Uncovering Bias Paths with LLM-guided Causal Discovery: An Active Learning and Dynamic Scoring Approach

Khadija Zanna[1], **Akane Sano**[1]
[1]Rice University
khzanna@rice.edu, akane.sano@rice.edu

Abstract

Ensuring fairness in machine learning requires understanding how sensitive attributes like race or gender causally influence outcomes. Existing causal discovery (CD) methods often struggle to recover fairness-relevant pathways in the presence of noise, confounding, or data corruption. Large language models (LLMs) offer a complementary signal by leveraging semantic priors from variable metadata. We propose a hybrid LLM-guided CD framework that extends a breadth-first search strategy with active learning and dynamic scoring. Variable pairs are prioritized for querying using a composite score combining mutual information, partial correlation, and LLM confidence, enabling more efficient and robust structure discovery. To evaluate fairness sensitivity, we introduce a semi-synthetic benchmark based on the UCI Adult dataset, embedding domain-informed bias pathways alongside noise and latent confounders. We assess how well CD methods recover both global graph structure and fairness-critical paths (e.g., sex→education→income). Our results demonstrate that LLM-guided methods, including our active, dynamically scored variant, outperform baselines in recovering fairness-relevant structure under noisy conditions. We analyze when LLM-driven insights complement statistical dependencies and discuss implications for fairness auditing in high-stakes domains.

Extended version —
https://doi.org/10.48550/arXiv.2506.12227

1 Introduction

Bias in machine learning (ML) systems affects decisions in hiring, lending, education, and healthcare. These biases often emerge through indirect, structural pathways where sensitive attributes (e.g., race or gender) influence outcomes via proxies or confounded relationships (Graetz, Boen, and Esposito 2022). Standard fairness audits rely on statistical disparity metrics, but such metrics overlook how bias propagates causally (Chinta et al. 2025). Without structural insight, interventions may be ineffective or misleading.

Causal discovery (CD) offers tools to identify such pathways, distinguishing genuine effects from those introduced by confounders or measurement artifacts. While recent methods apply structural causal models or fairness constraints to audit ML systems, classical CD techniques often fail under noisy data, latent confounding, or incomplete metadata. These limitations are acute in fairness contexts where both precision and interpretability are critical (Takayama et al. 2024).

Large language models (LLMs) have emerged as promising tools for CD, using their vast semantic knowledge to infer causal directions between variables from textual metadata (Le, Xia, and Chen 2024; Vashishtha et al. 2023). Prior work has used LLMs to infer causal links or orderings, and to reduce query cost via strategies like breadth-first search (BFS) (Jiralerspong et al. 2024). However, naive LLM use risks over or under attributing causality, especially involving sensitive attributes, raising concerns about spurious fairness conclusions.

Building on prior work, we introduce a fairness-driven CD framework that enhances the BFS approach with active learning (AL) and dynamic scoring. Our method prioritizes variable pairs using mutual information (MI), partial correlation (PCorr), and LLM-derived confidence, while discounting redundant queries and focusing on informative regions via a history-based weight. Unlike exhaustive querying or fixed-order search, our adaptive strategy balances semantic and statistical cues to recover fairness-relevant paths more efficiently.

A key challenge in evaluating CD methods in fairness contexts is the absence of real-world datasets with known ground-truth causal structures involving sensitive attributes (Loftus et al. 2018). Most, like the UCI Adult dataset, offer rich features but lack verified causal graphs. To enable robust evaluation, we construct a semi-synthetic benchmark from the UCI Adult dataset, embedding a known fairness-relevant causal graph with noise, corruption, and confounding. We assess how well CD methods recover both global structure and key fairness pathways (e.g., *sex* → *education* → *income*).

We also conduct a hyperparameter sensitivity analysis to examine how LLM-specific factors such as temperature, query prioritization weights, and stopping criteria affect CD performance. These findings offer practical insights into how LLM behavior and prompting influence graph reconstruction accuracy and fairness-path recovery. Our approach is designed for robustness, interpretability, and social ac-

countability, enabling stakeholders to trace how outcomes are structurally linked to sensitive variables and to intervene in meaningful ways.

Our key contributions:

- A hybrid LLM-CD framework with active learning and dynamic scoring to enhance fairness-aware structure discovery.
- A benchmark for fairness-sensitive CD grounded in real-world semantics and controlled ground truth.
- An evaluation of pathway recovery across methods, highlighting conditions where LLM-guided discovery improves fairness diagnostics.
- A sensitivity analysis on LLM parameters and their influence on recovery quality.

2 Related Work

Statistical and Optimization-Based CD. CD methods span statistical, optimization-based, and more recently, LLM-guided approaches (Long et al. 2025). Classical techniques such as PC (Spirtes and Glymour 1991) and GES (Meek 1997) rely on conditional independence tests and strong assumptions (e.g., faithfulness), often failing under latent confounding or limited data. Optimization-based methods like NOTEARS (Zheng et al. 2018) and DAGMA (Bello, Aragam, and Ravikumar 2022) offer greater precision but are computationally expensive and sensitive to hyperparameters, making them less practical for fairness-critical contexts.

