# CLAIM-FUSE: CAUSAL DISCOVERY WITH HETEROGENEOUS AND UNCERTAIN KNOWLEDGE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Causal discovery aims to recover cause-and-effect relationships from observational data, and large language models (LLMs) can help generate textual causal statements when data alone is not sufficient to identify the correct causal relationship. However, existing LLM-assisted methods either incur high computational costs by querying the causal relationship for every variable pair, or ask repeated questions despite some question types being more informative and more reliably answered than others. We therefore propose CLAIM-FUSE, which chooses which variables and which type of causal claim to query (directed edge, adjacency, ancestry, or order) at each step, minimizing the uncertainty within a fixed budget. Our empirical results show that CLAIM-FUSE achieves the lowest structural error among LLM-assisted methods, up to 19% below the strongest baseline, using orders of magnitude fewer tokens than exhaustive methods. Choosing among all claim types also improves recovery over querying only local or only ancestral claims, though it can degrade when the LLM is less reliable on the chosen type. Our work lays the foundation for knowledge elicitation that adapts not only the content, but also the type of the question.

## 1 INTRODUCTION

Identifying the causal graph underlying a system allows us to predict how intervening on one variable affects another. Correlation alone is insufficient, as two variables may be associated without one causing the other. Randomized controlled trials are the gold standard for establishing causal effects, but they are often infeasible, unethical, or prohibitively expensive. Hence, *causal discovery* (Spirtes et al., 2000) aims to infer causal structure from observational data by exploiting statistical patterns such as conditional independencies and asymmetries in noise distributions.

Prior work has shown that LLMs can provide domain knowledge to guide causal discovery (Kiciman et al., 2023; Long et al., 2023; Darvariu et al., 2024). We refer to this knowledge as *causal claims*. These claims can describe different types of relationships. *Local* claims describe specific relationships between two variables, such as "X causes Y" or "X is adjacent to Y." In contrast, *ancestral* claims describe relationships that may involve intermediate variables, such as "X is an ancestor of Y" or "X precedes Y." For example, if smoking causes cancer only through tar, the ancestral claim "smoking causes cancer" is true, while the local claim "smoking directly causes cancer" is false.

However, existing methods suffer from the following limitations: some (1) query the LLM about every possible pair or triplet of variables, which can be costly (Vashishtha et al., 2025; Takayama et al., 2025); others (2) select the most informative variable pair but always ask the same type of question, typically about a local edge (Bonilla et al., 2026; da Silva et al., 2026; Verma et al., 2026), or (3) switch between claim types according to a predetermined sequence (Tench & Demeester, 2026). Yet, the type of causal claim being queried matters. Ancestral questions can rule out many candidate graphs at once, whereas edge-level questions provide only local information, a limitation also noted by Bonilla et al. (2026). Moreover, LLMs struggle to distinguish between direct and indirect effects when asked directed causal questions between two variables (Vashishtha et al., 2025).

To address these problems, our work introduces CLAIM-FUSE, a causal discovery framework that chooses not only which variables to query but also which type of causal claim to query. CLAIM-FUSE maintains a weighted set of candidate graphs initialized from observational data and treats every claim, regardless of its type, as a yes/no test on a graph. When an answer arrives, graphs



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

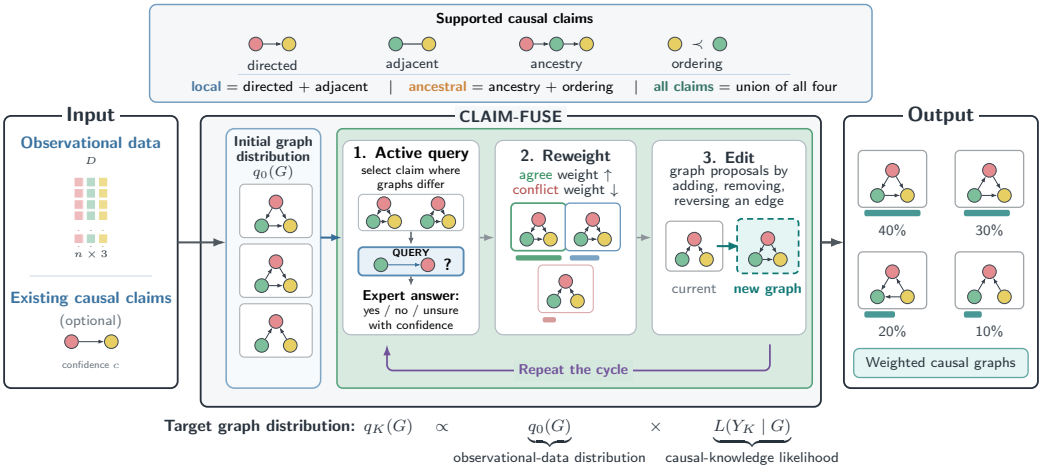


Figure 1. CLAIM-FUSE combines observational data with uncertain causal claims to recover a weighted distribution over causal graphs. Starting from an initial graph distribution $q_0(G)$, the method iteratively selects a causal claim where the graphs differ, obtains an expert answer with confidence, reweights graphs according to agreement or conflict, and proposes new graphs by adding, removing, or reversing an edge. Repeating this cycle yields weighted causal graphs approximating the target distribution $q_K(G)$.

that agree with it gain weight and graphs that disagree lose weight, in proportion to the answer's confidence. Because all claim types update the same distribution, their expected information gain can be compared directly. Within a fixed query budget, CLAIM-FUSE therefore asks, at each step, the question expected to reduce uncertainty the most across four types of causal claims (directed edge, adjacency, ancestry, ordering), including claims about variables that are not directly connected.

Based on this framework, we investigate three main research questions: (RQ1) Does CLAIM-FUSE outperform existing LLM-assisted causal discovery baselines? (RQ2) Does actively querying from all claim types improve causal discovery over querying only local claims or only ancestral claims? (RQ3) Is CLAIM-FUSE cost-effective in terms of tokens and runtime compared with existing LLM-assisted methods? To address them, we make the following contributions:

1. We introduce CLAIM-FUSE, an LLM-assisted causal discovery framework that treats directed-edge, adjacency, ancestry, and ordering claims as yes/no questions on the same set of candidate graphs. At each step, it picks the variables and the claim type with the highest expected information gain.
2. We show that CLAIM-FUSE achieves the lowest structural error among LLM-assisted baselines, with experiments on two causal discovery benchmarks (RQ1).
3. We show that choosing among all claim types improves recovery compared to querying only local or only ancestral claims, at the same budget (RQ2).
4. We show that CLAIM-FUSE is more cost-effective than existing LLM-assisted methods, achieving the same or better performance with 53–1,700× fewer tokens (RQ3).

## 2 RELATED WORK

**Knowledge-guided causal discovery.** Causal discovery aims to recover the causal graph underlying observational data, but data alone often leave many graphs equally plausible (Spirtes et al., 2000; Chickering, 2002). Traditionally, this ambiguity has been reduced with background knowledge, encoded as priors over graphs (O'Donnell et al., 2006), expert opinions on pairwise relations (Amirkhani et al., 2017), or structural constraints such as required and forbidden edges, orders, and ancestral relations (Li & Beek, 2018; Hasan & Gani, 2023). Since such knowledge can be wrong, later work made its use robust to errors, for example, through weighted constraints (Chen et al., 2025a), uncertain priors (Wang et al., 2025; Xu & Corter, 2026), or explicit mechanisms that detect and correct erroneous constraints (Chen et al., 2025b; Wang et al., 2026; Baldo et al., 2026). Recent work also scales knowledge integration to larger graphs (Schubert et al., 2026) and richer model classes (Björkman et al., 2025). However, these methods take the available knowledge and its types as fixed inputs. In contrast, CLAIM-FUSE decides which knowledge to acquire, and of which type, during inference.

**LLMs as sources of causal knowledge.** Early work showed that LLMs can answer many pairwise causal questions from variable descriptions alone (Kiciman et al., 2023). Recent methods combine LLM answers with observational data: LLMs orient ambiguous edges (Long et al., 2023), provide causal orders (Vashishtha et al., 2025), supply edge constraints derived from token probabilities (Takayama et al., 2025), propose a graph that reorders conditional independence tests (Hiremath et al., 2025), guide a breadth-first search over the graph (Jiralerspong et al., 2024), act as priors over graphs (Darvariu et al., 2024), or reweight an ensemble of discovery algorithms (Peng et al., 2026; Li et al., 2026). These answers are not equally reliable across claim types: LLMs are generally more reliable on ancestral and ordering relations than on direct edges (Vashishtha et al., 2025). Estimating how much to trust each answer is therefore central, and consistency across sampled or rephrased responses is a common signal of confidence (Kuhn et al., 2023; Lyu et al., 2025; Hou et al., 2024). However, most of these methods query every pair or triplet of variables, which is costly, or rely on a single claim type. Our work instead asks only the most informative questions, chooses their type, and weights each answer by its own confidence.

**Active knowledge elicitation.** Active learning of informative experiments for causal discovery (Tong, 2001; Agrawal et al., 2019) has recently been extended from interventions to expert queries. Ankan & Textor (2025) consult an expert on variable pairs where the data suggest a missing or superfluous edge, and da Silva et al. (2026) query a noisy expert about relations in an ancestral graph, choosing the pair that minimizes expected entropy. Closest to our work, Bonilla et al. (2026) select which pair of variables to query by expected information gain over a particle posterior, and Verma et al. (2026) adaptively query an LLM using a bandit-style score. Tench & Demeester (2026) ask about both edge existence and orientation, with a fixed rule deciding which to ask. However, all of these methods choose which variables to query but fix the type of question, typically the local edge between two variables, a limitation that Bonilla et al. (2026) argue can lead to excessive queries when resolving global uncertainty. Our work is distinct because CLAIM-FUSE chooses both which variables and which type of claim to query, including ancestral and ordering claims about variables that are not directly connected.

## 3 CLAIM-FUSE

We introduce CLAIM-FUSE, a method to estimate a graph distribution from both observational data and heterogeneous causal claims. It first uses observational data to initialize a set of candidate causal graphs, and then reweights them according to how well each one agrees with current claims. At each iteration, it queries the causal relation where the candidate graphs disagree most, and uses the answer to reweight the graphs again, continuing until it reaches a pre-specified query budget.

### 3.1 PROBLEM FORMULATION

Given an observational dataset $D = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, where $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$ is an observation of the variables $V = \{X_1, \ldots, X_p\}$, and a knowledge source, such as an expert or LLM, our goal is to infer a distribution over causal graphs on $V$ from $D$ and heterogeneous causal claims with claim-specific confidence. Claims may be supplied in advance or elicited during inference. Let $G$ denote a causal graph on $V$ and $q_K(G)$ denote the graph distribution after incorporating $K$ claims, which we approximate by a weighted set of causal graphs $\{(G_j, w_j)\}_{j=1}^{J}$:

$$\widehat{q}_K(G) = \sum_{j=1}^{J} w_j \, \mathbf{1}\{G = G_j\} \approx q_K(G), \tag{3.1}$$

where $w_j \geq 0$ and $\sum_{j=1}^{J} w_j = 1$, and $\mathbf{1}$ is the indicator function.

