CAUSAL REFEREE: HIERARCHICAL CAUSAL DISCOVERY WITH LLM-GUIDED EXPERT REWEIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Causal discovery aims to uncover causal structures from observational data, which is crucial for real-world decision-making. However, different causal discovery algorithms can produce divergent results that conflict with each other, complicating the identification of accurate causal graphs. Traditional approaches rely on numerical values and statistical assumptions, often ignoring rich domain-specific information, such as feature descriptions, which could also help structure learning. While recent works explore using Large Language Models (LLMs) to infer causal relations via direct queries, such methods can be unreliable due to a lack of alignment with the actual data. To address these limitations, we propose **Ca**usal **Re**feree (**CaRe**), a novel framework that aggregates structural insights from statistical discovery experts across different graph levels via linear opinion pooling, and uses an LLM as a meta-referee to dynamically reweight experts on disputed relations, thereby composing an improved and more complete causal graph. Extensive experiments on both synthetic and real-world datasets demonstrate that CaRe achieves the best average performance across a wide range of causal discovery methods, highlighting the effectiveness of using LLMs for meta-analysis in causal discovery.

1 INTRODUCTION

Causal discovery aims to identify the underlying cause-and-effect structure among variables from observational data (Pearl, 2009; Glymour et al., 2019), providing an essential foundation for decision-making in many real-world domains such as healthcare, economics, transportation, and epidemiology. Over the past decades, Statistical Causal Discovery (SCD) has made substantial progress, ranging from constraint-based (Spirtes et al., 2000) and score-based approaches (Chickering, 2002; Zheng et al., 2018) to functional causal models (Shimizu et al., 2006; Hoyer et al., 2008; Zhang & Hyvärinen, 2009), enabling scalable and theoretically grounded structure learning under well-defined assumptions. More recently, Large Language Models (LLMs), with strong natural language understanding and reasoning abilities (Zhao et al., 2025), have emerged as a promising complement for causal discovery (Long et al., 2022; Kıcıman et al., 2024). By integrating information from vast amounts of scientific text and domain knowledge, LLMs can interpret variable semantics and feature descriptions, reason about plausible mechanisms, and propose causal relations that may be overlooked by statistical, data-only discovery methods (Abdulaal et al., 2024).

Challenges. Despite these advances, accurately recovering causal graphs in real-world settings remains challenging (Stein et al., 2025). While SCD is mathematically rigorous and empirically stable under its assumptions, different SCD algorithms may return conflicting graphs on the same dataset due to distinct inductive biases and objective functions, and their performance can be highly scenario-dependent (i.e., a method that excels on one data domain may fail in another, Figure 1 Right), making it difficult to select a single "best" solver for an unseen dataset (Zhang et al., 2023; Vo et al., 2026). On the other hand, LLM-based Causal Discovery (LCD) often requires querying an LLM for causal relations over all possible pairs of variables, which is time-consuming and difficult to scale (Jiralerspong et al., 2024). Besides, although LLMs are capable of exploiting semantic cues and external knowledge, their predictions often lack alignment with the actual data and are therefore prone to unreliability and hallucination problems (Ji et al., 2023). This issue is further exacerbated when variable names are vague or feature descriptions are incomplete or misleading, in which case the model may produce incorrect causal assertions (Figure 1 Left).

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.