LLM-Guided CD. Recent LLM-based approaches use variable semantics to infer structure, with Kıcıman et al. (Kıcıman et al. 2023) proposing pairwise causal queries from metadata. However, such methods scale poorly due to quadratic query complexity. Others combine LLM-inferred causal orders with classical CD (Vashishtha et al. 2023), or use multi-agent prompting strategies (Khatibi et al. 2024; Le, Xia, and Chen 2024). Kampani et al. (Kampani et al. 2024) generate LLM-based priors refined through NOTEARS. Takayama et al. (Takayama et al. 2024) introduced statistical causal prompting, where LLMs enhance statistical methods by providing priors or plausible causal orders. Jiralerspong et al. (Jiralerspong et al. 2024) proposed a BFS strategy to reduce query complexity, but rely on fixed pair ordering, making them susceptible to early decision errors.

We build on this line by introducing AL and a dynamic scoring mechanism to adaptively prioritize informative variable pairs using MI, PCorr, and LLM confidence. Our method dynamically balances semantic and statistical signals, in contrast to prior work that uses LLMs for one-time priors or uniform querying. We focus on observational CD methods to maintain comparability with LLM-based approaches, leaving integration with interventional or experimental methods (e.g., do-calculus, IV-based CD) for future work.

Fairness and Causal Pathways. Beyond structural accuracy, recent work has emphasized the importance of CD for fairness auditing. Causal fairness frameworks (Nabi and Shpitser 2018; Kilbertus et al. 2017; Zhang and Bareinboim 2018; Chiappa 2019) focus on identifying discriminatory pathways, e.g., *sex* $\rightarrow$ *education* $\rightarrow$ *income* that explain how bias propagates through mediation or proxy variables. Our method is explicitly tuned to recover such fairness-relevant paths with greater robustness under noise and limited samples, enabling more actionable fairness assessments.

Social and Regulatory Context. Beyond computer science, our work connects to social science perspectives. Structural discrimination and intersectionality highlight how compounding disadvantage flows through proxy variables, not just direct links (Crenshaw 2013; Bonilla-Silva 1997). Regulatory frameworks such as GDPR's "right to explanation" and EEOC compliance requirements also motivate interpretable, path-level audits (Kaminski 2021).

Causal Inference and Accountability. We also draw from causal inference frameworks like the potential outcomes model (Rubin 1974; Imbens and Rubin 2015) and matching-based fairness analysis (Stuart 2010), which stress structural validity and transparency of assumptions. Our hybrid design reflects these values by surfacing interpretable causal hypotheses grounded in both observed data and language-based priors. Finally, we respond to calls for auditable AI systems through transparency-centered work on model documentation and sociotechnical accountability (Mitchell et al. 2019; Selbst et al. 2019).

3 Proposed LLM-Based Method

We build on the BFS-based CD framework of (Jiralerspong et al. 2024), where an LLM is queried iteratively to construct a causal graph. The method begins by identifying independent root variables and expands outward via LLM-guided queries, treating each confirmed edge as a BFS traversal step. To ensure acyclicity, edges are added only if they do not create cycles. While more efficient than exhaustive querying, this approach treats all variable pairs uniformly and can waste queries on uninformative or redundant relationships.

Active Learning and Dynamic Scoring for Efficient Querying

We introduce an AL strategy that selects variable pairs based on a dynamic informativeness score. This score integrates statistical signals, model confidence, and query history, prioritizing pairs likely to yield informative causal judgments. The query loop continues until a maximum iteration or informativeness threshold is reached. Figure 1 summarizes the overall architecture.

Dynamic Scoring Mechanism Each unqueried pair (x, y) receives a composite score:

$$S(x,y) = w_{\text{stat}} \cdot \text{StatScore}(x,y) + w_{\text{conf}} \cdot \text{LLMConf}(x,y) + w_{\text{hist}} \cdot \text{HistScore}(x,y) \quad (1)$$

Each term incorporates a distinct signal:

- **Statistical Score:** Combines MI and PCorr, measuring both linear and non-linear dependencies:

$$\text{StatScore}(x,y) = \frac{\text{MI}(x,y) + \text{PCorr}(x,y)}{2} \quad (2)$$

