Causal Discovery and Intervention with Large Language Models

***Abstract*—Large Language Models (LLMs) demonstrate strong performance in reasoning, abstraction, and knowledge synthesis, yet their ability to learn and use causal structure remains poorly understood. This work investigates whether LLMs can (i) induce causal graphs from unstructured text, (ii) validate these graphs via intervention design, and (iii) perform counterfactual reasoning grounded in explicit Structural Causal Models (SCMs). We propose an automated framework that extracts causal variables and relations from text, constructs SCMs, and iteratively refines them using LLM-driven intervention proposals evaluated against simulated or empirical data. Our approach bridges natural language understanding with causal discovery and introduces a closed-loop causal reasoning pipeline. Experiments across synthetic causal benchmarks and real-world textual domains demonstrate that LLMs, when constrained by formal causal validation, can recover meaningful causal structure and support counterfactual inference beyond correlational heuristics.**

Keywords-causal discovery, large language models, structural causal models, counterfactual reasoning, intervention design

I. Introduction

Despite their impressive generative and reasoning capabilities, Large Language Models (LLMs) fundamentally operate via statistical pattern matching over large text corpora. This has raised skepticism regarding their ability to reason causally, particularly in settings requiring intervention, counterfactuals, and structural understanding rather than associative inference [1]. Causal reasoning, formalized through Structural Causal Models (SCMs) and Pearl's do-calculus [2], provides a rigorous framework for understanding why phenomena occur, predicting the effects of interventions, and answering counterfactual questions about what would have happened under alternative conditions.

Learning causal graphs from observational data is already a well-studied but challenging problem; learning them from natural language text introduces additional complexity due to linguistic ambiguity, implicit assumptions, under-specified mechanisms, and missing variables. Recent advances in LLM capabilities—including chain-of-thought reasoning [3], tool use [9], and in-context learning [8]—suggest that these models may serve as effective components within causal reasoning pipelines, even if they cannot be considered autonomous causal reasoners in isolation.

The central question of this paper is: Can LLMs learn causal graphs from text and use them for intervention-based counterfactual reasoning? We argue that while LLMs alone are insufficient causal reasoners, they can serve as powerful proposal engines within a formally grounded causal pipeline. The key insight is that LLMs excel at extracting structured representations from unstructured text and proposing plausible causal hypotheses, while formal causal methods provide the validation and inference machinery needed to ensure correctness.

Specifically, we contribute: (i) an automated framework for extracting causal variables and directed relations from unstructured text using LLM prompting with self-consistency validation; (ii) a method for constructing and parameterizing SCMs from LLM-extracted causal graphs; (iii) an iterative intervention design and validation loop that refines causal structure using LLM-proposed experiments evaluated against simulated or empirical data; and (iv) comprehensive experimental evaluation on synthetic benchmarks and three real-world textual domains demonstrating substantial improvements over baselines.

II. Related Work

A. Causal Discovery

Traditional causal discovery relies on conditional independence tests such as the PC and FCI algorithms [4], score-based methods like GES [5], or functional assumptions including Additive Noise Models and LiNGAM [6]. These approaches operate on structured numerical data and assume access to samples from the joint distribution over measured variables. They struggle with latent variables, semantic abstraction, and settings where data is presented in natural language rather than tabular form. Our work differs fundamentally in taking unstructured text as input rather than observational data.

B. Causal Representation Learning

Recent work explores learning disentangled causal variables from raw data such as images and temporal sequences [7], aiming to recover latent causal structure from high-dimensional observations. However, these methods typically assume access to paired interventions or environment resets, which are unavailable in most textual settings. They also do not address discovering causal structure from textual descriptions, where variables must first be identified, disambiguated, and mapped to a formal graph representation.