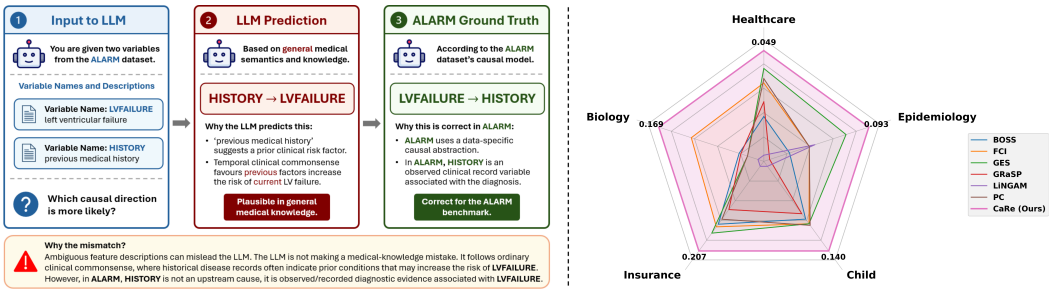

Figure 1: *Left:* Hallucination problem in direct LLM-based causal querying with misleading feature descriptions. *Right:* Causal discovery performance across different data domains. SCD algorithms can be domain-dependent, with performance degrading markedly when moving to a new domain, while our **CaRe** remains effective across all domains and consistently achieves the best results.

Motivations. Considering LLMs' strength in analyzing and coordinating myriad information, we seek a different but more reliable role for LLMs in causal discovery: ***instead of treating an LLM as a "causal expert" that directly infers causal relations, we can use it as a meta-referee that adjudicates among competing, data-driven hypotheses***. Concretely, running several SCD methods is often easy, but their disagreements may encode valuable information. Consistent causal relations across methods tend to be reliable, whereas conflicting ones highlight uncertainty that stems from data limitations and method-specific biases. In particular, expert reliability can vary across graph levels and individual relations, which uniform or globally fixed weights cannot fully capture. An LLM is particularly well-suited to interpret and judge these situations because it can jointly reason over heterogeneous method assumptions, failure modes, variable semantics, and dataset profiles, thereby deciding which methods should be trusted more for which parts of the graph.

In light of the aforementioned problems and motivations, we explore the potential of using LLMs for *meta-analysis* in causal discovery: leveraging an LLM to *selectively* coordinate and calibrate multiple statistical causal discovery experts. In this paper, we propose **Ca**usal **Re**feree (**CaRe**), a hierarchical ensemble framework that performs relation-level linear opinion pooling over heterogeneous SCD experts, and invokes an LLM as a meta-referee to dynamically reweight experts *only* when the aggregated confidence is close to the decision boundary. Unlike existing approaches that either aggregate causal graphs with numerically computed weights or query LLMs to directly infer causal relations, we use an LLM to estimate expert credibility by jointly considering variable semantics, dataset characteristics, and expert-data compatibility. In this way, CaRe mitigates benchmark-leakage concerns by avoiding direct edge prediction and restricting the LLM's role to expert reweighting, while keeping the overall procedure grounded in data and leaving expert decisions unchanged. Experimentally, CaRe is compatible with a diverse set of mainstream causal discovery methods and achieves the best average performance across eight synthetic and real-world benchmarks, demonstrating superior robustness while requiring substantially fewer LLM queries than exhaustive LLM-based causal discovery. The contributions of our work are threefold:

- We introduce a new role for LLMs in causal discovery: *meta-referees* that adjudicate among competing, data-driven hypotheses from multiple SCD experts, enabling reliable integration of domain knowledge without sacrificing data alignment.
- We propose CaRe, a hierarchical causal discovery ensemble framework that aggregates expert beliefs in a coarse-to-fine manner over skeletons, v-structures, and edge orientations, with margin-aware LLM-guided expert reweighting for disputed relations.
- Extensive experiments on both synthetic and real-world datasets show that CaRe outperforms competitive baselines, highlighting remarkable improvements in causal discovery accuracy and cross-domain robustness with substantially fewer LLM calls.

2 Related Work

Beyond statistical structure learning from observational data, causal discovery has also been approached from two alternative perspectives: ensemble-based methods and LLM-based methods.