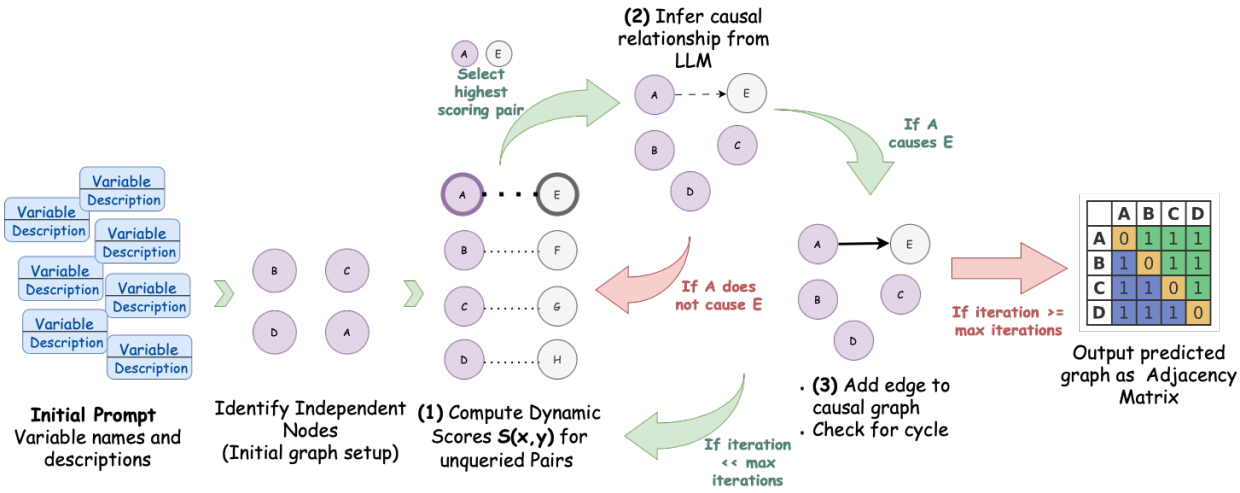

Figure 1: Overview of the proposed LLM-guided BFS framework with dynamic scoring and AL.

PCorr is computed conditionally on the current discovered parent sets, following the iterative conditioning procedure of the PC algorithm.

- **LLM Confidence Score:** Reflects the certainty of the model's previous response for a given pair, where higher token-level confidence leads to higher scores:

$$\text{LLMConf}(x, y) = \frac{1}{1 + e^{-\text{confidence}}} \tag{3}$$

- **Query History Score:** Penalizes repeated queries to encourage broader exploration:

$$\text{HistScore}(x, y) = \frac{1.5}{1 + \text{query\_count}(x, y)} \tag{4}$$

The weights $w_{\text{stat}}, w_{\text{conf}}, w_{\text{hist}}$ are treated as tunable hyperparameters. These weights are optimized via Bayesian optimization (see Section 6), allowing the framework to adapt scoring behavior to different graph structures or domain characteristics.

Querying the LLM At each step, the pair (x^*, y^*) with the highest score is selected:

$$(x^*, y^*) = \arg \max_{(x,y) \in \text{Unqueried}} S(x, y) \tag{5}$$

The LLM is queried using a simple natural language prompt:

Listing 1: Simplified LLM Query Format

```
System: You are a domain expert.
User: You assist the expert in
    evaluating possible causal links.

Metadata: Each variable is described
    using a short name and a natural
    language description.

Query: Does VAR_A cause VAR_B?
Respond with <Answer>Yes</Answer> or <
    Answer>No</Answer>.
```

Domain-specific instructions and variable descriptions are included to contextualize each query. See Appendix for full example prompts.

If the LLM returns "Yes" and adding $x^* \to y^*$ does not form a cycle, the edge is added to the graph. To estimate the LLM uncertainty, we extract token-level log-probabilities and compute the average probability of the model's response ("Yes" or "No"). This score, scaled to [0,1], informs the dynamic scoring function. If log-probs are unavailable, a default confidence of 0.5 is used. The process terminates after a query limit or when scores fall below a threshold.

To ensure the output is a valid DAG, each edge addition is immediately followed by a cycle check. If the graph becomes cyclic, the addition is reversed.

The final causal graph is returned both as a directed edge list and an adjacency matrix:

$$A(i, j) = \begin{cases} 1 & \text{if } X_i \to X_j \text{ is predicted,} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The entire procedure is conducted in a multi-turn chat session, allowing the LLM to maintain contextual awareness of previously inferred relationships and rationales.

4 Fairness Evaluation via Pathway Analysis

We assess the fairness utility of learned causal graphs by analyzing how sensitive attributes S (e.g., `race`, `sex`) influence outcomes Y (e.g., `income`). Unlike statistical fairness metrics, pathway analysis provides a causal perspective by distinguishing direct, mediated, and spurious effects (Pearl 2009, 2022).

Path Classification We enumerate all directed paths from S and classify them as:

- **Direct:** Edges $S \to Y$.
- **Indirect:** Paths $S \to \cdots \to Y$ via mediators.
- **Spurious:** Paths involving S that do not reach Y.

Comparing these across methods and against the ground-truth graph reveals how well fairness-relevant mechanisms are recovered.

Effect Decomposition We estimate the causal effects of S on Y using either structural equations or interventional estimators:

$$\begin{aligned} DE &: \text{Direct effect (via } S \to Y) \\ IE &: \text{Indirect effect (via mediators)} \\ TE &= DE + IE \end{aligned}$$

We normalize by outcome variance to obtain:

$$C_{\text{bias}} = \frac{TE}{\text{Var}(Y)} \tag{7}$$

Here, C_{bias} quantifies the fairness-relevant contribution of S to Y, enabling method comparison across datasets. Higher values indicate a greater risk of bias propagation.