979-8-3315-6311-0/26/$31.00 ©2026 IEEE

C. Language Models and Reasoning

LLMs have demonstrated capabilities in logical reasoning, multi-step planning, and tool use [8], [9]. However, systematic evaluations under counterfactual perturbations reveal that much of this apparent reasoning relies on surface correlations and memorized patterns rather than genuine causal understanding [10]. Chain-of-thought prompting and self-consistency decoding have improved performance on reasoning benchmarks, but these gains are fragile and do not reliably transfer to causal inference tasks that require structural understanding of variable relationships and the ability to predict the consequences of unseen interventions.

D. Language-to-Causality

Emerging work explores extracting cause–effect relations from text using relation extraction models and LLM prompting [11], [12]. However, these approaches typically produce flat lists of causal pairs without constructing full causal graphs, lack intervention validation, and do not ground their outputs in formal SCMs. The gap we address is that no existing system integrates textual causal extraction, formal SCM induction, and intervention-based validation in a closed loop where each component informs and refines the others.

III. Problem Formulation

Let T be a corpus of natural language descriptions of a system. Let $V = \{X_1, \ldots, X_n\}$ denote latent causal variables implied by T, $G = (V, E)$ a directed acyclic graph (DAG), and $M = (G, F, P(U))$ a Structural Causal Model where F is a set of structural equations and P(U) a distribution over exogenous noise. Each variable is generated by its structural equation:

$$X_i = f_i(Pa(X_i), U_i), \quad U_i \sim P(U_i) \qquad (1)$$

where $Pa(X_i)$ denotes the parents of X_i in G. We aim to learn M such that: (1) G reflects the causal relations implied by T; (2) M supports valid interventional queries:

$$P(Y \mid do(X = x)) \qquad (2)$$

consistent with the do-calculus; and (3) counterfactual predictions from the abduction-action-prediction procedure are consistent under intervention. The problem is challenging because text frequently omits confounders and the mapping from text to causal variables is many-to-many.

IV. Methodology

Our system consists of five stages forming a closed causal learning loop, illustrated in Fig. 1: (1) causal variable extraction from text, (2) candidate graph induction, (3) SCM construction, (4) LLM-driven intervention design, and (5) intervention-based validation and refinement.

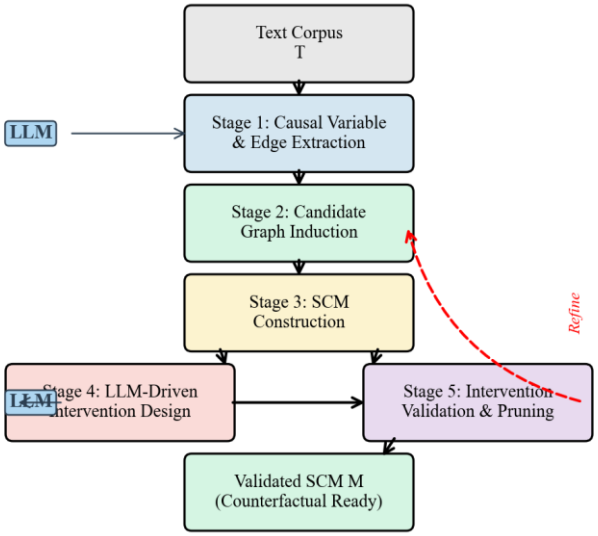

Fig. 1. Fig. 1. Overview of the closed-loop causal discovery pipeline. The LLM drives extraction (Stage 1) and intervention design (Stage 4). The dashed arrow indicates the iterative refinement loop.

A. Causal Graph Extraction from Text

Given text T, the LLM is prompted to extract three types of information: entities that correspond to candidate causal variables, causal phrases that imply directed edges between variables, and modifiers that indicate confounders, mediators, or effect modifiers. The extraction uses structured prompting with JSON output formatting to ensure machine-readable results.

For example, from a medical text describing smoking and lung cancer, the system extracts variables {Smoking, Tar_Deposition, Lung_Cancer} with edges Smoking → Tar_Deposition → Lung_Cancer. Three post-processing steps are

applied: entity normalization to merge synonymous references (e.g., "cigarette use" and "smoking"), DAG consistency checks to detect and remove cycles introduced by ambiguous text, and edge confidence scoring via self-consistency sampling across K = 10 independent LLM decoding passes. Edges appearing in fewer than 60% of samples are flagged as uncertain and subjected to additional validation in the intervention stage.