2.1 Ensemble-based Causal Discovery

Ensemble-based methods aim to improve the robustness of causal discovery by aggregating multiple candidate graphs into a more reliable consensus structure. Early studies mainly follow the bootstrap aggregation paradigm, where the same causal discovery algorithm is repeatedly applied to resampled datasets, and the resulting graphs are then aggregated to reduce estimation variance. Friedman et al. (2013) propose an analysis framework for Bayesian networks, where edge-selection frequencies are used to measure structural stability. DAGBag (Wang & Peng, 2014) learns an ensemble of Directed Acyclic Graphs (DAGs), aiming to find an average graph that minimizes the distance to all graphs. Recent studies further extend causal discovery ensembles to heterogeneous graph aggregation, where outputs from multiple algorithms are aggregated. The work from Guo et al. (2021) introduces a scalable and flexible two-phase ensemble framework that combines data partitioning with ensemble learning to mitigate the divergence across algorithms. Aslani & Mohebbi (2023) formulate heterogeneous DAG aggregation from an optimization perspective, using a two-stage pipeline and lazy constraints to obtain a consensus DAG. WECD (Zhang et al., 2023) utilizes weighted ensemble causal discovery for effective connectivity estimation while improving robustness in noisy and small-sample settings. Despite their effectiveness, existing ensemble-based causal discovery methods usually determine expert weights through numerical criteria or manually designed weighting schemes. In contrast, our CaRe framework uses an LLM as a meta-referee to estimate relation-specific expert credibility from rich natural language evidence, making the ensemble process more context-aware.

2.2 LLM-based Causal Discovery

Large Language Models (LLMs) have recently been explored as a new source of causal knowledge, owing to their ability to interpret variable semantics and leverage domain knowledge acquired from large-scale scientific corpora. Existing LLM-based causal discovery methods can be broadly categorized into two lines. One is using *LLMs for direct inference* that queries LLMs to infer causal relations from variable names, feature descriptions, or retrieved textual evidence. Willig et al. (2022) examine whether foundation models can reason about causality, while Long et al. (2022) evaluate whether LLMs can build causal graphs by scoring candidate causal relations. Kıcıman et al. (2024) further study LLMs on a range of causal reasoning tasks, including pairwise causal discovery and counterfactual reasoning. To improve scalability, Jiralerspong et al. (2024) reduce the number of LLM queries through a breadth-first search. To improve grounding, Zhang et al. (2024) introduce a retrieval-augmented generation framework that incorporates scientific evidence for structure discovery. The other is using *LLMs for posterior correction*, which uses LLMs to refine, validate, or correct causal structures produced by SCD methods. Long et al. (2023) treat language models as imperfect experts for orienting causal relations beyond Markov equivalence classes. Takayama et al. (2025) integrate LLM-derived background knowledge into SCD through prior knowledge augmentation. CMA (Abdulaal et al., 2024) synergizes metadata-based LLM reasoning with data-driven structural causal modelling. Both lines still rely on the LLM to directly validate, orient, add, remove, or correct causal edges, which may introduce unreliable judgments into the final graph. In contrast, our CaRe does not treat the LLM as a causal expert that directly predicts or modifies causal relations. Instead, by restricting the LLM's role to estimating expert weights, CaRe keeps the final graph grounded in data-driven SCD outputs while still leveraging domain knowledge for context-aware calibration.

3 Causal Referee (CaRe)

Figure 2 illustrates the overall framework of our Causal Referee (CaRe), which constructs the causal graph in a coarse-to-fine manner through three hierarchical layers that aggregate (a) graph skeletons, (b) v-structures, and (c) edge orientations across multiple SCD experts. At each layer, expert beliefs are pooled via *linear opinion pooling* of bootstrap-based support scores, and an LLM is invoked as a meta-referee on disputed relations to estimate *relation-specific, context-dependent expert weights*. The resulting relations are integrated across layers to produce the final graph $\widehat{G}$.