5 Datasets

We evaluate our framework on three benchmark networks of varying size and realism.

Synthetic Adult-Based Network. We construct a semi-synthetic dataset based on the UCI Adult dataset (Kohavi et al. 1996), a standard benchmark in fairness research. The graph includes 15 original variables, with causal edges informed by prior work (Kilbertus et al. 2017; Nabi and Shpitser 2018). Demographics (e.g., `race`, `sex`) influence education and occupation, which in turn affect income, both directly and via mediators such as capital gains. Age and marital status act as confounders or mediators.

To simulate structural bias, we inject direct edges from `race` and `sex` to `income`, alongside indirect paths (e.g., `sex` $\to$ `education` $\to$ `income`). These allow evaluation of direct, indirect, and total effects. Details on graph construction and data generation are in the Appendix.

Child Causal Network. A 20-node, 25-edge Bayesian network modeling clinical, environmental, and parental factors in congenital heart disease (Spiegelhalter et al. 1993). Variables include both categorical and continuous types.

Neuropathic Pain Network. A large-scale clinical graph with 221 nodes and 770 edges capturing pathophysiological and symptomatic relationships in neuropathic pain diagnoses, derived from real-world patient data (Tu et al. 2019).

Since the latter two datasets lack sensitive attributes, we focus on structural accuracy rather than fairness.

6 Experiments

We evaluate our LLM-based CD framework across the benchmark datasets, focusing on graph accuracy and fairness-relevant path recovery under varying noise and hyperparameter settings.

Baselines

We compare against the following methods:

- **PC Algorithm** (Spirtes and Glymour 1991): Constraint-based method using independence tests.
- **GES** (Meek 1997): Score-based approach optimizing a score function (e.g., BIC) via forward-backward search.
- **NOTEARS** (Zheng et al. 2018): Continuous optimization using a differentiable acyclicity constraint.
- **DAGMA** (Bello, Aragam, and Ravikumar 2022): Neural network-based CD with sparsity and acyclicity constraints.
- **LLM Pairwise** (Kıcıman et al. 2023): Uses LLM to infer pairwise causality from metadata.
- **LLM BFS** (Jiralerspong et al. 2024): Explores graph structure via LLM-guided breadth-first querying.

All baselines are run using the open-source code from (Jiralerspong et al. 2024).

Experimental Setup

We use Bayesian optimization (`gp_minimize`) with a Gaussian Process surrogate to tune key hyperparameters (Mockus 1994; Snoek, Larochelle, and Adams 2012):

- **Scoring Weights:** Weights for MI, PCorr, and query history (Eq. 1)
- **Score Threshold:** Minimum score required for querying a pair
- **LLM Temperature:** Controls randomness in responses
- **Max Iterations:** AL rounds per run

Each query uses a multi-turn GPT-4 (0125-preview) prompt with prior discoveries and variable metadata (Section 3). A directed edge $X \to Y$ is added if the reply contains `<Answer>Yes</Answer>` and does not create a cycle. If ambiguous, no edge is added. Confidence scores are derived from the average probability of top-5 tokens based on log-probs; defaulting to 0.5 if unavailable.

Trials are divided into chunks across hyperparameter space. We report results from the best-performing configuration for each run. Full prompt design and search ranges are detailed in the Appendix.

We evaluate all methods on four synthetic variants of the Adult-based dataset (Section 5), varying random seeds to preserve the underlying structure while introducing independent data realizations. Fairness analysis targets paths from `sex`, `race`, and `age` to `income`. Structural accuracy is also assessed on the Child and Neuropathic networks, with LLM-based methods averaged over five independent runs.

Evaluation Metrics

We assess reconstruction accuracy using standard metrics:

- **Precision & Recall:** Defined as Precision $= \frac{\text{TP}}{\text{TP + FP}}$ and Recall $= \frac{\text{TP}}{\text{TP + FN}}$.
- **F1 Score:** Harmonic mean of precision and recall.
- **Edge Count:** Number of predicted edges vs. ground truth.
- **Acyclicity:** Ensures output is a valid DAG.

- **Normalized Hamming Distance (NHD):** NHD $= \frac{\text{Mismatches}}{n^2}$, where n is the number of nodes.
- **Adjacency Accuracy:** Accuracy $= \frac{\text{Correct entries}}{n^2}$.

7 Results and Analyses

Graph Performance Comparison

Synthetic Adult-based Network. Our method achieves top scores across F1 (0.585), accuracy (0.413), precision (0.792), and NHD (0.109), recovering a compact graph (17 edges) with strong alignment to ground truth. LLM Pairwise attains high recall (0.813) but overpredicts, while GES and PC show moderate F1 but higher NHDs. Optimization-based methods struggle with fairness structure.

Child Network. Again, our method leads in F1 (0.533), accuracy (0.364), and NHD ratio (0.467), balancing edge precision and coverage. NOTEARS achieves slightly better raw NHD but worse recall. LLM Pairwise finds many indirect effects, but at the cost of low precision.