**Causal claims.** A causal claim is a statement about a causal relation that can be evaluated on a causal graph. Local causal claims concern **directed edges** ($X \to Y$) or **adjacencies** ($X - Y$). Ancestral causal claims concern **ancestry** ($X \rightsquigarrow Y$) or **topological ordering** ($X \prec Y$). Each causal claim $k$ defines a binary graph predicate $f_k : \mathcal{G} \to \{0, 1\}$. For example, when evaluated on graph $G : X \to Y \to Z$, ancestry claim $X \rightsquigarrow Z$ returns 1 while directed edge claim $X \to Z$ returns 0. The extension to maximal ancestral graphs is described in Appendix G.

**Query on causal claims.** For a pair of variables $(X, Y)$, we define claim-specific question templates. Four question types corresponding to the predicates defined above are considered. Respectively: *"Does X directly cause Y?"*, *"Are X and Y directly connected, regardless of direction?"*, *"Does X cause Y, possibly through intermediate variables?"*, and *"Does X precede Y in the causal order?"*. The LLM returns a yes/no judgement or abstains, together with a confidence score. Full prompt templates are provided in Appendix E.2.

**Response likelihoods.** When causal claim $k$ is queried, let $Y_k \in \{0, 1\}$ denote the response variable, $y_k$ its observed value and $c_k \in (0, 1)$ the associated confidence, interpreted as the probability that the answer is correct. Under a symmetric error model, the source answers correctly with probability $c_k$ for either outcome. The likelihood of the observed response is

$$\ell_k(y_k \mid G) = P(Y_k = y_k \mid f_k(G), c_k) = \begin{cases} c_k, & f_k(G) = y_k, \\ 1 - c_k, & f_k(G) \neq y_k. \end{cases} \tag{3.2}$$

### 3.2 Combining Claims with Observations

Let $q_0(G)$ denote the data-only graph distribution and let $\mathbf{y}_{1:K} = (y_1, \ldots, y_K)$ denote the observed responses to the $K$ causal claims. Assuming the responses are conditionally independent, their joint likelihood is $L(\mathbf{y}_{1:K} \mid G) = \prod_{k=1}^{K} \ell_k(y_k \mid G)$, yielding the weighted distribution

$$q_K(G) = \frac{q_0(G) L(\mathbf{y}_{1:K} \mid G)}{Z_K}, \qquad Z_K = \sum_{G' \in \mathcal{G}} q_0(G') L(\mathbf{y}_{1:K} \mid G') > 0. \tag{3.3}$$

Observations and claim responses therefore jointly determine graph probabilities: high-confidence agreement increases graph weight, while responses with $c_k = 0.5$ leave $q_0$ unchanged. Equation 3.3 is an exact Bayesian update if $q_0$ is a data posterior, or likelihood-based reweighting otherwise.

### 3.3 Exploring Beyond the Initial Graphs

To recover graphs outside the candidate set, Claim-Fuse explores DAGs using sequential Monte Carlo (SMC), introducing claims gradually and proposing graph edits through Metropolis–Hastings (MH). At step $t$, SMC uses $J$ particles to approximate the distribution $q_t(G) \propto q_0(G) L(\mathbf{y}_{1:t} \mid G)$, where $\mathbf{y}_{1:t}$ denotes the responses incorporated up to step $t$. The graph particles are initialized using data-only discovery methods or a Markov chain without causal claims. Claims are then combined with observational data according to Equation 3.3. A new graph $G'$ is proposed by adding, deleting, or reversing an existing edge with probability

$$Q_\gamma(G' \mid G) = (1 - \gamma) Q_{\text{ord}}(G' \mid G) + \gamma Q_{\text{guide}}(G' \mid G). \tag{3.4}$$

The ordinary proposal $Q_{\text{ord}}$ randomly adds, deletes, or reverses edges. The guided proposal selects a contradicted claim, prioritizing larger likelihood odds, and attempts a local edit that changes disagreement into agreement. Candidate graphs are accepted with probability:

$$\alpha_t(G, G') = \min \left\{ 1, \ \frac{q_t(G') Q_\gamma(G \mid G')}{q_t(G) Q_\gamma(G' \mid G)} \right\}. \tag{3.5}$$

The details can be found in Appendix A.

### 3.4 Choosing the Next Causal Claim to Query

Exhaustive elicitation is usually sparse and may contain information that is already given by observation data. To avoid this, we sequentially query causal claims that maximize expected information gain (EIG). At step $t$, with current graph particles $\{(G_j, w_j)\}_{j=1}^{J}$ and assuming a constant response confidence $c$, the predictive distribution for answer $Y_k \in \{0, 1\}$ to an unqueried causal claim $k$ is

$$\mathbb{P}_t(Y_k = y) = \sum_{j=1}^{J} w_j \, \mathbb{P}(Y_k = y \mid f_k(G_j), c). \tag{3.6}$$

The particle-based EIG measures the expected reduction in graph uncertainty:

$$\mathrm{EIG}_t(k) = H(\mathbb{P}_t(Y_k)) - \sum_{j=1}^{J} w_j \, H(\mathbb{P}(Y_k \mid f_k(G_j), c)), \tag{3.7}$$

where $H$ is Shannon entropy. Under constant accuracy $c$ for Equation 3.6, this prioritizes queries with maximal particle disagreement

$$k^* = \arg\max_k \mathrm{EIG}_t(k). \tag{3.8}$$

The search space is the causal claim pool, which can be local, ancestral, or their union. The graph weights are updated based on the elicited answer and confidence, and repeated until reaching the query budget $B$ (Algorithm 1).

**Algorithm 1** CLAIM-FUSE with active querying

1: **Input:** Data $D$, answer source $\mathcal{E}$, candidate graphs number $J$, query budget $B$, ESS fraction $\tau$
2: Initialize weighted graph set $\{(G_j, w_j)\}_{j=1}^{J}$ from data $D$.
3: **for** $t = 0, \ldots, B-1$ **do**
4: **Query:** Choose causal relation $k^* \in \arg\max_k \mathrm{EIG}_t(k)$.
5: Query $\mathcal{E}$ for claim $k^*$, obtaining response $y_{k^*}$ and confidence $c_{k*}$.
6: Update weights $w_j \leftarrow w_j, \ell_{k^*}(y_{k^*} \mid G_j)$ and then normalize.
7: **Resample:** If ESS $< \tau J$, resample and set $w_j \leftarrow 1/J$.
8: **Edit:** After resampling, apply MH moves (Equation 3.5).
9: **end for**
10: **return** Weighted particles $\{(G_j, w_j)\}_{j=1}^{J}$.

### 3.5 KNOWLEDGE-DRIVEN IMPROVEMENTS IN DISCOVERY

The graph discovery performance may degrade because of wrong answers or overconfidence. We show by numerically bounding every confidence $c_k$ used in Algorithm 1, CLAIM-FUSE preserves consistency of the data-only discovery method used to initialize candidate graphs.

**Assumption 1** (Bounded update weights)**.** *There exists a fixed $\delta \in (0, 1/2)$ such that*

$$\delta \le c_k \le 1 - \delta \text{ for every query } k. \tag{3.9}$$

**Proposition 1** (Preservation of consistency under bounded claim updates)**.** *Under Assumption 1, let $q_0^{(n)}$ be a graph distribution based on $n$ observations over a fixed finite graph space $\mathcal{G}$, and let $q_B^{(n)}$ be its exact normalized update after $B$ claims, where $B$ is fixed independently of $n$. For a fixed target set $\mathcal{S} \subseteq \mathcal{G}$, write $q(\mathcal{S}) = \sum_{G \in \mathcal{S}} q(G)$. Then*

$$1 - q_B^{(n)}(\mathcal{S}) \le \left(\frac{1-\delta}{\delta}\right)^B \left[1 - q_0^{(n)}(\mathcal{S})\right]. \tag{3.10}$$

*Consequently, if $q_0^{(n)}(\mathcal{S}) \to 1$ as $n \to \infty$, then $q_B^{(n)}(\mathcal{S}) \to 1$, with convergence in probability or almost surely inherited from the initial distribution.*

Let $\mathcal{Q}_{\mathrm{homo}}$ be a query pool restricted to one causal-claim type, and let $\mathcal{Q}_{\mathrm{het}}$ contain this pool together with queries of additional claim types. Thus, $\mathcal{Q}_{\mathrm{homo}} \subseteq \mathcal{Q}_{\mathrm{het}}$. The following proposition shows the benefit of incorporating more causal claim types for CLAIM-FUSE.

**Proposition 2** (Information gain from heterogeneous claims)**.** *Under Assumption 1, fix a current graph distribution $q$. Write*

$$p_k = \mathbb{P}_{G \sim q}\big(f_k(G) = 1\big). \tag{3.11}$$

*Then the best expected information gain under the response model satisfies*

$$\max_{k \in \mathcal{Q}_{\mathrm{het}}} \mathrm{EIG}(k) \ge \max_{k \in \mathcal{Q}_{\mathrm{homo}}} \mathrm{EIG}(k). \tag{3.12}$$

*If $c \neq 1/2$, the inequality is strict whenever an added claim $k \in \mathcal{Q}_{\mathrm{het}} \setminus \mathcal{Q}_{\mathrm{homo}}$ satisfies*

$$\left|p_k - \tfrac{1}{2}\right| < \min_{j \in \mathcal{Q}_{\mathrm{homo}}} \left|p_j - \tfrac{1}{2}\right|. \tag{3.13}$$

*Thus, with a common predictive confidence different from $1/2$, an additional claim type can strictly increase the modelled information gain from the next query by providing a more balanced split of the current graph probabilities.*

# 4 EXPERIMENTS

We evaluate CLAIM-FUSE across three key research questions (RQs). We begin by assessing its **performance**, benchmarking causal discovery accuracy against both data-only and LLM-assisted baselines (RQ1). Next, we investigate **heterogeneous claims** by measuring the gains achieved when combining diverse claim types over a single claim type (RQ2). Finally, we analyze **cost trade-offs** to verify whether these improvements hold under constrained query, token, and runtime budgets without exhaustive elicitation (RQ3).

**Benchmarks and metrics.** We evaluate our approach across sample sizes of $n \in \{200, 1000\}$ on three distinct domains: synthetic 20-node linear Gaussian SCMs, a 20-node binary neuropathic pain subnetwork (Tu et al., 2019), and the 37-variable categorical ALARM network (Beinlich et al., 1989). Evaluation relies on two primary metrics: expected distance (ED), computed as the mean structural Hamming distance (SHD) between the estimated and true DAGs, and the orientation F1 score over directed edges in the highest-weight graph.

**LLM Confidence.** We use stated confidence and semantic consistency to assign confidence to each causal claim required by CLAIM-FUSE. Stated confidence relies on a single LLM query and its self-reported confidence. Semantic consistency (Lyu et al., 2025) measures response stability across paraphrased but semantically consistent queries. Confidence is calculated as the proportion of valid judgments that support the majority. See Appendices C, E.2, and F.2 for full details.

### RQ1: DOES CLAIM-FUSE OUTPERFORM EXISTING LLM-ASSISTED CAUSAL DISCOVERY BASELINES?

We test whether CLAIM-FUSE improves causal graph discovery compared to data-only and LLM-assisted baselines. Tables 1 and 2 report ED and orientation F1 at $n \in \{200, 1000\}$ over ten seeds on neuropathic pain with GPT-5.6 Terra and Qwen3.5-9B and ALARM with Qwen3.5-9B, while retaining each baseline's native query and confidence procedure. CLAIM-FUSE achieves the lowest mean ED and highest mean orientation F1 in each model–sample-size comparison.