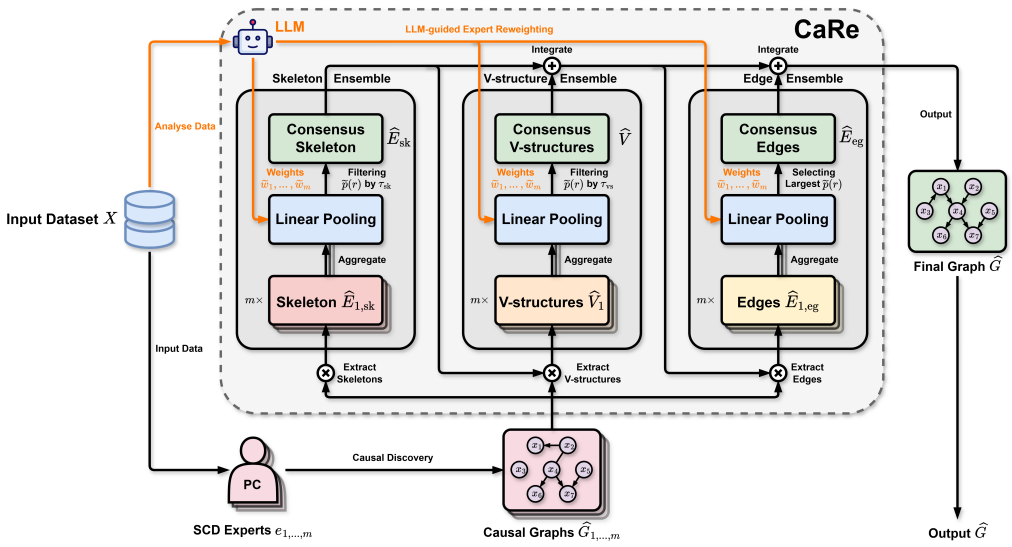

Figure 2: The overview of CaRe. Given input data, multiple SCD experts first produce candidate causal graphs. CaRe then aggregates expert beliefs through linear opinion pooling and constructs the final graph in a hierarchical, coarse-to-fine manner. For causal relations whose aggregated confidence lies near the decision boundary, an LLM is invoked as a meta-referee to perform context-aware expert reweighting, resolving disagreements while keeping the overall procedure data-driven.

Problem Definition. Consider m SCD experts $\{e_1, \ldots, e_m\}$, each expert $e_i : \mathcal{X} \mapsto \mathcal{G}$ can recover a causal graph $\widehat{G}_i = e_i(X)$ from the observational data $X \in \mathcal{X} := \mathbb{R}^{n \times |V|}$, where n is the sample size, $|V|$ is the number of variables. The estimated graph can be a DAG $G_{\text{dag}} = (V, E)$, where $E \subseteq V \times V$ denotes the set of directed edges, a Completed Partially Directed Acyclic Graph (CPDAG) $G_{\text{cpdag}} = (V, E \cup E_u)$ with directed edges E and undirected edges E_u, or a Partial Ancestral Graph (PAG) $G_{\text{pag}} = (V, E^{\{-,\blacktriangleright,\circ\}})$ whose endpoint marks are in $\{-, \blacktriangleright, \circ\}$. Our goal is to ensemble $\{\widehat{G}_1, \ldots, \widehat{G}_m\}$ into a single higher-quality DAG estimate $\widehat{G}$ under the guidance of an LLM.

3.1 Expert Aggregation via Linear Pooling

We study the problem of aggregating heterogeneous expert beliefs on a candidate causal relation r among a subset of variables. Let $\{e_1, \ldots, e_m\}$ denote the set of experts. Given observational data X, we estimate an empirical support score $\hat{p}_i(r) \in [0, 1]$ via repeated subsampling. Let $X^{(1)}, \ldots, X^{(B)}$ denote B resampled datasets and $\widehat{G}_i^{(b)}$ the graph returned by expert e_i on $X^{(b)}$. We define

$$\hat{p}_i(r) := \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{r \in T_\kappa(\widehat{G}_i^{(b)})\}, \tag{1}$$

where $\kappa \in \{\text{sk}, \text{vs}, \text{eg}\}$ indexes the relation family (skeleton, v-structure, or edge orientation) and $T_\kappa(\cdot)$ extracts the corresponding relation set from a graph. Thus $\hat{p}_i(r)$ is the selection frequency with which expert e_i supports relation r across resamples, and larger values indicate stronger support.

Linear Opinion Pooling. Given *normalized* expert credibility weights $\alpha_i \geq 0$ with $\sum_{i=1}^{m} \alpha_i = 1$, the pooled support score of relation r is

$$\hat{p}(r) = \sum_{i=1}^{m} \alpha_i \, \hat{p}_i(r). \tag{2}$$

Unless otherwise specified, we initialize $\alpha_i = 1/m$, treating all experts with equal credibility.

