdot 0.1}{0.66} \approx 0.0455$ 4. Final: $P(Z=1) \approx 0.2059 \cdot 0.34 + 0.0455 \cdot 0.66 \approx 0.10$ **Flaw:** Wrong causal direction (X causes Z), leads to incorrect inversion logic. **Correct Answer** : 0.025	X $\rightarrow$ Z Z $\rightarrow$ Y X $\rightarrow$ Y	Z $\rightarrow$ X Z $\rightarrow$ Y X $\rightarrow$ Y
Graph Perturbation(Delete Edge S $\rightarrow$ E)	25% of students received tutoring. 80% of them passed; 50% passed without it. 40% of students have strong ability and always pass, regardless of tutoring. **Question:** What is the pass rate? **Incorrect reasoning:** Enumerate combinations: S=1,T=1: $0.4 \cdot 0.25 \cdot 0.8 = 0.08$ S=1,T=0: $0.4 \cdot 0.75 \cdot 1 = 0.30$ S=0,T=1: $0.6 \cdot 0.25 \cdot 0.8 = 0.12$ S=0,T=0: $0.6 \cdot 0.75 \cdot 0.5 = 0.225$ Total: $P(E=1) = 0.08 + 0.30 + 0.12 + 0.225 = 0.725$ **Flaw:** S is a confounder for both T and E, and not controlling for it leads to misleading observational estimates. **Correct Answer:** 0.745	T $\rightarrow$ E	S $\rightarrow$ E T $\rightarrow$ E