Not all methods perform evenly on the neuropathic pain and ALARM datasets. BFS+Pearson performs worse on ALARM at both sample sizes, flattening out around an expected distance (ED) of 40 to 45 no matter the value of $n$. BFS+Pearson fails because it depends on an early choice made by the LLM that more data cannot fix. The method sets edge directions using a breadth-first search

Table 1. Performance of CLAIM-FUSE and baselines on the neuropathic pain dataset. Values are mean (SE) over ten seeds with masked identifiers. All CLAIM-FUSE rows use all claim types and 100 attempted queries; bold marks the best mean within each model block. SCP Pattern 2 and PC+Long-$S_{\text{risk}}$ have no GPT-5.6 Terra results because their native LLM decision steps require output-token log probabilities unavailable through the Terra interface used here.

| | $n = 200$ | | $n = 1000$ | |
|---|---|---|---|---|
| Method | ED ↓ | Orient. F1 ↑ | ED ↓ | Orient. F1 ↑ |
| *Data-Only* | | | | |
| PC | 23.75 (0.63) | 0.149 (0.029) | 21.00 (0.39) | 0.241 (0.025) |
| GES | 20.05 (0.34) | 0.443 (0.012) | 18.55 (0.28) | 0.528 (0.009) |
| *GPT-5.6 Terra* | | | | |
| gPC-Guess | 24.00 (0.71) | 0.237 (0.032) | 22.95 (0.83) | 0.309 (0.025) |
| BFS+Pearson | 28.80 (0.79) | 0.280 (0.016) | 28.70 (0.37) | 0.282 (0.010) |
| Causal Order+PC | 18.90 (0.64) | 0.559 (0.020) | 16.00 (0.30) | 0.632 (0.013) |
| CLAIM-FUSE (stated) | **15.23 (0.38)** | **0.712 (0.007)** | 13.57 (0.50) | 0.751 (0.016) |
| CLAIM-FUSE (consistency) | 15.77 (0.49) | 0.695 (0.007) | **12.59 (0.31)** | **0.758 (0.007)** |
| *Qwen3.5-9B* | | | | |
| gPC-Guess | 23.55 (0.62) | 0.256 (0.029) | 20.50 (0.67) | 0.404 (0.032) |
| SCP Pattern 2 | 18.80 (0.49) | 0.547 (0.016) | 15.70 (0.54) | 0.647 (0.016) |
| BFS+Pearson | 17.50 (1.42) | 0.571 (0.064) | 20.90 (1.72) | 0.485 (0.082) |
| Causal Order+PC | 18.90 (0.64) | 0.559 (0.020) | 16.00 (0.30) | 0.632 (0.013) |
| PC+Long-$S_{\text{risk}}$ | 21.65 (0.66) | 0.431 (0.025) | 18.83 (0.33) | 0.511 (0.016) |
| CLAIM-FUSE (stated) | **15.40 (0.72)** | **0.705 (0.020)** | 14.29 (0.68) | 0.731 (0.017) |
| CLAIM-FUSE (consistency) | 15.52 (0.28) | 0.701 (0.012) | **12.66 (0.62)** | **0.764 (0.012)** |

from one start point with Pearson correlation filtering, meaning an early mistake is never fixed. This is a much bigger problem on ALARM than on neuropathic pain. The longest path in ALARM is ten edges long, while in neuropathic pain it is only three edges long. As a result, a wrong choice near the start can spread across only two more steps in neuropathic pain, but it can ruin up to nine steps in ALARM.

CLAIM-FUSE avoids this failure mode. Rather than relying on one upfront estimation, it keeps a weighted population of candidate graphs and lets every acquired claim reweight all of them at once, so an early wrong belief loses weight instead of being locked in. The resampling and guided Metropolis-Hastings edits in the Edit step can then drop a low-weight graph and propose a new one as further claims arrive, so the search stays free to change. This is why CLAIM-FUSE improves on ALARM as $n$ grows rather than plateauing.

Table 2. Performance of CLAIM-FUSE and baselines on the ALARM dataset. Results are mean (SE) over ten paired seeds (gPC-Guess at $n = 1000$: three verified CPDAGs); Qwen3.5-9B is the LLM. All CLAIM-FUSE rows use all claim types and 50 attempted queries; bold marks the best ten-seed mean per column.

| | $n = 200$ | | $n = 1000$ | |
|---|---|---|---|---|
| Method | ED ↓ | Orient. F1 ↑ | ED ↓ | Orient. F1 ↑ |
| *Data-Only* | | | | |
| PC | 30.60 (1.03) | 0.352 (0.031) | 16.05 (0.56) | 0.687 (0.015) |
| GES | 22.65 (1.75) | 0.642 (0.033) | 11.85 (1.40) | 0.808 (0.023) |
| NOTEARS | 46.70 (0.87) | 0.234 (0.010) | 46.20 (0.84) | 0.231 (0.017) |
| *Qwen3.5-9B* | | | | |
| gPC-Guess | 44.60 (1.09) | 0.253 (0.020) | 39.33 (2.13) | 0.409 (0.026) |
| BFS+Pearson | 43.30 (2.11) | 0.315 (0.021) | 43.70 (1.02) | 0.296 (0.019) |
| PC+Long-$S_{\text{risk}}$ | 29.80 (0.97) | 0.466 (0.024) | 15.02 (0.57) | 0.727 (0.015) |
| Causal Order+PC | 27.30 (0.98) | 0.543 (0.021) | 13.60 (0.64) | 0.780 (0.013) |
| SCP Pattern 2 | 27.65 (0.93) | 0.509 (0.024) | 12.20 (0.60) | 0.808 (0.011) |
| CLAIM-FUSE (stated) | **22.63 (1.86)** | **0.694 (0.032)** | **11.55 (1.17)** | 0.834 (0.021) |
| CLAIM-FUSE (consistency) | 24.65 (2.10) | 0.686 (0.027) | 12.13 (1.34) | **0.835 (0.023)** |

Additional experiments run on synthetic data show similar improvements, see Appendix D.

**RQ2:** DOES ACTIVELY QUERYING FROM ALL CLAIM TYPES IMPROVE CAUSAL DISCOVERY OVER QUERYING ONLY LOCAL CLAIMS OR ONLY ANCESTRAL CLAIMS?

We evaluate whether, for CLAIM-FUSE, access to the union of causal claims results in better graph discovery performance rather than restricting to one. On the neuropathic pain and ALARM datasets, the local, ancestral, and all-claim pools share their initialization, active querying rule, LLM inference seeds, and confidence treatment at a matched query budget, so the union only enlarges the relations

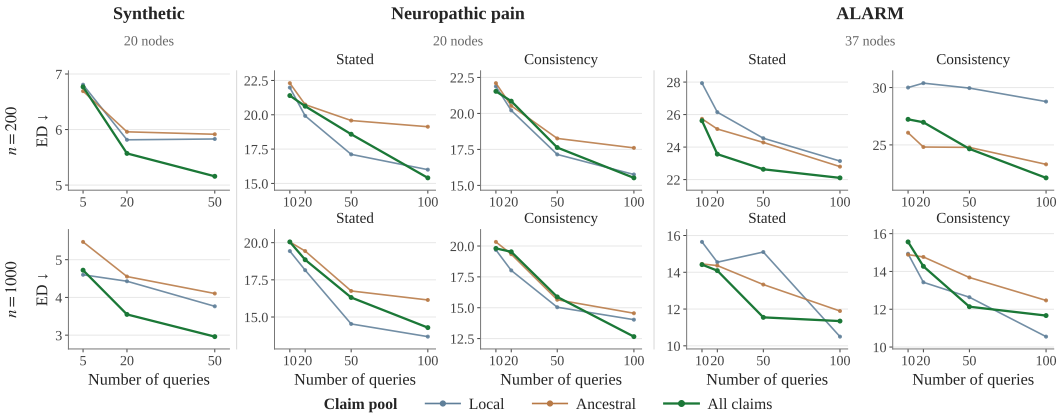


Figure 2. Performance of CLAIM-FUSE with different query budget and claim pools. Each panel shows mean ED against the number of queries for the local, ancestral, and all-claim pools (color), at $n = 200$ (solid) and $n = 1000$ (dashed), with simulated-expert accuracy $a \in \{0.6, 0.8, 1.0\}$.

available to EIG without increasing the number of acquired answers. Figure 2 shows that the all-claim pool achieves the lowest mean ED in most matched conditions on the synthetic, neuropathic pain, and ALARM datasets. The advantage of using all claims is more prominent with more queries.

This advantage is because supporting more types of claims gives CLAIM-FUSE more ways to distinguish candidate graphs. Local and ancestral claims can each fail to distinguish two graphs that the other type tells apart. This is why the all-claims pool reaches the lowest mean ED on the synthetic dataset once the budget is large enough, and why it is also the best pool in most matched conditions on neuropathic pain and ALARM.

However, the advantage depends on the type-specific accuracy. On ALARM at $n = 1000$ and 100 queries, the local pool alone reaches a lower ED than the union. This is because, on the ALARM dataset, Qwen3.5-9B has higher accuracy on local claims than on ancestral claims, while CLAIM-FUSE is more likely to query ancestral relations than local ones, as shown in Table 3. If we know in advance the accuracy of an answer source on different claim types, then CLAIM-FUSE allows us to restrict claims to a specific group.

Table 3 also shows that the measured accuracy of the same LLM changes with sample size. For GPT-5.6 Terra with stated confidence, local accuracy drops from 86.1% at n = 200 to 71.8% at n = 1000. Since the model and prompts are unchanged, this reflects which questions are selected: with more data, easy relations are already resolved, so EIG directs queries toward relations that are harder for both the data and the LLM.

Table 3. Accuracy and proportion of local and ancestral claims queried by CLAIM-FUSE, with 100 queries across all types. Each cell reports accuracy/proportion, where accuracy is the correctness of the corresponding type's answers, and proportion is the percentage of queries that were of that type.

| $n$ | Confidence | Local | Ancestral |
|---|---|---|---|
| *Neuropathic pain: GPT-5.6 Terra* | | | |
| 200 | Stated | 86.1% / 39.4% | 92.5% / 60.6% |
| 200 | Consistency | 86.2% / 38.2% | 94.0% / 61.8% |
| 1000 | Stated | 71.8% / 35.1% | 91.5% / 64.9% |
| 1000 | Consistency | 75.1% / 34.8% | 92.2% / 65.2% |
| *Neuropathic pain: Qwen3.5-9B* | | | |
| 200 | Stated | 81.3% / 40.8% | 79.4% / 59.2% |
| 200 | Consistency | 81.9% / 41.2% | 84.7% / 58.8% |
| 1000 | Stated | 75.4% / 32.6% | 75.8% / 67.4% |
| 1000 | Consistency | 77.6% / 33.9% | 83.1% / 66.1% |
| *ALARM: Qwen3.5-9B* | | | |
| 200 | Stated | 72.8% / 21.7% | 64.4% / 78.3% |
| 200 | Consistency | 70.4% / 26.4% | 68.4% / 73.6% |
| 1000 | Stated | 81.1% / 20.7% | 67.7% / 79.3% |
| 1000 | Consistency | 70.6% / 22.0% | 70.5% / 78.0% |

**RQ3:** IS CLAIM-FUSE COST-EFFECTIVE IN TERMS OF TOKENS AND RUNTIME COMPARED WITH EXISTING LLM-ASSISTED METHODS?

We ask whether CLAIM-FUSE achieves the best performance on the neuropathic pain and ALARM datasets with a favorable cost-accuracy trade-off. Using Qwen3.5-9B on neuropathic pain and ALARM at $n \in \{200, 1000\}$, Figure 3 plots mean ED against runtime with marker area proportional to input-plus-output tokens, comparing all-claim CLAIM-FUSE at 10, 20, 50, and 100 queries under stated and consistency confidence with LLM-assisted baselines. Figure 3 shows a favorable accuracy-cost trade-off: the 100-attempt all-claim stated variant has lower mean ED than SCP Pattern 2 and Causal Order+PC with fewer tokens and less estimated runtime on both datasets, while some cheaper baselines have higher ED.