Margin-aware Disputed Set. For $\kappa \in \{\text{sk}, \text{vs}\}$, let $\tau_\kappa \in (0, 1)$ be the acceptance threshold and $\delta > 0$ a margin coefficient. We call a candidate relation $r \in \mathcal{R}_\kappa$ *disputed* whenever

$$\left|\hat{p}(r) - \tau_\kappa\right| \leq \delta \, \tau_\kappa. \tag{3}$$

Only disputed relations are sent to the LLM for expert reweighting. The edge orientation family ($\kappa = \text{eg}$) uses a different dispute criterion based on the strict-maximum rule (which will be discussed later in Section 3.3). Intuitively, bootstrap averaging makes $\hat{p}(r)$ stable, so relations near the threshold τ_κ are more susceptible to threshold-crossing fluctuations and are prioritized for LLM reweighting. Proposition A.1 formalizes this intuition: the base pooled support satisfies a Hoeffding-type concentration bound, and the probability of crossing the threshold decays exponentially with both the number of resamples B and the squared distance of its mean from the threshold.

LLM-guided Reweighting. For each disputed relation r, we invoke an LLM as a meta-referee to produce raw expert-weight scores $w_i(r) \in \{0, 1, \ldots, 10\}$ for all experts, with $\sum_{j=1}^{m} w_j(r) > 0$ enforced by our structured prompt, which requires the LLM to identify at least one credible expert for the relation. We then normalize these scores into relation-specific linear pooling weights:

$$\alpha_i'(r) = \frac{w_i(r)}{\sum_{j=1}^{m} w_j(r)}, \tag{4}$$

and recompute the refined pooled support as

$$\hat{p}'(r) = \sum_{i=1}^{m} \alpha_i'(r)\, \hat{p}_i(r). \tag{5}$$

The reweighting changes only the expert coefficients and does not alter the underlying expert outputs $\hat{p}_i(r)$. For notational convenience in the ensemble procedure, we define the *final* pooled score as

$$\tilde{p}(r) := \begin{cases} \hat{p}'(r), & \text{if } r \text{ is disputed,} \\ \hat{p}(r), & \text{otherwise.} \end{cases} \tag{6}$$

Candidates outside the dispute margin retain the base pooling score $\hat{p}(r)$, so LLM calls are reserved for borderline cases only, reducing unnecessary inference cost without sacrificing ensemble quality.

3.2 LLM AS A META-REFEREE

Rather than directly querying LLMs to infer pairwise causal relations as previous methods (Kıcıman et al., 2024), our approach positions the LLM as a *meta-referee* that intervenes only when data-driven methods are disputed and resolves discrepancies by reweighting experts. Structurally, the linear-pooling form guarantees that LLM reweighting can only redistribute trust among experts, with the resulting score always lying between the smallest and largest individual expert scores (Lemma A.2). For each disputed causal relation r, the LLM performs a meta-analysis over the following information:

Meta Context. To ground the LLM's adjudication with precise information and sufficient background evidence, we first prompt the LLM with a *meta context*. This includes: (i) ***Causal Relation Type*** (sk / vs / eg), which explicitly specifies the graph structural level under consideration, encouraging the LLM to apply the appropriate graph rules (e.g., adjacency vs. orientation) and, crucially, to differentiate which experts are expected to be more reliable at that level (i.e., some methods are stronger at skeleton recovery while others are more accurate in orienting edges); (ii) ***Variable Names and Descriptions***, which enable the LLM to leverage domain knowledge (e.g., time order, physical constraints, measurement meanings) to assess whether a candidate adjacency or direction is causally plausible, thereby reducing spurious inferences; (iii) ***Dataset Profile***. Rather than providing only a high-level domain narrative, we provide *automatically* summarized dataset statistics so that the LLM can check whether each expert's assumptions and methodology plausibly hold for the observed data, and thus reweight experts more appropriately; and (iv) ***Agreement Summary*** (how many experts support vs. oppose), which provides a coarse but stable aggregation signal, helping the LLM calibrate its adjudication and avoid overreacting to a single expert when the ensemble exhibits broad consensus.