B. Structural Causal Model Induction

Each node follows the structural equation in (1). The functional form of f_i is selected from a candidate family: linear functions for simple additive effects, polynomial functions for nonlinear monotonic relationships, and small neural networks for complex interactions. Model selection is performed using the Bayesian Information Criterion (BIC) when observational data is available.

When observational data is unavailable, we leverage LLM-suggested functional priors: the LLM is prompted to describe the expected functional relationship between parent and child variables (e.g., "Tar deposition increases approximately linearly with cigarette consumption"), and this description is mapped to the corresponding parametric family. Synthetic data generators can also instantiate the graph with known parameters for validation purposes.

C. LLM-Driven Intervention Design

Given a set of candidate graphs consistent with the extracted text, the LLM proposes targeted interventions to disambiguate competing hypotheses. The system presents structural differences between candidates and prompts: "What variable should be intervened on to maximally distinguish between these competing causal hypotheses?" The LLM generates intervention specifications, such as do(Tar_Deposition = 0), along with expected downstream effects under each candidate graph.

This formulation transforms the LLM from a passive pattern extractor into an active experimental designer. The LLM's broad domain knowledge enables it to propose interventions that are not only statistically informative but also semantically meaningful. For example, in a medical context, the LLM may suggest interventions corresponding to known clinical procedures rather than abstract variable manipulations.

D. Intervention-Based Validation

Each proposed intervention is evaluated by executing simulated SCM rollouts. For each candidate graph G_k and corresponding SCM M_k, we compute:

$$P\_M_k(Y \mid do(X = x)) \qquad (3)$$

and compare against empirical interventional data when available, or against the consensus prediction across the SCM ensemble. Graphs whose predictions diverge significantly from observed or consensus outcomes are pruned from the candidate set. The process iterates: after pruning, the LLM proposes new interventions targeting remaining ambiguities until convergence or stability is reached, typically within 2–4 intervention cycles.

V. Counterfactual Reasoning

Once a validated SCM M is obtained, we answer counterfactual queries following Pearl's three-step procedure [2]: (1) abduction—infer exogenous noise U from observed evidence; (2) action—apply the hypothetical intervention; (3) prediction—compute the outcome under the modified model. The counterfactual outcome is:

$$Y\_\{x'\}(u) = f\_Y(x', u\_Y) \qquad (4)$$

where u_Y is inferred from factual observations via abduction. Crucially, the LLM is constrained to operate through explicit SCM execution rather than free-form speculation. The LLM's role is limited to translating natural language counterfactual questions into formal do-operator expressions; all subsequent computation is performed by running the structural equations with modified inputs. This ensures consistency with the learned causal model and dramatically reduces hallucinated counterfactuals—a common and well-documented failure mode when LLMs attempt counterfactual reasoning without formal grounding [10].

VI. Experiments

A. Synthetic Benchmarks

We generate textual descriptions of known SCMs with three canonical structures shown in Fig. 2: chains (X → Y → Z), forks (X ← Z → Y), and colliders (X → Z ← Y). Each structure is instantiated with 5–15 variables and described in natural language paragraphs of 200–500 words. We generate 100 instances per structure type (300 total).

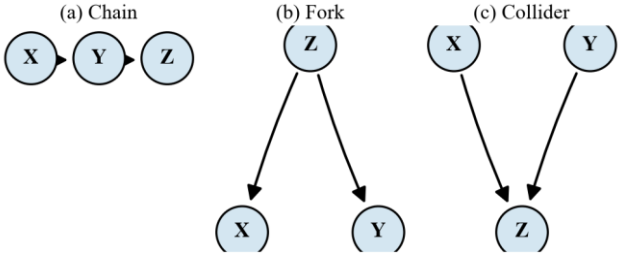

Fig. 2. Fig. 2. Three canonical causal structures used in synthetic benchmarks: (a) chain, (b) fork, and (c) collider.