This advantage comes from spending the query budget adaptively instead of covering every relation. Causal Order+PC asks about every unordered variable triplet, and SCP Pattern 2 asks about every ordered pair, so both cost more as the graph grows. Causal Order+PC's triplet numbers are 1,140 on

the 20-node neuropathic subnetwork and 7,770 on the 37-node ALARM network, and SCP Pattern 2 let the LLM evaluate 1,332 variable pairs on ALARM. At $n = 1000$ this costs Causal Order+PC 2.00 million tokens and 21.7 hours on neuropathic pain and 4.79 million tokens and 2.6 hours on ALARM, and it costs SCP Pattern 2 13.89 million tokens and 9.6 hours on neuropathic pain and 67.0 million tokens and 9.8 hours on ALARM. CLAIM-FUSE instead reaches a lower ED on both datasets using about 38 to 39 thousand tokens and 8 to 9 minutes, a reduction of roughly 53 to over 1,700 times in tokens and 20 to 155 times in runtime.

Two properties explain this gap. First, exhaustive querying spends cost on many uninformative queries that may already be resolved by observation data. In contrast, active query checks every unasked claim across all supported types and asks whichever one most reduces uncertainty over the current candidate graphs. Supporting different types of causal claims also contributes to this. A directed-edge test cannot separate $X \to Y \to Z$ from $X \leftarrow Y \leftarrow Z$, since both agree that neither $X$ nor $Z$ has a direct edge to the other, while an ancestral test still carries a signal. Moreover, each query only names two variables, without restating the graph, the sample, or earlier answers, so its cost stays about the same no matter the number of nodes, the sample size, or how many questions came before, unlike the baselines above, whose total number of questions is fixed by the graph size.

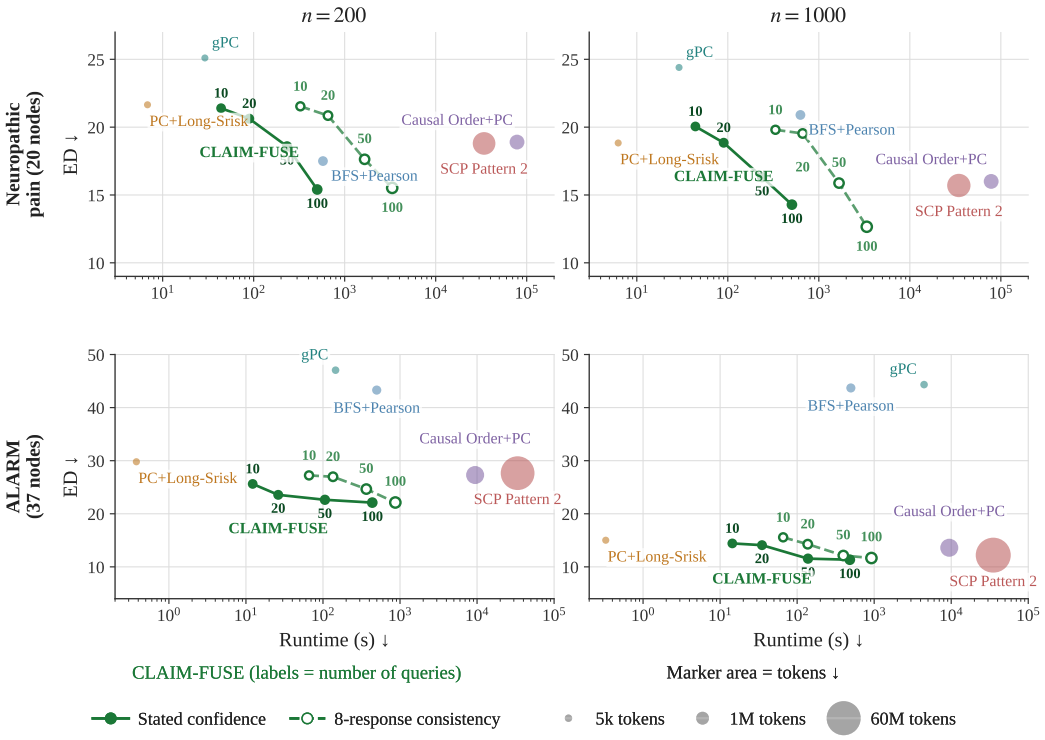


Figure 3. Performance, runtime, and token usage of CLAIM-FUSE and baselines using Qwen3.5-9B on neuropathic pain (top) and ALARM (bottom) at $n = 200$ (left) and $n = 1000$ (right). Green trajectories show all-claim CLAIM-FUSE with stated confidence (solid) and eight-response consistency (dashed); labels give attempted-query budgets. Other labeled points are LLM-assisted baselines. Marker area encodes input-plus-output tokens, and runtime is logarithmic; lower ED, shorter runtime, and fewer tokens are better.

## 5 CONCLUSION

In causal discovery, how external knowledge is represented and integrated with observational data determines the accuracy and efficiency of the resulting method. This becomes even more important when the knowledge source is an LLM, since exhaustive elicitation quickly becomes slow and expensive. CLAIM-FUSE represents and elicits knowledge as heterogeneous causal claims, including directed edges, adjacency, ancestry, and topological ordering. Compared with evaluated baselines, it recovers causal graphs more accurately with lower cost. Our results show that this combination of accuracy and efficiency is achieved by actively querying the most informative causal relations from a richer set of claim types.

### AI USE STATEMENT

In this work, we used generative AI tools for formulating mathematical claims, assisting in the writing of proofs, implementing methods, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, interpreting results, and assisting with translation is not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images, creating or editing software code, summarizing or analyzing existing literature, brainstorming, identifying relevant literature.

We have reviewed all AI-assisted work. LLMs are used only to assist in technical part of the mathematical proofs, with a clear breakdown of the problem and target clearly specified by the authors. All results are independently evaluated by the authors. LLM-generated code was reviewed and tested for correctness by 2 authors. For the related work, LLMs are only used to identify and retrieve relevant papers. The authors read the original paper rather than LLM-generated summaries. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

All datasets used in this work are publicly available and, to the best of our knowledge, accurately reported. We made a best effort to compare against contemporary baselines in a fair manner, using each method's native query and confidence procedures. We do not foresee direct harmful impact from this work. However, because our approach uses LLMs as a source of causal claims, and because two of our benchmarks are drawn from clinical and biomedical settings, we urge suitable prudence before applying CLAIM-FUSE or similar LLM-assisted causal discovery methods in consequential applications, given the risk of miscalibrated or unreliable claims.

### REPRODUCIBILITY STATEMENT

Anonymized code is included in the supplementary material. Section 3 specifies the proposed method and its assumptions, while Appendix A provides additional algorithmic details and Appendix B contains complete proofs of the theoretical results. Section 4 describes the evaluation design, with metrics defined in Appendix C. The synthetic data generation procedure is documented in Appendix D. The neuropathic pain and ALARM datasets, preprocessing, prompts are detailed in Appendices E and F. The anonymous supplementary code provides the implementation, dependency requirements, tests, fixed seeds, and commands for rerunning the reported experiments. It also records used prompts, model-provider settings, responses, and result files. However, because the hosted model may change over time, rerunning the experiments with new API requests may produce results that differ from those reported in the paper.

## REFERENCES

Raj Agrawal, Chandler Squires, Karren D. Yang, Karthikeyan Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3400–3409. PMLR, 2019. URL http://proceedings.mlr.press/v89/agrawal19b.html.

R. Ayesha Ali, Thomas S. Richardson, and Peter Spirtes. Markov equivalence for ancestral graphs. *The Annals of Statistics*, 37(5B), October 2009. ISSN 0090-5364. doi: 10.1214/08-aos626. URL http://dx.doi.org/10.1214/08-AOS626.

Hossein Amirkhani, Mohammad Rahmati, Peter J. F. Lucas, and Arjen Hommersom. Exploiting experts' knowledge for structure learning of bayesian networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2154–2170, 2017. doi: 10.1109/TPAMI.2016.2636828. URL https://doi.org/10.1109/TPAMI.2016.2636828.

Ankur Ankan and Johannes Textor. Expert-in-the-loop causal discovery: Iterative model refinement using expert knowledge. In Silvia Chiappa and Sara Magliacane (eds.), *Conference on Uncertainty in Artificial Intelligence, Rio Othon Palace, Rio de Janeiro, Brazil, 21-25 July 2025*, volume 286 of *Proceedings of Machine Learning Research*, pp. 172–183. PMLR, 2025. URL https://proceedings.mlr.press/v286/ankan25a.html.

Federico Baldo, Simon Ferreira, and Charles K. Assaad. Retrieving classes of causal orders with inconsistent knowledge bases. In *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pp. 393–422. PMLR, 2026. URL https://proceedings.mlr.press/v323/baldo26a.html.

Ingo A. Beinlich, H. J. Suermondt, R. Martin Chavez, and Gregory F. Cooper. The alarm monitoring system: A case study with two probabilistic inference techniques for belief networks. In Jim Hunter, John Cookson, and Jeremy Wyatt (eds.), *AIME 89*, pp. 247–256, Berlin, Heidelberg, 1989. Springer Berlin Heidelberg. ISBN 978-3-642-93437-7.

Zachris Björkman, Jorge Loría, Sophie Wharrie, and Samuel Kaski. Incorporating expert knowledge into bayesian causal discovery of mixtures of directed acyclic graphs. *CoRR*, abs/2510.06735, 2025. doi: 10.48550/ARXIV.2510.06735. URL https://doi.org/10.48550/arXiv.2510.06735.

Edwin V. Bonilla, He Zhao, and Daniel M. Steinberg. Causal preference elicitation. *CoRR*, abs/2602.01483, 2026. doi: 10.48550/ARXIV.2602.01483. URL https://doi.org/10.48550/arXiv.2602.01483.

Lyuzhou Chen, Taiyu Ban, Derui Lyu, Yijia Sun, Kangtao Hu, Xiangyu Wang, and Huanhuan Chen. Continuous structure constraint integration for robust causal discovery. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Mohammad Emtiyaz Khan (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2025, Mai Khao, Thailand, 3-5 May 2025*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3160–3168. PMLR, 2025a. URL https://proceedings.mlr.press/v258/chen25g.html.

Lyuzhou Chen, Taiyu Ban, Xiangyu Wang, Derui Lyu, and Huanhuan Chen. Mitigating prior errors in causal structure learning: A resilient approach via bayesian networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(12):10990–11002, 2025b. doi: 10.1109/TPAMI.2025.3594755. URL https://doi.org/10.1109/TPAMI.2025.3594755.

David Maxwell Chickering. Optimal structure identification with greedy search. *J. Mach. Learn. Res.*, 3:507–554, 2002. URL https://jmlr.org/papers/v3/chickering02b.html.

Tiago da Silva, Bruna Bazaluk, Eliezer de Souza da Silva, António Góis, Salem Lahlou, Dominik Heider, Samuel Kaski, Diego Mesquita, and Adèle Helena Ribeiro. Expert-aided causal discovery of ancestral graphs. *Inf. Sci.*, 756:123816, 2026. doi: 10.1016/J.INS.2026.123816. URL https://doi.org/10.1016/j.ins.2026.123816.