Neuropathic Network. Despite the challenge of 221 nodes, our method yields highest F1 (0.136) and precision (0.690), outperforming baselines. All baselines degrade sharply; LLM BFS fails entirely (F1 = 0). Dynamic scoring enables better query selection under sparsity.

Scalability. Our method scales better than NOTEARS, DAGMA, or LLM Pairwise while offering greater structural fidelity. Complexity analysis (detailed in Appendix) shows favorable scaling theoretically in large sparse graphs, common in real-world networks. GES and LLM Pairwise were excluded from Neuropathic due to resource limits.

Reproducibility. LLM-BFS and Pairwise underperform vs. Jiralerspong et al. (2024), likely due to (1) GPT-4 updates and nondeterminism, (2) token limits or message handling, and (3) lack of access to original variable descriptions for the Neuropathic dataset, affecting metadata quality. Our implementation adheres to the BFS protocol and exposes its limitations under scale, motivating our enhancements.

Fairness Pathway Recovery

We evaluate fairness path recovery on the Adult-based synthetic graph, where the true structure contains 2 direct and 25 indirect paths from `sex`, `age`, and `race` to `income`. TE = 4.89 and $C_{\text{bias}} = 28.46$.

Our method uniquely recovers both true direct paths (from `sex` and `race`) while omitting the spurious `age` path, showing strong fairness alignment. LLM methods overattribute (3 direct paths) due to LLM priors, inflating C_{bias}. PC and GES miss direct effects; NOTEARS and DAGMA capture none, yielding near-zero C_{bias}.

Overall, our method prioritizes high-confidence discovery, avoiding false positives and maintaining robustness across seeds as reflected by its consistently low standard deviation on key metrics such as indirect path count, total effect, and normalized bias as seen in the table in Appendix. Its conservative nature may underestimate bias but enhances reliability in fairness-critical settings.

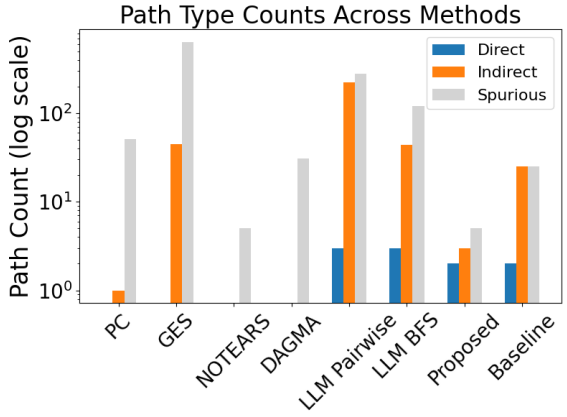

Figure 2: Log-scaled path counts (direct, indirect, spurious) per method.

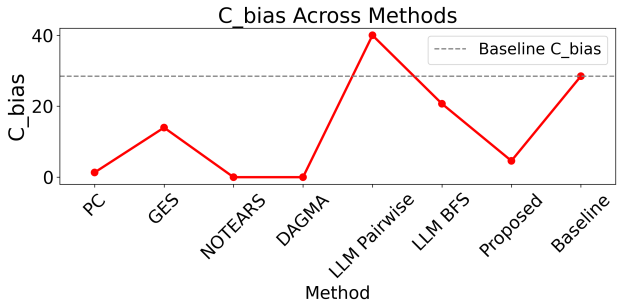

Figure 3: Normalized bias (C_{bias}) across methods.

Ablation and Parameter Influence

Bayesian optimization trials were analyzed via Random Forest regression to assess parameter impact on F1 score.

Findings. Max iterations have the strongest effect, suggesting deeper AL improves recovery. Score threshold and temperature also influence performance. Correlation analysis (Figure 5) shows anti-correlation among scoring weights, indicating competitive trade-offs. In small graphs, deeper querying matters more; in large ones (e.g., Neuropathic), MI/PCorr dominate (details in Appendix). This highlights how our dynamic scoring mechanism could adapt to scale, favoring semantic guidance in small graphs and statistical grounding in larger ones.

8 Discussion

Our findings demonstrate that combining LLM-based CD with AL and dynamic scoring improves both structural accuracy and fairness sensitivity across diverse datasets. The proposed method consistently outperforms statistical, LLM-only, and hybrid baselines on both small and large networks.