Expert Opinions. Analogous to a meta-review process, we cast the LLM as a *Meta-referee* and treat SCD experts as *Reviewers* that provide assessments for the candidate relation, enabling the LLM to compare arguments rather than raw votes. For each expert, we construct an *opinion* containing: (i) ***Methodology Profile***, which includes the expert name and method type (e.g., constraint-based, score-based or FCM), key assumptions (e.g., causal sufficiency, faithfulness, functional form, noise

model), and any latent confounder handling capability so the LLM can distinguish what statistical signal the expert is using and whether its characteristics are compatible with the dataset; (ii) ***Expert Stance***, which shows the expert's explicit support/oppose information for the candidate relation, removing ambiguity and making the opinion directly comparable across experts; and (iii) ***Expert Configuration***, which contains the score function or conditional-independence test method that the expert used, with regularization and penalty choices, and known strengths or failure modes. These factors help the LLM attribute disagreements among experts to method-specific biases and better weight experts by calibrating how suitable the expert is for the current data.

LLM Response. Given the meta context and expert opinions, the LLM conducts a relation-specific meta-analysis to 1) evaluate *expert-dataset compatibility*, 2) assess *causal plausibility*, and 3) judge which expert is more reliable at the current graph level. The LLM then outputs a raw weight score $w_i(r)$ for each expert e_i, which is normalized to linear pooling weights $\alpha_i'(r)$ via Eq. (4) and used to compute the refined support $\hat{p}'(r)$ via Eq. (5). We require a JSON response from the LLM, together with a brief per-expert analysis and a short weight justification for decision transparency.

3.3 Hierarchical Ensemble

We consider three types of structural relations in the causal graph, indexed by $\kappa \in \{\mathrm{sk}, \mathrm{vs}, \mathrm{eg}\}$ for graph skeletons, v-structures, and edge orientations, respectively. CaRe ensembles these relations layer by layer in a coarse-to-fine, structure-to-direction manner, detailed below.

Skeleton Ensemble. We start by ensembling the skeletons of causal graphs produced by different experts. Given the skeleton edge set $\widehat{E}_{i,\mathrm{sk}}$ (a set of undirected edges) extracted from each expert's causal graph $\widehat{G}_i$, we pool their beliefs on every skeleton relation r and define the ensembled edge set

$$\widehat{E}_{\mathrm{sk}} = \Big\{ r \in \bigcup_{i=1}^{m} \widehat{E}_{i,\mathrm{sk}} : \tilde{p}(r) \geq \tau_{\mathrm{sk}} \Big\}, \tag{7}$$

where $\tilde{p}(r)$ is the final pooled score defined in Eq. (6) and τ_{sk} is the skeleton acceptance threshold. This yields the consensus graph skeleton $\widehat{G}_{\mathrm{sk}} = (V, \widehat{E}_{\mathrm{sk}})$.

V-structure Ensemble. Based on $\widehat{G}_{\mathrm{sk}}$, let $\widehat{\mathcal{V}}_i$ denote the set of v-structures extracted from expert e_i's graph. Each v-structure $r = (a, b, c)$ corresponds to $a \to b \leftarrow c$ and is *valid* only if $\{a, b\}, \{b, c\} \in \widehat{E}_{\mathrm{sk}}$ and a, c are nonadjacent in $\widehat{G}_{\mathrm{sk}}$. The consensus v-structure set $\widehat{\mathcal{V}}$ can be obtained by:

$$\widehat{\mathcal{V}} = \Big\{ r \in \bigcup_{i=1}^{m} \widehat{\mathcal{V}}_i : \tilde{p}(r) \geq \tau_{\mathrm{vs}} \Big\}. \tag{8}$$