Victor-Alexandru Darvariu, Stephen Hailes, and Mirco Musolesi. Large language models are effective priors for causal graph discovery. *CoRR*, abs/2405.13551, 2024. doi: 10.48550/ARXIV.2405.13551. URL https://doi.org/10.48550/arXiv.2405.13551.

Uzma Hasan and Md Osman Gani. Optimizing data-driven causal discovery using knowledge-guided search, 2023. URL https://arxiv.org/abs/2304.05493. Revised 2024.

Sujai Hiremath, Dominik Janzing, Philipp Michael Faller, Patrick Blöbaum, Elke Kirschbaum, Shiva Prasad Kasiviswanathan, and Kyra Gan. From guess2graph: When and how can unreliable experts safely boost causal discovery in finite samples? *CoRR*, abs/2510.14488, 2025. doi: 10.48550/ARXIV.2510.14488. URL https://doi.org/10.48550/arXiv.2510.14488.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In Ruslan Salakhutdinov, Zico Kolter, Katherine A. Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix

Berkenkamp (eds.), *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19023–19042. PMLR / OpenReview.net, 2024. URL https://proceedings.mlr.press/v235/hou24b.html.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. *CoRR*, abs/2402.01207, 2024. doi: 10.48550/ARXIV.2402.01207. URL https://doi.org/10.48550/arXiv.2402.01207.

Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *CoRR*, abs/2305.00050, 2023. doi: 10.48550/ARXIV.2305.00050. URL https://doi.org/10.48550/arXiv.2305.00050.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Andrew C. Li and Peter Beek. Bayesian network structure learning with side constraints. In Milan Studený and Václav Kratochvíl (eds.), *International Conference on Probabilistic Graphical Models, PGM 2018, 11-14 September 2018, Prague, Czech Republic*, volume 72 of *Proceedings of Machine Learning Research*, pp. 225–236. PMLR, 2018. URL http://proceedings.mlr.press/v72/li18a.html.

Xinyu Li, Yuanyuan Wang, Haoxuan Li, Chuan Zhou, Erdun Gao, Bo Han, Tongliang Liu, Kun Zhang, Howard D. Bondell, and Mingming Gong. Causal ensemble agent: Hierarchical causal discovery with llm-guided expert reweighting. *CoRR*, abs/2606.10607, 2026. doi: 10.48550/ARXIV.2606.10607. URL https://doi.org/10.48550/arXiv.2606.10607.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. *CoRR*, abs/2307.02390, 2023. doi: 10.48550/ARXIV.2307.02390. URL https://doi.org/10.48550/arXiv.2307.02390.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating large language models with sample consistency. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 19260–19268. AAAI Press, 2025. doi: 10.1609/AAAI.V39I18.34120. URL https://doi.org/10.1609/aaai.v39i18.34120.

Rodney T. O'Donnell, Ann E. Nicholson, B. Han, Kevin B. Korb, M. J. Alam, and Lucas R. Hope. Causal discovery with prior information. In Abdul Sattar and Byeong-Ho Kang (eds.), *AI 2006: Advances in Artificial Intelligence, 19th Australian Joint Conference on Artificial Intelligence, Hobart, Australia, December 4-8, 2006, Proceedings*, volume 4304 of *Lecture Notes in Computer Science*, pp. 1162–1167. Springer, 2006. doi: 10.1007/11941439_141. URL https://doi.org/10.1007/11941439_141.

Bo Peng, Kaiwen Wu, Sirui Chen, Zhiheng Wang, Yu Qiao, and Chaochao Lu. Caution: Knowing when to trust llms for ensemble causal discovery. *CoRR*, abs/2606.03602, 2026. doi: 10.48550/ARXIV.2606.03602. URL https://doi.org/10.48550/arXiv.2606.03602.

Mátyás Schubert, Theofanis Aslanidis, Tom Claassen, and Sara Magliacane. Integrating background knowledge for scalable causal discovery. *CoRR*, abs/2607.10456, 2026. doi: 10.48550/ARXIV.2607.10456. URL https://doi.org/10.48550/arXiv.2607.10456.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press, 2000. ISBN 978-0-262-19440-2.

Masayuki Takayama, Tadahisa Okuda, Thong Pham, Tatsuyoshi Ikenoue, Shingo Fukuma, Shohei Shimizu, and Akiyoshi Sannai. Integrating large language models in causal discovery: A statistical causal approach. *Trans. Mach. Learn. Res.*, 2025, 2025. URL https://openreview.net/forum?id=Reh1S8rxfh.

Adrick Tench and Thomas Demeester. Dynamic expert-guided model averaging for causal discovery. *CoRR*, abs/2601.16715, 2026. doi: 10.48550/ARXIV.2601.16715. URL https://doi.org/10.48550/arXiv.2601.16715.

Simon Tong. *Active learning: theory and applications*. PhD thesis, Stanford University, USA, 2001. URL https://searchworks.stanford.edu/view/4762444.

Ruibo Tu, Kun Zhang, Bo C. Bertilson, Hedvig Kjellström, and Cheng Zhang. Neuropathic pain diagnosis simulator for causal discovery algorithm evaluation. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 12773–12784, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/4fdaa19b1f22a4d926fce9bfc7c61fa5-Abstract.html.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. Causal order: The key to leveraging imperfect experts in causal inference. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=9juyeCqL0u.

Prakhar Verma, David Arbour, Sunav Choudhary, Harshita Chopra, Arno Solin, and Atanu R. Sinha. Sequential causal discovery with noisy language model priors. *Trans. Mach. Learn. Res.*, 2026, 2026. URL https://openreview.net/forum?id=wFs71JzEO7.

Zidong Wang, Xiaoguang Gao, and Qingfu Zhang. Uncertain priors for graphical causal models: A multi-objective optimization perspective. *IEEE Trans. Knowl. Data Eng.*, 37(12):7426–7439, 2025. doi: 10.1109/TKDE.2025.3608723. URL https://doi.org/10.1109/TKDE.2025.3608723.

Zidong Wang, Xi Lin, Chuchao He, and Xiaoguang Gao. Robust causal discovery under imperfect structural constraints. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 36757–36765. AAAI Press, 2026. doi: 10.1609/AAAI.V40I43.41001. URL https://doi.org/10.1609/aaai.v40i43.41001.

Guoliang Xu and James E. Corter. Kg-softmap: Soft knowledge-graph priors for bayesian network structure learning from sparse discrete data. *CoRR*, abs/2606.10358, 2026. doi: 10.48550/ARXIV.2606.10358. URL https://doi.org/10.48550/arXiv.2606.10358.

Jiji Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artif. Intell.*, 172(16-17):1873–1896, 2008. doi: 10.1016/J.ARTINT.2008.08.001. URL https://doi.org/10.1016/j.artint.2008.08.001.

## APPENDIX CONTENTS

**A Claim-Fuse** **15**

A.1 Different types of causal relations . . . . . 15

A.2 CLAIM-FUSE-FS . . . . . 15

A.3 CLAIM-FUSE-SMC-PASSIVE and CLAIM-FUSE-SMC-ACTIVE . . . . . 15

**B Theoretical Analysis** **16**

B.1 Proof of Proposition 1 . . . . . 16

B.2 Proof of Proposition 2 . . . . . 17

**C Evaluation Metrics** **17**

C.1 Graph Discovery Metrics . . . . . 17

C.2 Confidence and Query-Cost Metrics . . . . . 18

**D Synthetic Experiments** **18**

D.1 Shared Pairwise Expert Knowledge . . . . . 19

D.2 Claim Families with Uncertain Experts . . . . . 20

D.3 Miscalibration of Causal Claims . . . . . 20

**E Neuropathic Pain Dataset Experiments** **21**

E.1 20-node sub-network . . . . . 21

E.2 CLAIM-FUSE causal query interface . . . . . 22

**F ALARM Dataset Experiments** **22**

F.1 Overview of the ALARM dataset . . . . . 22

F.2 CLAIM-FUSE active querying interface . . . . . 23

**G Extension to Latent Confounding** **23**

G.1 Graph and Claims under Hidden Confounding . . . . . 23

G.2 Experiment Setup . . . . . 24

G.3 Results . . . . . 24

# A CLAIM-FUSE

## A.1 DIFFERENT TYPES OF CAUSAL RELATIONS

Table 4 presents causal claim types used in prior work, grouped by their graphical implications. CLAIM-FUSE-FS supports local, ancestral, and latent-confounding causal claims through finite-support inference over DAG or MAG candidates. CLAIM-FUSE-SMC-PASSIVE and CLAIM-FUSE-SMC-ACTIVE assume causal sufficiency and support only local and ancestral causal claims.

Table 4. External causal-claim types and representative prior work.

| Causal claim | Native form | Representative prior work |
|---|---|---|
| **Local causal claims: edge and adjacency** | | |
| Directed-edge inclusion | $X_i \to X_j$ | (Hasan & Gani, 2023; Darvariu et al., 2024; Ankan & Textor, 2025; Schubert et al., 2026; Wang et al., 2026; Xu & Corter, 2026; Bonilla et al., 2026) |
| Directed-edge exclusion | $X_i \not\to X_j$ | (Hasan & Gani, 2023; Wang et al., 2026) |
| Adjacency | $X_i \sim X_j$ | (Hasan & Gani, 2023; Hiremath et al., 2025; Schubert et al., 2026; Chen et al., 2025a; Tench & Demeester, 2026; Peng et al., 2026) |
| Non-adjacency | $X_i \not\sim X_j$ | (Darvariu et al., 2024; Verma et al., 2026; da Silva et al., 2026; Schubert et al., 2026; Bonilla et al., 2026) |
| **Ancestral causal claims: ancestry and ordering** | | |
| Ancestral relation | $X_i \in \mathrm{An}(X_j)$ | (da Silva et al., 2026) |
| Topological ordering | $X_i \prec X_j$ | (Vashishtha et al., 2025; Tench & Demeester, 2026) |
| **Latent-confounding causal claims: ancestral graphs** | | |
| Bidirected MAG adjacency | $X_i \leftrightarrow X_j$ | (Verma et al., 2026; da Silva et al., 2026) |
| Endpoint relation | Specified tails or arrowheads, unspecified endpoints, or absence | (Verma et al., 2026) |

## A.2 CLAIM-FUSE-FS

Algorithm 2 provides an update method for the finite-support case. Starting with candidate graphs and their data-only probabilities, the algorithm evaluates the provided claims for each graph and multiplies the claim weight by the likelihood determined by Equation 3.2. The algorithm employs log-weights to ensure numerical stability.

**Algorithm 2** CLAIM-FUSE-FS: passive fusion on a finite candidate set

**Input:** Candidate graphs $\{G_j\}_{j=1}^{J}$, normalized data-only masses $q_0(G_j)$, causal claims $\{(f_k, y_k, c_k)\}_{k=1}^{K}$ with $0 < c_k < 1$
$a_j \leftarrow \log q_0(G_j)$ for all $j$, with $\log 0 = -\infty$
**for** $k = 1, \ldots, K$ **do**
    **for** $j = 1, \ldots, J$ **do**
        $\ell_{kj} \leftarrow c_k$ if $f_k(G_j) = y_k$, otherwise $\ell_{kj} \leftarrow 1 - c_k$
        $a_j \leftarrow a_j + \log \ell_{kj}$
    **end for**
**end for**
$a_{\max} \leftarrow \max_j a_j$
$\widetilde{w}_j \leftarrow \exp(a_j - a_{\max})$ and $w_j \leftarrow \widetilde{w}_j / \sum_{i=1}^{J} \widetilde{w}_i$ for all $j$
**return** Weighted graphs $\{(G_j, w_j)\}_{j=1}^{J}$ representing $q_K$

## A.3 CLAIM-FUSE-SMC-PASSIVE AND CLAIM-FUSE-SMC-ACTIVE

CLAIM-FUSE-SMC-PASSIVE incorporates predefined claims, whereas CLAIM-FUSE-SMC-ACTIVE draws claims during inference. Both use resampling to control weight concentration and perform MH moves guided by Equation 3.4 to explore new DAGs.