Dynamic scoring balances LLM semantic priors with empirical signals, adapting to graph complexity and data quality. In early iterations, LLM judgments guide exploration when statistical signals are weak. As more information is gathered, mutual information and partial correlation become

Adult-based Synthetic	Method	Acc. (↑)	F1 Score (↑)	Precision	Recall	NHD (↓)	NHD Ratio (↓)	# Pred. Edges
(15 nodes, 28 edges)	PC	0.239	0.382	0.352	0.420	0.193	0.743	33
	GES	0.296	0.473	0.368	0.580	0.203	0.782	44
	NOTEARS ($\lambda = 0.01$)	0.021	0.039	0.035	0.045	0.260	0.650	27
	DAGMA ($\lambda = 0.05$)	0.099	0.180	0.141	0.250	0.283	0.794	50
	LLM Pairwise	0.307	0.470	0.331	**0.813**	0.253	**0.530**	69
	LLM BFS	0.299	0.456	0.332	0.750	0.305	0.539	64
	Proposed Method	**0.413**	**0.585**	**0.792**	0.464	**0.109**	**0.415**	17
Child	PC	0.146	0.255	0.273	0.239	0.097	0.745	22
(20 nodes, 25 edges)	GES	0.206	0.341	0.438	0.279	0.119	0.659	16
	NOTEARS	0.216	0.356	0.403	0.319	**0.080**	0.644	20
	DAGMA	0.179	0.304	0.333	0.279	0.089	0.696	21
	LLM Pairwise	0.130	0.229	0.144	**0.559**	0.235	0.770	97
	LLM BFS	0.150	0.261	0.286	0.240	0.085	0.739	21
	Proposed Method	**0.364**	**0.533**	**0.601**	0.479	0.082	**0.467**	20
Neuropathic	PC	0.041	0.078	0.092	0.068	**0.025**	0.922	563
(221 nodes, 770 edges)	NOTEARS	0.022	0.044	0.500	0.023	0.334	0.955	36
	DAGMA	0.020	0.039	0.421	0.021	0.351	0.960	38
	LLM BFS	0.000	0.000	0.000	0.000	0.903	1.000	43
	Proposed Method	**0.073**	**0.136**	**0.690**	**0.075**	0.109	**0.864**	84

Table 1: Performances of Methods on the Adult-based Synthetic, Child, and Neuropathic Causal Networks. All reported values are averaged over four random seeds. Metrics exhibit minimal variance across runs ($\pm 2 \times 10^{-4}$ in F1 score).

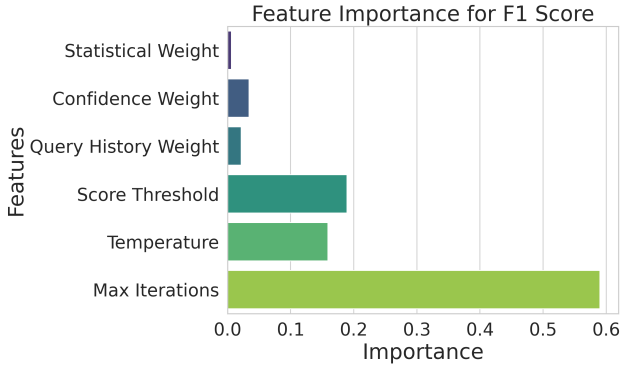

Figure 4: Hyperparameter importance on F1 score.

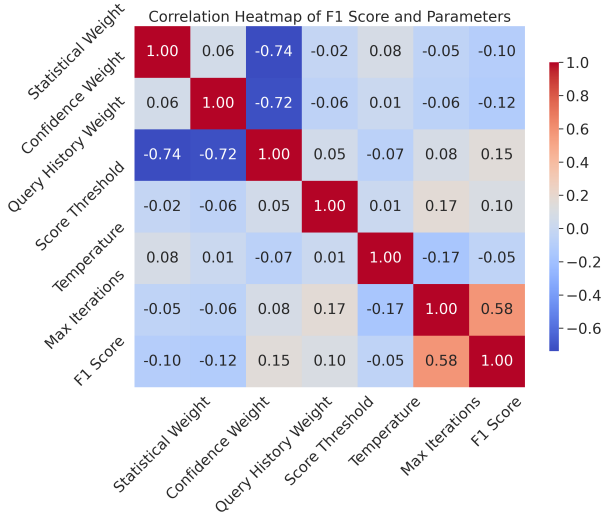

Figure 5: F1 vs. hyperparameter correlation (Adult).

dominant, reflecting a shift from semantic exploration to empirical refinement. AL enhances efficiency by prioritizing uncertain, low-redundancy variable pairs, avoiding the limitations of fixed-order strategies.

Our method uniquely recovers fairness-critical paths, e.g., `sex` → `education` → `income`—while avoiding spurious ones like `age` → `income`, often introduced by LLM-only baselines. This improves fairness diagnostics for impact assessments in domains like hiring or lending. While our fairness evaluation relies on a semi-synthetic benchmark, this is necessary as real-world datasets lack ground-truth causal graphs for sensitive attributes and outcomes.

However, our results also underscore broader concerns: CD methods, including ours, may introduce spurious or incomplete structures if used without scrutiny. In fairness contexts, such errors carry high stakes. False positives (e.g., unjustified bias paths) can trigger overcorrections, while false negatives obscure actual harms. We find that even underperforming methods can appear fairness-aware using naive metrics like C_{bias}, which may be inflated by erroneous causal paths. This highlights the need for path-level interpretability rather than summary statistics alone.