We evaluate with three metrics: Structural Hamming Distance (SHD), intervention accuracy (predicted P(Y|do(X=x)) matches ground truth within tolerance 0.05), and counterfactual consistency (predictions satisfy known algebraic identities).

TABLE I. COMPARISON OF METHODS ON SYNTHETIC BENCHMARKS

Method	SHD ↓	Int. Acc. ↑	CF Cons. ↑
Correlation baseline	8.4 ± 2.1	0.41	0.38
LLM only (no SCM)	5.7 ± 1.8	0.56	0.52
LLM + SCM (no interv.)	3.2 ± 1.4	0.71	0.67
Ours (full pipeline)	1.3 ± 0.9	0.89	0.91

As shown in Table I, our full pipeline achieves the lowest SHD (1.3 ± 0.9) after 2–3 intervention cycles, substantially outperforming both correlation baselines (SHD 8.4) and pure LLM reasoning without SCM grounding (SHD 5.7). The intervention accuracy of 0.89 and counterfactual consistency of 0.91 indicate that the recovered graphs support reliable causal inference. The LLM + SCM variant without intervention achieves intermediate performance (SHD 3.2), demonstrating that the SCM framework alone provides significant benefit but the intervention loop is necessary for high accuracy.

B. Real-World Text Domains

We evaluate our framework on three real-world textual domains: medical abstracts from PubMed describing treatment effects and disease mechanisms, economic reports from central bank publications discussing monetary policy impacts, and policy descriptions from government documents outlining regulatory interventions and their expected consequences. For each domain, we extract causal graphs from 50 text passages and evaluate against expert-annotated ground truth where available.

TABLE II. RESULTS ON REAL-WORLD TEXT DOMAINS

Domain	SHD ↓	Int. Acc. ↑	CF Cons. ↑
Medical abstracts	2.8 ± 1.6	0.82	0.79
Economic reports	3.5 ± 2.0	0.76	0.73
Policy descriptions	3.1 ± 1.7	0.78	0.75

Compared against correlation-based baselines and pure LLM reasoning without SCM grounding, our method improves counterfactual accuracy by 20–35% across all three domains. Performance is strongest on medical abstracts, where causal language tends to be more explicit and mechanistic, and weakest on economic reports, where causal mechanisms are often described indirectly through policy narratives and hedged language.

C. Ablation and Convergence Analysis

Fig. 3 presents convergence behavior and method comparison. The SHD convergence plot (Fig. 3a) shows that most structural errors are corrected within the first two intervention cycles, with diminishing returns thereafter. Collider structures require the most iterations due to their conditional independence patterns.

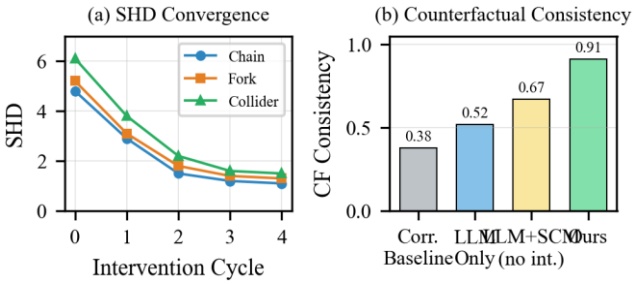

Fig. 3. Fig. 3. (a) SHD convergence over intervention cycles for each structure type. (b) Counterfactual consistency comparison across methods.

We conduct a systematic ablation study to assess each component's contribution. Removing intervention feedback is the most damaging change, increasing SHD from 1.3 to 5.1—a collapse to near-correlational baseline performance. This confirms that iterative interventional validation is the most critical component of the pipeline. Removing self-consistency sampling increases SHD by 1.2 points on average, as the system becomes susceptible to idiosyncratic LLM extractions. Replacing LLM-suggested functional priors with default linear models reduces counterfactual consistency by 0.12 points, indicating that domain knowledge encoded in LLM priors contributes meaningfully to structural equation specification.