**CLAIM-FUSE-SMC-PASSIVE.** Algorithm 3 receives a fixed set of $K$ claims with joint response likelihood $L(G) = \prod_{k=1}^{K} \ell_k(G)$ from Equation 3.2. It increases the tempering parameter $\lambda$ from zero to one to incorporate the supplied claims gradually. At each stage, incremental likelihood factors update the weights, and MH acceptance uses the current tempered target and the forward and reverse proposal probabilities.

**CLAIM-FUSE-SMC-ACTIVE.** Algorithm 1 selects the next question based on expected information gain. After $t$ queries, the target distribution is $\pi_t(G) \propto \exp\{S_D(G)\} \prod_{k=1}^{t} L(Y_t \mid G)$, where $k$ index queries. At each update, the current particle weights are multiplied by the likelihood of the latest answer, $w_j \leftarrow w_j \ell_t(G_j)$, and then normalized. MH moves are performed with respect to this updated distribution. Abstention consumes a query opportunity but does not change the distribution.

**Algorithm 3** CLAIM-FUSE-SMC-PASSIVE with likelihood tempering

**Require** Data score $S_D$, joint likelihood $L$ of supplied claims with $0 < c_k < 1$, particle count $J$, ESS fraction $\tau$, MH steps $M$, proposal mixture $Q_\gamma$

Initialize DAG particles $\{(G_j, w_j)\}_{j=1}^{J}$ approximating the data-only distribution
Normalize weights; set $\lambda \leftarrow 0$
**while** $\lambda < 1$ **do**
  Choose $\lambda' \in (\lambda, 1]$ from a fixed schedule or by predicted ESS; end the schedule at 1
  $\widetilde{w}_j \leftarrow w_j L(G_j)^{\lambda' - \lambda}$ for all $j$; normalize to obtain $w_j$
  $R \leftarrow \mathbb{I}\{(\sum_j w_j^2)^{-1} < \tau J\}$
  **if** $R = 1$ **then**
    Resample $J$ particles with probabilities $w_j$; set $w_j \leftarrow 1/J$
  **end if**
  **if** $R = 1$ or $\lambda' = 1$ **then**
    **for** each particle $G_j$ and each of $M$ MH steps **do**
      Propose $G' \sim Q_\gamma(\cdot \mid G_j)$
      Reject if $G'$ is inadmissible or $Q_\gamma(G_j \mid G') = 0$
      Otherwise set $G_j \leftarrow G'$ with probability $\min\Big\{1, \frac{\pi_{\lambda'}(G') Q_\gamma(G_j \mid G')}{\pi_{\lambda'}(G_j) Q_\gamma(G' \mid G_j)}\Big\}$
    **end for**
  **end if**
  $\lambda \leftarrow \lambda'$
**end while**
**return** Weighted DAGs, summing weights of duplicate graphs

# B THEORETICAL ANALYSIS

## B.1 PROOF OF PROPOSITION 1

*Proof.* Using the symmetric response model defined by Equation 3.2, Assumption 1 gives

$$\delta \le \ell_k^{(n)}(G) \le 1 - \delta, \qquad G \in \mathcal{G}, \quad k = 1, \dots, B. \tag{B.1}$$

Multiplying these bounds over the $B$ claims gives

$$\delta^B \le \prod_{k=1}^{B} \ell_k^{(n)}(G) \le (1-\delta)^B, \qquad G \in \mathcal{G}. \tag{B.2}$$

Applying these bounds to the numerator and denominator of the normalized update yields

$$1 - q_B^{(n)}(\mathcal{S}) = \frac{\sum_{G \notin \mathcal{S}} q_0^{(n)}(G) \prod_{k=1}^{B} \ell_k^{(n)}(G)}{\sum_{G \in \mathcal{G}} q_0^{(n)}(G) \prod_{k=1}^{B} \ell_k^{(n)}(G)} \le \frac{(1-\delta)^B \sum_{G \notin \mathcal{S}} q_0^{(n)}(G)}{\delta^B \sum_{G \in \mathcal{G}} q_0^{(n)}(G)} = \left(\frac{1-\delta}{\delta}\right)^B \big[1 - q_0^{(n)}(\mathcal{S})\big]. \tag{B.3}$$

For a fixed $k$ and $n$, abbreviate $q = q_k^{(n)}$ and let $P$ denote the MH transition kernel defined in equation 3.5. The Hastings acceptance rule gives detailed balance for $G \neq G'$:

$$\begin{aligned} q(G)P(G, G') &= \min\big\{q(G)Q_\gamma(G' \mid G),\, q(G')Q_\gamma(G \mid G')\big\} \\ &= q(G')P(G', G). \end{aligned} \tag{B.4}$$

Summing detailed balance over the previous graph and using that transition probabilities sum to one establishes invariance

$$\sum_{G\in\mathcal{G}} q(G)P(G,G') = q(G')\sum_{G\in\mathcal{G}} P(G',G) = q(G'). \tag{B.5}$$

Repeated application therefore preserves the current exact target for any fixed number $m$ of MH moves:

$$qP^m = q, \qquad m = 0,1,2,\dots. \tag{B.6}$$

Starting from $q_0^{(n)}$, each exact claim update produces $q_k^{(n)}$ and its following MH block preserves that distribution, so the final law remains $q_B^{(n)}$.

Since $B$ and $\delta$ are fixed, the same bound proves consistency with or without these MH blocks:

$$0 \le 1 - q_B^{(n)}(\mathcal{S}) \le \left(\frac{1-\delta}{\delta}\right)^B \left[1 - q_0^{(n)}(\mathcal{S})\right] \longrightarrow 0. \tag{B.7}$$

□

**Remark.** Proposition 1 analyzes the exact target distribution. In practice, CLAIM-FUSE represents this distribution with a finite weighted particle approximation, which introduces additional Monte Carlo approximation error that vanishes as the particle approximation becomes exact.

### B.2 PROOF OF PROPOSITION 2

*Proof.* Averaging the response model with confidence $c$ over the current graph distribution gives

$$\mathbb{P}_{\text{model}}(Y_k = 1) = cp_k + (1-c)(1-p_k) = \tfrac{1}{2} + (2c-1)(p_k - \tfrac{1}{2}). \tag{B.8}$$

Writing $h(p) = -p\log p - (1-p)\log(1-p)$ for binary entropy. With common confidence $c$, $H(Y_k \mid G) = h(c)$, computing all entropies under the working response model gives

$$\mathrm{EIG}(k) = H(Y_k) - H(Y_k \mid G) = h\big(\tfrac{1}{2} + (2c-1)(p_k - \tfrac{1}{2})\big) - h(c). \tag{B.9}$$

Since every homogeneous query also belongs to the heterogeneous pool,

$$\max_{k\in\mathcal{Q}_{\text{het}}} \mathrm{EIG}(k) \ge \max_{k\in\mathcal{Q}_{\text{homo}}} \mathrm{EIG}(k). \tag{B.10}$$

For $c \neq 1/2$, the distance of the modeled answer probability from $1/2$ satisfies

$$\left|\mathbb{P}_{\text{model}}(Y_k = 1) - \tfrac{1}{2}\right| = |2c-1|\left|p_k - \tfrac{1}{2}\right|, \qquad |2c-1| > 0. \tag{B.11}$$

Since binary entropy strictly decreases with this distance, the more uncertain query has strictly greater modelled information gain:

$$|p_k - \tfrac{1}{2}| < |p_j - \tfrac{1}{2}| \quad\Longrightarrow\quad \mathrm{EIG}(k) > \mathrm{EIG}(j). \tag{B.12}$$

Applying the balance condition to every homogeneous query therefore proves strict improvement:

$$\max_{j\in\mathcal{Q}_{\text{het}}} \mathrm{EIG}(j) \ge \mathrm{EIG}(k) > \max_{j\in\mathcal{Q}_{\text{homo}}} \mathrm{EIG}(j). \tag{B.13}$$

□

## C EVALUATION METRICS

### C.1 GRAPH DISCOVERY METRICS

Let $\mathcal{V} = \{1,\dots,p\}$ be the node set. For a DAG $G$, define the pairwise edge state $s_G(i,j) \in \{\varnothing, i\to j, j\to i\}$ for $i<j$. For a MAG $G$, let $m_G(i \mid j) \in \{\varnothing, \text{tail}, \text{arrow}\}$ denote the endpoint mark at

node $i$ on the pair $\{i, j\}$, with $\varnothing$ indicating non-adjacency. We use structural error counts

$$d_{\mathrm{DAG}}(G, G') = \sum_{i<j} \mathbf{1}\{s_G(i,j) \neq s_{G'}(i,j)\}\,, \tag{C.1}$$

$$d_{\mathrm{MAG}}(G, G') = \sum_{i<j} [\mathbf{1}\{m_G(i \mid j) \neq m_{G'}(i \mid j)\} + \mathbb{I}\{m_G(j \mid i) \neq m_{G'}(j \mid i)\}]\,. \tag{C.2}$$

Thus $d_{\mathrm{DAG}} \in [0, \binom{p}{2}]$ and $d_{\mathrm{MAG}} \in [0, 2\binom{p}{2}]$. In the DAG regime, an added, deleted, or reversed edge contributes one pairwise error. Thus $d_{\mathrm{DAG}}$ is equivalent to structural Hamming distance (SHD) with an edge reversal being counted as one error. In the MAG regime, errors are counted at individual endpoints, so a missing or additional edge contributes two errors.

For graph distribution $q_a$ and target $G_a^\star$, expected SHD (ED) is defined as

$$\mathrm{ED}(q_a, G_a^\star) = \mathbb{E}_{G \sim q_a}[d_a(G, G_a^\star)]\,, \tag{C.3}$$

SHD measures mean structural error. Adjacency and orientation F1 are evaluated on the maximum-mass graph, with endpoint-mark F1 replacing orientation F1 under latent confounding.

Let $H$ be a partially oriented DAG, e.g. CPDAG. Let $m_H(i,j) \in \{\varnothing, \mathrm{tail}, \mathrm{arrow}\}$ be the mark at node $i$ on pair $\{i,j\}$ in $H$, and represent the target DAG $G^\star$ using the same marks. Define the end error distance:

$$d_{\mathrm{end}}(H, G^\star) = \frac{1}{2} \sum_{i<j} [\mathbf{1}\{m_H(i,j) \neq m_{G^\star}(i,j)\} + \mathbf{1}\{m_H(j,i) \neq m_{G^\star}(j,i)\}]\,. \tag{C.4}$$

For two DAGs this equals structural Hamming distance. For partially oriented DAGs, an undirected edge with the correct adjacency but unresolved direction contributes $1/2$, whereas a missing or reversed edge contributes one. Let $A(H)$ be the set of adjacent pairs in $H$ and $D(H)$ contain ordered edges in $H$. Then

$$\mathrm{F1}_{\mathrm{adj}}(H, G^\star) = \frac{2|A(H) \cap A(G^\star)|}{|A(H)| + |A(G^\star)|}, \qquad \mathrm{F1}_{\mathrm{orient}}(H, G^\star) = \frac{2|D(H) \cap D(G^\star)|}{|D(H)| + |D(G^\star)|}. \tag{C.5}$$

An undirected edge therefore counts as an adjacency prediction but missing orientation. For CLAIMFUSE, Equations C.3 uses estimated graph distribution represented by weighted graph particles, while F1 scores use the maximum-mass graph .