We then sort candidates by descending $\tilde{p}(r)$ and iteratively insert *non-conflicting* (i.e., not contradicting any previously oriented edge) v-structures into the current $\widehat{G}_{\mathrm{sk}}$. After each insertion, we check whether the newly oriented edges introduce a directed cycle; if so, the insertion is rejected and the graph is rolled back. This yields a partially directed acyclic graph $\widehat{G}_{\mathrm{vs}}$. Since Eqs. (7)–(8) only threshold expert-generated candidates, the resulting skeleton edges and v-structures remain within the expert candidate pool, and any adjacency excluded from $\widehat{E}_{\mathrm{sk}}$ cannot be reintroduced by later orientation steps (Proposition A.3). Under candidate coverage and limiting score separation, the threshold-based skeleton and v-structure selections are asymptotically consistent (Theorem A.4).

Edge Orientation Ensemble. In this final stage, we aim to resolve all remaining undirected edges in $\widehat{G}_{\mathrm{vs}}$. For each undirected pair $\{a, b\}$ with three candidate relations $r \in \mathcal{C}_{ab} := \{a \to b, b \to a, a\text{–}b\}$, each expert e_i provides a support score $\hat{p}_i(r)$, where $\hat{p}_i(r) = 0$ when expert e_i's graph does not include the edge $\{a, b\}$. If one of the two directed candidates $a \to b$ or $b \to a$ attains a strict maximum $\hat{p}(r)$ over $\mathcal{C}_{ab}$, we tentatively orient $\{a, b\}$ accordingly. Otherwise, the pair is treated as orientation-disputed and is forwarded to the LLM, which estimates pair-specific expert weights $\alpha_i'(r)$ via Eq. (4). Using these weights, we recompute $\hat{p}'(r)$ and re-check whether a directed candidate now strictly dominates. For each resolved pair with a determined direction, we define its evidence strength as

$$\pi(a, b) = \sum_{r \in \mathcal{C}_{ab}} \tilde{p}(r). \tag{9}$$

We then insert the resolved orientations into $\widehat{G}_{\text{vs}}$ in descending order of $\pi(a, b)$. After each insertion, we apply Meek-rule closure to propagate compelled directions, and skip any tentative orientation that would conflict with an already-compelled direction or introduce a directed cycle. If some edges still remain undirected after this process, typically because the available expert evidence supports $a-b$ only, we use the LLM only as a last-resort completion aid and accept only acyclicity-preserving directions, yielding the final DAG $\widehat{G}$; this validity is guaranteed by Proposition A.5.

4 Experiments

Datasets. We comprehensively include eight benchmark datasets commonly used in causal discovery for our experiments, covering both synthetic and real-world scenarios across various domains. Specifically, we use five synthetic datasets: **Alarm**, **Asia**, **Child**, **Insurance**, and **Sachs** (Lauritzen & Spiegelhalter, 1990), and three real-world datasets: **Alzheimer** (Abdulaal et al., 2024), **Arctic Sea** (Huang et al., 2021), and **Sangiovese** (Kıcıman et al., 2024). For the synthetic benchmarks, the ground-truth causal graphs are specified as Bayesian networks with conditional probability tables from the `bnlearn` repository, and observational data are generated by sampling from the corresponding network distributions (Scutari & Denis, 2021). To evaluate performance under different data regimes, we simulate four sample sizes $n \in \{500, 1000, 5000, 10000\}$ for each synthetic dataset. For real-world datasets, we use the fixed sample sizes and ground-truth graphs provided by the original benchmarks. Detailed dataset descriptions are provided in Appendix C.1.