To mitigate these risks, our approach integrates LLMs as prior-informed agents within a selective AL loop. This reduces unnecessary queries, avoids early commitment errors,

and improves efficiency over fixed-query strategies such as BFS. Like traditional DAG methods, our framework embraces uncertainty and treats CD as hypothesis generation.

While it is argued that LLMs may not perform causal reasoning per se, we treat them as semantic priors filtered through statistical signals and query history. This is analogous to expert elicitation—subject to bias, but useful when paired with empirical validation. As argued by (Imbens 2020) and (McCoy et al. 2023), epistemic uncertainty is intrinsic to CD. LLM-guided proposals can still surface plausible structures, especially in fairness-critical settings. Prior work has emphasized this role (Kıcıman et al. 2023); we build on it with a dynamic querying process that improves recovery while reducing false positives in bias pathways.

Our broader aim is to support interpretable, efficient, and socially aware fairness diagnostics, offering complementary insight to notions like demographic parity or counterfactual fairness. By uncovering how sensitive variables influence outcomes through direct and mediated paths, our method moves beyond outcome disparities toward causal insight. This aligns with calls for path-level transparency in AI audits from fairness, accountability, and legal communities. We see this work as a step toward embedding causal reasoning into policy evaluations, model and risk assessments, tools needed to build more trustworthy socio-technical systems.

This approach provides actionable insights for non-technical stakeholders. By surfacing fairness-relevant pathways, the method enables practitioners, compliance teams, or policy analysts to trace how sensitive attributes indirectly shape outcomes. These insights can guide interventions, from revising decision rules to informing accountability processes in hiring, lending, or regulatory oversight.

9 Limitations

Simplified Synthetic Causal Structure. Our synthetic graph models `income` as a terminal node to isolate fairness pathways, aiding evaluative clarity. However, this excludes realistic downstream effects (e.g., on health, opportunity), omitting feedback or temporal dynamics. Future work should extend this to longitudinal or post-outcome settings.

Hyperparameter Sensitivity and Compute Costs. Our method requires tuning multiple hyperparameters (e.g., score weights, temperature), which can be resource-intensive. LLM querying is computationally costly, particularly for large graphs like Neuropathic. Token limits restrict the ability to encode full metadata, particularly in complex graphs like Neuropathic, leading to incomplete or inconsistent responses. Specialized variable names may be misinterpreted, especially at higher temperatures, introducing semantic errors and prediction noise. These factors collectively pose scalability and usability challenges, particularly for resource-constrained settings.

Reproducibility of LLM-based CD. While our method is LLM-agnostic, we observe discrepancies with results from Jiralerspong et al. (2024), highlighting broader reproducibility challenges. LLMs are stochastic and sensitive to prompt design, versioning, query order, and API behavior, factors rarely standardized or reported. Without standardized prompting protocols or access to latent states, consistent benchmarking remains difficult.

Risk of Social Bias in LLMs. LLMs trained on web-scale data may encode social biases, leading to causal inferences that reflect stereotypes rather than ground-truth mechanisms. This risks semantic hallucinations and biased edge assignments, particularly in fairness-critical applications. This risk is addressed through statistical weighting and confidence-based filtering, as described in Section 3.

Domain and Metadata Dependence. Our framework depends on interpretable variable metadata to guide LLM queries. It performs best in domains like healthcare or census data. In sparse, technical, or ambiguous domains (e.g., genomics, sensor data), performance degrades. While LLM priors scale efficiently, human-in-the-loop priors could be incorporated for low-interpretability domains. Future work could explore hybrid prompting strategies or domain adaptation techniques to extend semantic CD to less interpretable or evolving data environments.

10 Conclusion

We propose a fairness-driven CD framework that integrates LLMs with AL. Across benchmarks including synthetic and semi-synthetic real-world networks, our method outperforms prior baselines in structural accuracy and reduces false positives in fairness-critical path recovery. By combining statistical dependencies with LLM-based semantic priors and prioritizing informative queries, our framework improves both reliability and fairness-awareness in CD.

To support reproducible evaluation, we introduce a semi-synthetic benchmark based on the UCI Adult dataset, embedding structural bias, latent confounding, and noise. This provides a realistic yet controlled testbed with ground-truth graphs for consistent comparison across CD methods.

Our analysis shows that exploration budget (e.g., number of iterations) significantly affects performance, with the relative importance of LLM-derived versus statistical signals shifting with graph size and complexity. While our method improves fairness-path recovery, it does not offer end-to-end guarantees for downstream fairness outcomes. Instead, it uncovers explainable causal pathways, particularly those involving sensitive attributes intended to guide audits, impact assessments, or stakeholder review.

This positions our framework as an improvement in diagnostic tools for fairness evaluations. Use cases include hiring audits, policy design, and regulatory compliance reporting. By recovering interpretable causal paths, our method offers a step toward tools that support fairness-focused exploration, potentially informing future integration into dashboards or algorithmic auditing pipelines. We plan to extend the framework to dynamic graphs and release open-source pipelines for scalable, reproducible fairness audits.