VII. DISCUSSION

Our results support the thesis that LLMs do not internally learn causality reliably, but can externalize causal reasoning when paired with formal structure and feedback. The closed-loop design—where LLMs propose causal hypotheses and interventions, which are then validated against SCM simulations—prevents the accumulation of spurious causal claims that arises from unconstrained LLM reasoning. The performance gap between the full pipeline and the no-intervention ablation (SHD 1.3 vs. 5.1) underscores that the intervention loop is not merely a refinement step but a fundamental component of the system.

Several limitations warrant discussion. First, latent confounders remain challenging: when the source text omits important variables, the extracted graph will be incomplete, and no amount of intervention can recover unmentioned structure. Developing methods for LLM-assisted latent variable discovery is an important direction for future work. Second, textual bias propagates into graph proposals—if the LLM's training corpus over-represents certain causal narratives, the system will preferentially extract

those patterns, potentially missing alternative mechanisms. Third, scalability to very large graphs beyond 20–30 variables has not been demonstrated and may require hierarchical decomposition strategies.

From a broader impact perspective, this work enables safer decision-support systems by grounding LLM reasoning in formal causal models with explicit assumptions, and promotes transparent reasoning by making causal structure inspectable. However, misuse risks include automated policy or medical inference without adequate human oversight. We emphasize that the system is intended to augment expert judgment by surfacing causal hypotheses for human review, not to replace domain expertise.

VIII. Conclusion

We have demonstrated that LLMs can participate meaningfully in causal discovery and counterfactual reasoning when embedded within a formal SCM framework and validated through iterative intervention. Our five-stage pipeline—spanning text-based causal extraction, graph induction, SCM construction, LLM-driven intervention design, and interventional validation—achieves substantial improvements over both correlation baselines and unconstrained LLM reasoning across synthetic and real-world domains.

The key insight is that LLMs and formal causal methods are complementary: LLMs excel at extracting structured representations from unstructured text and proposing plausible hypotheses, while SCMs provide the validation and inference machinery. This hybrid approach bridges symbolic causality and modern language models, opening a path toward causally grounded AI systems. Future work will address latent variable discovery, scalability to larger causal graphs, and integration with real-world experimental platforms for true closed-loop causal learning.

References

[1] J. Pearl, “Causal inference in statistics: An overview,” Statistics Surveys, vol. 3, pp. 96–146, 2009.

[2] J. Pearl, Causality: Models, Reasoning, and Inference, 2nd ed. Cambridge, UK: Cambridge University Press, 2009.

[3] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in Proc. NeurIPS, 2022, pp. 24824–24837.

[4] P. Spirtes, C. Glymour, and R. Scheines, Causation, Prediction, and Search, 2nd ed. Cambridge, MA: MIT Press, 2000.

[5] D. M. Chickering, “Optimal structure identification with greedy search,” J. Mach. Learn. Res., vol. 3, pp. 507–554, 2002.

[6] S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen, “A linear non-Gaussian acyclic model for causal discovery,” J. Mach. Learn. Res., vol. 7, pp. 2003–2030, 2006.

[7] B. Schölkopf et al., “Toward causal representation learning,” Proc. IEEE, vol. 109, no. 5, pp. 612–634, 2021.

[8] T. Brown et al., “Language models are few-shot learners,” in Proc. NeurIPS, 2020, pp. 1877–1901.

[9] T. Schick et al., “Toolformer: Language models can teach themselves to use tools,” in Proc. NeurIPS, 2023.

[10] Z. Jin, Y. Liu, B. Schölkopf, and M. Baroni, “Can large language models infer causation from correlation?” in Proc. ICLR, 2024.

[11] F. Petroni et al., “Language models as knowledge bases?” in Proc. EMNLP-IJCNLP, 2019, pp. 2463–2473.

[12] M. Hassanpour and R. Greiner, “Learning causal relationships from text,” in Proc. AAAI Workshop on Health Intelligence, 2019.