### C.2 CONFIDENCE AND QUERY-COST METRICS

For the $M$ answers, let $V \subseteq \{1, \ldots, M\}$ be the set of valid responses. For response $i$, $c_i \in \{0, 1\}$ indicates whether it matches the ground truth graph or not. Let $p_i \in [1/2, 1]$ be the estimated confidence. We use the following metrics: the number of valid responses $|V|$, coverage $|V|/M$, precision $|V|^{-1} \sum_{i \in V} c_i$, Brier score $|V|^{-1} \sum_{i \in V} (p_i - c_i)^2$, and the expected calibration error with ten bins $\sum_b |I_b||V|^{-1} |\mathrm{acc}(I_b) - \mathrm{conf}(I_b)|$. Invalid responses consume the query budget but do not update the distribution.

Token cost is defined as the sum of all recorded input and output tokens, including invalid responses, repairs using LLM, and semantically consistent prompt variants. Runtime measures the actual elapsed time (wall-clock time) from start to finish.

## D SYNTHETIC EXPERIMENTS

For each seed, we sample a directed acyclic graph (DAG) $G^*$ over $p$ variables in the following way: first, we draw a random topological order, and then we include each admissible edge independently with probability $2/(p-1)$. This yields an expected degree of two. The data follows the linear Gaussian structural model,

$$X_j = \sum_{i \in \mathrm{pa}_{G^\star}(j)} \beta_{ij} X_i + \varepsilon_j, \qquad \beta_{ij} \sim \mathrm{Unif}([-2, -0.5] \cup [0.5, 2])\,, \tag{D.1}$$

where coefficients are sampled independently and the noise variables $\varepsilon_j \sim \mathcal{N}(0,1)$ are mutually independent. We generate 50 independently seeded $p = 20$ SCMs using master seed 20260810 and replicate indices 0 through 49. A realization of size 1000 is generated for each SCM. Its first observations form the nested smaller sample condition. The two datasets are standardized separately. This keeps the graph, coefficients, and data fixed across different sample sizes.

### D.1 Shared Pairwise Expert Knowledge

In this experiment, we compare data-only and LLM-assisted causal discovery methods with the same observational data and the same set of causal claims.

**Dataset generation.** We use ten different SCM types. For each SCM, we sample twenty variable pairs uniformly without replacement from the $\binom{20}{2} = 190$ possibilities. An oracle expert classifies each of the pairs here as $A \to B$, $B \to A$ or no direct edge in either direction. No direct edge doesn't rule out an indirect causal path.

**How causal claims are encoded.** Claim-Fuse-FS and Claim-Fuse-SMC-Passive encode a directed answer as a positive directed-edge claim and an absent edge as a negative adjacency claim. Their internal soft confidence is capped at 0.99, so agreement and contradiction receive likelihood factors 0.99 and 0.01. CaMML receives local pair marginals $(p_{A\to B}, p_{B\to A}) = (1,0), (0,1)$, or $(0,0)$, corresponding to probabilities $(1,0,0,0)$, $(0,1,0,0)$, or $(0,0,1/2,1/2)$ over its two directed and two absent-with-order states. ECA represents a directed answer by one required edge and its forbidden reverse, and represents absence by two forbidden edges; thus its twenty pairwise facts produce forty signed entries. ECA also caps its native assertion confidence at 0.99. Marg receives twenty categorical opinions attributed to one shared expert. Unqueried pairs retain each method's default treatment.

**Setup for data-only causal discovery method** For our data-only causal discovery methods, Stable Fisher–$Z$ PC uses $\alpha = 0.05$, Gaussian-BIC GES uses penalty multiplier 0.5, and linear NOTEARS uses $\ell_1 = 0.05$ with edge threshold 0.30. Each method is fit once to the same standardised observations without bootstrap resampling, repeated fitting, retries or target-dependent tuning.

PC and GES's outputs are partially directed graphs while NOTEARS returns a point DAG. For initializing Claim-Fuse-FS and Claim-Fuse-SMC-Passive, the partially directed acyclic graph from PC contributes at most 200 certified distinct compatible DAGs with maximum in-degree 12.

**Evaluation.** The ED columns in Table 5 report the expected structural error, equivalently the expected SHD, as defined in Appendix C. For Claim-Fuse-FS, Claim-Fuse-SMC-Passive, and CaMML, ED is computed over the complete returned graph distribution, while adjacency and orientation F1 are computed using the canonical maximum-mass graph. Marg, ECA, and NOTEARS return point DAGs, whereas PC and GES are evaluated as raw, partially directed graphs.

Table 5. Graph recovery on ten 20-node SCMs, mean (SE). All methods share observations; parentheses give the number of shared pairwise facts. Bold marks the best mean. Claim-Fuse-FS and Claim-Fuse-SMC-Passive use the shared PC initialization described above.

| | $n = 200$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|
| Method | ED ↓ | Adj. F1 ↑ | Orient. F1 ↑ | ED ↓ | Adj. F1 ↑ | Orient. F1 ↑ |
| PC (0) | 7.40 (1.05) | **0.897 (0.025)** | 0.567 (0.051) | 5.55 (0.82) | **0.937 (0.021)** | 0.671 (0.029) |
| GES (0) | 7.30 (0.64) | 0.880 (0.014) | 0.686 (0.048) | 5.30 (0.91) | 0.925 (0.021) | 0.736 (0.049) |
| NOTEARS (0) | 16.30 (1.38) | 0.767 (0.025) | 0.315 (0.030) | 16.80 (1.25) | 0.761 (0.026) | 0.284 (0.023) |
| CaMML (20) | 23.82 (0.99) | 0.628 (0.023) | 0.480 (0.028) | 14.04 (0.89) | 0.826 (0.023) | 0.695 (0.044) |
| Marg (20) | 11.50 (1.01) | 0.821 (0.021) | 0.568 (0.040) | 8.50 (0.89) | 0.891 (0.017) | 0.630 (0.040) |
| ECA (20) | 15.30 (1.48) | 0.792 (0.027) | 0.357 (0.035) | 15.50 (1.23) | 0.779 (0.024) | 0.360 (0.026) |
| Claim-Fuse-FS (20) | **6.91 (0.98)** | **0.897 (0.025)** | 0.682 (0.039) | **5.07 (0.76)** | **0.937 (0.021)** | 0.769 (0.032) |
| CLAIM-FUSE-SMC-PASSIVE (20) | 7.07 (1.00) | **0.897 (0.025)** | **0.701 (0.038)** | 5.16 (0.76) | **0.937 (0.021)** | **0.795 (0.024)** |

### D.2 CLAIM FAMILIES WITH UNCERTAIN EXPERTS

In this section, we investigate the performance of local, ancestral vs. all causal claims when expert answers are imperfect but calibrated.

**Dataset construction** This experiment is solely on synthetic data. As in our other experiments, we use sample sizes of n = 200 and n = 1000. We create a "simulated experts" that returns the correct binary answer with probabilities $a \in \{0.6, 0.8, 1.0\}$ or the opposite answer otherwise. This $a$ is the confidence score, which is capped at $0.99$.

**Results.** Our results show that when the expert is moderately reliable (accuracy 0.8), letting CLAIM-FUSE ask any type of claim works better than asking only local or only ancestral claims. After 50 queries, the union of local and ancestral claims (all claims) lowers ED compared with local claims by 0.67 at n = 200 and by 0.80 at n = 1000. The advantage needs a few queries to appear: at 5 queries, the union is slightly worse than the best single claim family. At n = 200, local and ancestral claims stop improving after 20 queries, while the union keeps improving. When the expert is very noisy (0.6) or perfect (1.0), the choice of claim family makes no clear difference.

### D.3 MISCALIBRATION OF CAUSAL CLAIMS

In this section, we check for miscalibration, e.g., when the claim source says "I'm 70% correct" but is actually correct for a different number. We set the confidence fixed at $0.70$ but change the response accuracy.

**Setup** We use a set of fifty queries for our experiments. We compare local, ancestral and all causal claims. Each run has a budget of 50 EIG-selected queries. We compare local claims (directed edges and adjacencies), ancestral claims (ancestry and order), and their union. All runs use the same ten SCMs, two sample sizes, five paired seeds, and a shared PC+GES initialization. Responses have confidence $0.70$ and actual accuracy $0.4$, $0.7$, or $1.0$. We record results after 5, 10, 20, and 50 queries.

**Robustness experiments with 10 queries** Figure 4 reports graph recovery after ten queries on the same ten SCMs, sample sizes, and PC+GES initialization. Queries are selected by EIG from the union of local and ancestral claims. Every response has confidence $0.70$, while its actual accuracy is $0.4$, $0.7$, or $1.0$, giving overconfident, calibrated, and underconfident sources. The three regimes share one uniform draw per query and use their accuracy as a threshold, so a response that is correct at a lower accuracy is also correct at a higher one whenever the same query is selected. We report ED, energy score (ES), adjacency F1, and orientation F1. Five paired seeds are averaged within each SCM, and error bars show one standard error across SCMs. Intervals in the text are paired percentile intervals from $10{,}000$ SCM-level bootstrap resamples. Because confidence is fixed, accuracy and calibration change together, so these comparisons describe each source regime as a whole rather than the separate effects

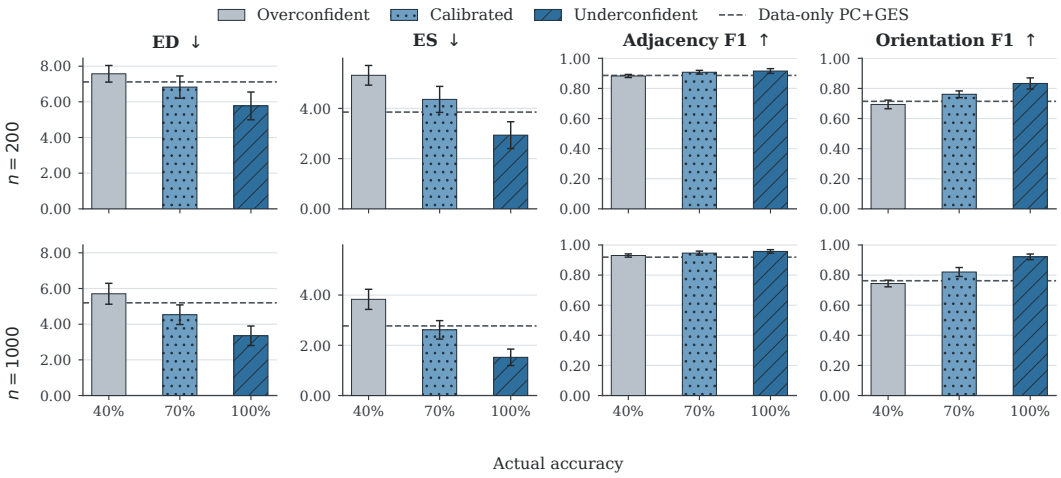


Figure 4. Recovery after ten EIG queries at fixed confidence 0.70. Bars show means with one SE across ten SCMs after averaging five repetitions. Dashed lines denote the PC+GES initialization.