Acknowledgments

This study was funded by NSF #2047296 and OpenAI's Researcher Access Program.

References

Bello, K.; Aragam, B.; and Ravikumar, P. 2022. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *Advances in Neural Information Processing Systems*, 35: 8226–8239.

Bonilla-Silva, E. 1997. Rethinking racism: Toward a structural interpretation. *American sociological review*, 465–480.

Chiappa, S. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Chinta, S. V.; Wang, Z.; Palikhe, A.; Zhang, X.; Kashif, A.; Smith, M. A.; Liu, J.; and Zhang, W. 2025. AI-driven healthcare: Fairness in AI healthcare: A survey. *PLOS Digital Health*, 4(5): e0000864.

Crenshaw, K. 2013. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *Feminist legal theories*, 23–51. Routledge.

Graetz, N.; Boen, C. E.; and Esposito, M. H. 2022. Structural racism and quantitative causal inference: a life course mediation framework for decomposing racial health disparities. *Journal of Health and Social Behavior*, 63(2): 232–249.

Imbens, G. W. 2020. Potential outcome and directed acyclic graph approaches to causality: Relevance for empirical practice in economics. *Journal of Economic Literature*, 58(4): 1129–1179.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

Jiralerspong, T.; Chen, X.; More, Y.; Shah, V.; and Bengio, Y. 2024. Efficient causal graph discovery using large language models. *arXiv preprint arXiv:2402.01207*.

Kaminski, M. E. 2021. The right to explanation, explained. In *Research handbook on information law and governance*, 278–299. Edward Elgar Publishing.

Kampani, S.; Hidary, D.; van der Poel, C.; Ganahl, M.; and Miao, B. 2024. LLM-initialized Differentiable Causal Discovery. *arXiv preprint arXiv:2410.21141*.

Khatibi, E.; Abbasian, M.; Yang, Z.; Azimi, I.; and Rahmani, A. M. 2024. ALCM: Autonomous LLM-Augmented Causal Discovery Framework. *arXiv preprint arXiv:2405.01744*.

Kıcıman, E.; Ness, R.; Sharma, A.; and Tan, C. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*.

Kilbertus, N.; Caruana, R.; Janzing, D.; Valera, I.; Hardt, M.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, volume 30.

Kohavi, R.; et al. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, 202–207.

Le, H. D.; Xia, X.; and Chen, Z. 2024. Multi-Agent Causal Discovery Using Large Language Models. *arXiv preprint arXiv:2407.15073*.

Loftus, J. R.; Russell, C.; Kusner, M. J.; and Silva, R. 2018. Causal reasoning for algorithmic fairness. *arXiv preprint arXiv:1805.05859*.

Long, S.; Tan, J.; Mao, B.; Tang, F.; Li, Y.; Zhao, M.; and Kato, N. 2025. A Survey on Intelligent Network Operations and Performance Optimization Based on Large Language Models. *IEEE Communications Surveys & Tutorials*.

McCoy, R. T.; Smolensky, P.; Linzen, T.; Gao, J.; and Celikyilmaz, A. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using raven. *Transactions of the Association for Computational Linguistics*, 11: 652–670.

Meek, C. 1997. *Graphical Models: Selecting causal and statistical models*. Ph.D. thesis, Carnegie Mellon University.

Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.

Mockus, J. 1994. Application of Bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4: 347–365.

Nabi, R.; and Shpitser, I. 2018. Fair inference on outcomes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Pearl, J. 2009. *Causality: Models, Reasoning, and Inference*. Cambridge University Press.

Pearl, J. 2022. Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, 373–392. Wiley.

Rubin, D. B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5): 688.

Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.

Snoek, J.; Larochelle, H.; and Adams, R. P. 2012. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25.

Spiegelhalter, D. J.; Dawid, A. P.; Lauritzen, S. L.; and Cowell, R. G. 1993. Bayesian analysis in expert systems. *Statistical science*, 219–247.

Spirtes, P.; and Glymour, C. 1991. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1): 62–72.

Stuart, E. A. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1): 1.

Takayama, M.; Okuda, T.; Pham, T.; Ikenoue, T.; Fukuma, S.; Shimizu, S.; and Sannai, A. 2024. Integrating large language models in causal discovery: A statistical causal approach. *arXiv preprint arXiv:2402.01454*.

Tu, R.; Zhang, K.; Bertilson, B.; Kjellstrom, H.; and Zhang, C. 2019. Neuropathic pain diagnosis simulator for causal

discovery algorithm evaluation. *Advances in Neural Information Processing Systems*, 32.

Vashishtha, A.; Reddy, A. G.; Kumar, A.; Bachu, S.; Balasubramanian, V. N.; and Sharma, A. 2023. Causal inference using llm-guided discovery. *arXiv preprint arXiv:2310.15117*.

Zhang, J.; and Bareinboim, E. 2018. Fairness in decision-making—the causal explanation formula. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zheng, X.; Aragam, B.; Ravikumar, P. K.; and Xing, E. P. 2018. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31.