**Findings.** Recovery improves as source accuracy increases. Calibrated responses outperform overconfident ones on all four metrics, with paired intervals excluding zero. They also improve on data-only PC+GES, reducing ED from 7.12 to 6.83 at $n = 200$ and from 5.20 to 4.53 at $n = 1000$. Fully accurate but underconfident responses give the best results, so reliable claims help even when weighted cautiously. In contrast, overconfident responses from a worse-than-chance source make recovery worse than data alone, raising ES from 3.85 to 5.32 at $n = 200$.

## E Neuropathic Pain Dataset Experiments

Due to scalability issue of some evaluated baselines, we use a 20-node sub-network of the full 222-node neuropathic pain dataset.This appendix gives information about this 20-node sub-network, the masking variable names to avoid data contamination, and an prompt example of CLAIM-FUSE of actively eliciting causal knowledge from LLM.

### E.1 20-node sub-network

**20-node sub-network.** The original neuropathic-pain simulator uses a fixed 222-node binary DAG, constructed using medical domain knowledge. With depth of 3, it contains pathophysiology, radiculopathy, and symptom layers (Tu et al., 2019). We extract a `L Nackbesvär` and all its ancestors, which creates a sub-network with 20 nodes and 30 edges, as shown in Figure 5. It contains six disc-level nodes, `Kraniocervikal ledskada`, left and right C2 through C7 radiculopathy nodes, and the target symptom.

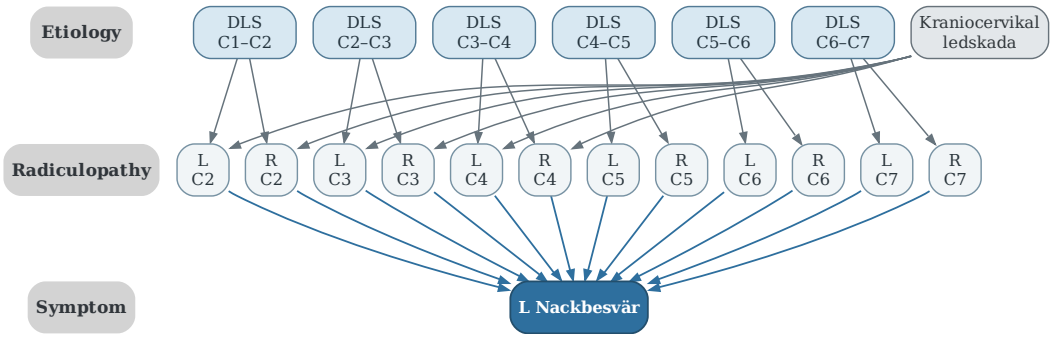


Figure 5. Neuropathic pain sub-network with 20 variables and 30 directed edges.

**Observation generation.** For experiments on the 20-node sub-network, we generate ten observations of sample size 1000, then use its first 200 data-points for the sample size 200 condition. For each data-seed, we run the code from the original paper, which uses ancestral sampling to draw the samples.

**Masked variable names.** Since the paper and data may already present in the pre-training corpus of current LLMs, there could be data contamination issue that LLM does not give causal claims based on its domain knowledge but by memorizing the dataset. We address this by replacing variable names using some non-informative names and providing meaning-preserving description to each of them. The variable are masked by `V01` through `V20`. The accompanying description could be like "disc lesion or disorder at cervical level C2–C3," without revealing graph membership or direction.

**Data-only causal discovery method setup.** PC likelihood-ratio $G^2$ test at $\alpha = 0.01$ as its conditional independence test. GES uses BDeu with equivalent sample size one and structure prior one.

### E.2 CLAIM-FUSE CAUSAL QUERY INTERFACE

In this subsection we give an example of the prompt we use to elicit causal claim from LLMs for the neuropathic pain dataset and how we compute confidence.

**Example prompt.** Each causal query provides LLM with general domain context and description of the two masked variables, and ends with the question on the target causal relation and instruction to answer. The prompt does not supply observations, graph structure, or the dataset name.

> **System:** You are a careful medical causal expert. Do not infer the answer from variable ordering. Answer the causal question and state the probability that your selected yes/no answer is correct. Abstain when you cannot make a defensible judgment.
>
> **User:** Context: a clinical diagnostic system. Glossary: `V03` = disc lesion or disorder at cervical level C2–C3; `V13` = left C3 radiculopathy. Assess whether `V03` directly causes `V13` in the stated context. Association, shared symptoms, a common cause, or a pathway through an intermediate variable does not by itself establish a direct edge. Answer yes only when a specific direct mechanism is supported, answer no only when absence of the edge is supported, and otherwise abstain. Return JSON with `answer`, `confidence`, and a one-sentence `justification`.

**Stated and consistency confidence.** Stated confidence simply send the question once and let LLM answer yes, no or unsure together with the self-reported probability of the given answer. Semantic consistency instead reports the confidence as LLM's consistency over perturbed by semantically consistent queries. Specifically, we perturb the prompt using 4 paraphrasing and yes/no order swap for each of them, resulting 8 prompts. The final response use the majority of the answers when 6 of 8 are yes or no. The confidence is then computed as the proportion of the majority answer. For example, LLM answers "yes" six times, "no" and "unsure" once for each. This is then treated as valid answer giving "yes", with confidence $6/8$.

**Claim pools.** For the 20-node neuropathic-pain sub-network, the local pool contains 380 directed-edge and 190 adjacency questions. The ancestral pool contains 380 ordered ancestral questions. Their union contains 950 distinct questions.

## F ALARM DATASET EXPERIMENTS

This appendix gives an overview of the ALARM dataset and an example prompt of CLAIM-FUSE to elicit causal claim from LLM.

### F.1 OVERVIEW OF THE ALARM DATASET

The ALARM DAG has 37 categorical variables, 46 directed edges, and a longest directed path of ten edges. We draw ten samples of size 1000 using ancestral sampling and use first 200 data-points

for size 200 samples. Similar to the treatment on the neuropathic pain dataset, we mask the variable names and provide meaning preserving description instead. Prompts contain individual descriptions but no graph structure, dataset name, or reference answer.

### F.2 CLAIM-FUSE ACTIVE QUERYING INTERFACE

**Example prompt.** The following is an example prompt to elicit causal claims from LLM on the ALARM dataset:

> **System:** You are a careful medical causal expert. Do not infer from variable ordering. Answer the causal question and state the probability that your selected yes/no answer is correct. Abstain when you cannot make a defensible judgment.
>
> **User:** Context: physiology and monitoring during anesthesia and mechanical ventilation. Glossary: `V35` = whether anesthesia or pain relief is insufficient; `V36` = overall peripheral blood-vessel resistance, with ordered levels low, normal, and high. Assess whether `V35` directly causes `V36`. Association, shared symptoms, a common cause, or a pathway through an intermediate variable does not by itself establish a direct edge. Answer yes only when a specific direct mechanism is supported, answer no only when absence of the edge is supported, and otherwise abstain. Return JSON with `answer`, `confidence`, and a one-sentence `justification`.

Stated and consistency confidence are computed using the same method as in neuropathic pain dataset.

**Claim pools.** For the 37-node ALARM network, the local pool contains 1,332 directed-edge and 666 adjacency questions. The ancestral pool contains 1,332 ordered ancestral questions. Their union contains 3,330 distinct questions.

## G EXTENSION TO LATENT CONFOUNDING

In this appendix, we extend the supported claim types from local and ancestral to hidden confounding. We first define the causal claims under hidden confounding and how it is evaluated on maximal ancestral graphs (MAGs), then evaluate CLAIM-FUSE-SMC-PASSIVE on synthetic causal graphs with hidden confounding. Thus, $X \leftrightarrow Y$ states that neither endpoint is an ancestor of the other and can represent a hidden common cause. Different from DAG, a directed MAG edge $X \rightarrow Y$ means $X$ is an ancestor of $Y$ but not necessarily a direct cause.

### G.1 GRAPH AND CLAIMS UNDER HIDDEN CONFOUNDING

**Maximal ancestral graphs.** A maximal ancestral graph (MAG) represents causal relations and hidden confounding among observed variables. We consider latent confounding without selection bias, so MAGs contain directed and bidirected edges. An arrowhead at $X$ on an edge between $X$ and $Y$ means $X$ is not an ancestor of Y.

**Partial ancestral graphs.** A partial ancestral graph (PAG) represents an equivalence class of MAGs with the same conditional-independence relations (Ali et al., 2009; Zhang, 2008). It use tails, arrowheads and circles as edge marks.

**Claim evaluation.** In a MAG $G$, local directed relation and ancestral relation are represented by a directed edge, whereas a latent-confounding relation is a bidirected edge. These evaluations on MAGs produce the same binary values $f_k(G)$, used in Equation 3.2. So, the claim likelihood and normalized graph update can be directly transferred to this setting. The posterior then becomes a distribution over MAGs on the observed variables.

**Scope of the extension.** The implemented extension performs passive fusion over a fixed finite set of MAG candidates. The mutation and active-query of CLAIM-FUSE remain restricted to DAGs.

### G.2 Experiment Setup

**Question and data.** We test whether uncertain local, ancestral, and latent-confounding claims improve recovery with fixed candidate MAGs. We use 20 fixed linear-Gaussian SCM seeds with ten observed variables and two hidden common causes. Each SCM provides samples of 200 and 1,000 observations. We first generate the full graph and random hold common causes unobserved. The ground truth MAG is obtained by marginalizing DAG on the hidden variables.

**Candidate construction.** FCI is fitted once per dataset with Fisher's $Z$ test at significance level $0.05$. We retain at most 200 distinct MAG completions of its PAG.

**Matched evidence.** We compare the data-only distribution $q_0$ with passive CLAIM-FUSE-FS on exactly the same candidate graphs. Each SCM receives ten claims from each of the local, ancestral, and latent-confounding pools. Claims are fixed across sample sizes. Every answer has accuracy $0.8$ with supplied confidence also being $0.8$.

**Evaluation.** Expected distance is the endpoint-error count from Appendix C. Endpoint F1 and bidirected-adjacency F1 evaluate the highest-mass graph against the ground truth graph.

### G.3 Results

Table 6. Performance of CLAIM-FUSE-FS with latent confounding.

| $n$ | Metric | Data-only | CLAIM-FUSE-FS |
|---|---|---|---|
| 200 | Endpoint ED | 16.85 | 16.00 |
| 200 | Endpoint F1 | 0.4695 | 0.5611 |
| 200 | Bidirected F1 | 0.1696 | 0.2530 |
| 1000 | Endpoint ED | 14.00 | 13.07 |
| 1000 | Endpoint F1 | 0.5575 | 0.6289 |
| 1000 | Bidirected F1 | 0.2260 | 0.3441 |

**Recovery.** Table 6 shows lower expected endpoint error and higher endpoint and bidirected F1 after incorporating uncertain claims at both sample sizes. The experiment shows reweighing using causal claims improves the expected accuracy in MAGs, with claims about local, ancestral and hidden confounding.

**Candidate-support limitation.** The main limitation of finite support is that if the ground truth graph is not present in the candidate graphs, then it is impossible to recover it even with infinite number of correct causal claims. Assuming causal sufficiency, we fix this by using MH to propose new graphs that fix the disagreement between candidate graphs and causal claims. Future work can extend to hidden confounding regime by developing new statistical techniques and LLM elicitation method to account for hidden confounding.