Baselines. We carefully select three groups of causal discovery methods as our baselines for comparison: SCD methods, ensemble-based methods, and LLM-based methods. For SCD baselines, we cover a diverse range of search paradigms, including: (i) constraint-based methods, **PC** (Spirtes et al., 2000) and **FCI** (Spirtes et al., 1995), (ii) the score-based method, **GES** (Chickering, 2002), (iii) permutation-based methods, **BOSS** (Andrews et al., 2023) and **GRaSP** (Lam et al., 2022), and (iv) the functional causal model, **ICA-LiNGAM** (Shimizu et al., 2006). These six methods also serve as the SCD experts integrated by CaRe. For ensemble-based baselines, we include **DAGBag** (Wang & Peng, 2014) and **Uniform-CaRe**, where Uniform-CaRe is a controlled non-LLM variant of CaRe that keeps the same hierarchical ensemble pipeline but replaces LLM-guided relation-specific weights with uniform weights, serving as a strong ensemble method. For LLM-based baselines, we include **LLM-Greedy** (Long et al., 2023) and **LLM-BFS** (Jiralerspong et al., 2024), which directly query an LLM to infer causal relations under different strategies. All methods are evaluated under the same environment and settings to ensure fairness. Further details on baselines are in Appendix C.2.

4.1 Main Results

Table 1: Causal discovery results, synthetic datasets are averaged over respective sample sizes of $n \in \{500, 1000, 5000, 10000\}$ and real-world datasets are marked with $*$. Full Table in Appendix F.

Datasets	Alarm		Asia		Child		Insurance		Sachs		Alzheimer*		Arctic-Sea*		Sangiovese*	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
BOSS	0.381	0.713	0.531	0.645	0.345	0.747	0.351	0.670	0.515	0.694	**0.125**	**0.900**	0.542	0.483	**0.343**	**0.683**
FCI	0.212	0.795	0.422	0.588	0.315	0.711	0.337	0.648	0.309	0.761	0.281	0.734	0.490	0.523	0.480	0.532
GES	0.139	0.865	0.219	0.806	0.315	0.772	0.303	0.688	0.551	0.597	0.219	0.807	0.562	0.439	0.569	0.451
GRaSP	0.307	0.737	0.641	0.497	0.380	0.721	0.430	0.594	0.522	0.659	0.250	0.775	0.531	0.490	0.431	0.597
ICA-LiNGAM	0.576	0.553	0.391	0.628	0.685	0.440	0.663	0.376	0.596	0.443	0.438	0.585	0.594	0.453	0.765	0.393
PC	0.190	0.808	0.422	0.588	0.305	0.712	0.378	0.600	0.551	0.525	0.406	0.592	0.438	0.583	0.441	0.557
DAGBag	0.709	0.430	0.406	0.599	0.630	0.476	0.445	0.566	0.360	0.677	0.375	0.633	0.531	0.443	0.539	0.431
Uniform CaRe	0.147	0.865	0.281	0.732	0.260	0.755	0.300	0.683	0.456	0.691	0.219	0.798	0.469	0.549	0.412	0.612
LLM-Greedy	0.457	0.551	0.625	0.414	0.440	0.574	0.327	0.696	0.412	0.617	0.438	0.606	0.448	0.591	0.471	0.591
LLM-BFS	0.489	0.545	0.188	0.834	0.340	0.677	0.279	0.717	0.559	0.070	0.156	0.861	0.458	0.583	0.539	0.421
CaRe (Ours)	**0.049**	**0.952**	**0.093**	**0.918**	**0.140**	**0.864**	**0.207**	**0.793**	**0.169**	**0.838**	**0.125**	**0.900**	**0.427**	**0.597**	0.373	0.652

Table 1 reports the causal discovery results on eight benchmark datasets, where synthetic datasets are averaged over four sample sizes and real-world datasets are marked with $*$. We employ `GPT-5.5` as the backbone LLM and evaluate all methods using two metrics: **RNB** (Ratio between Normalized Hamming Distance and Baseline Hamming Distance) and **AUC** (Area Under the Precision-Recall Curve), where lower RNB and higher AUC indicate better performance. The best values are in **bold**, and the second-best are underlined. Overall, CaRe achieves the strongest performance across diverse

domains, obtaining an average relative RNB reduction of 28.3% and an average AUC improvement of 6.2% compared with the strongest non-CaRe baselines over datasets. These gains indicate that the proposed hierarchical ensemble with LLM-guided expert reweighting can effectively exploit complementary signals from heterogeneous SCD methods, rather than relying on a single solver. The gains are particularly pronounced on *Alarm*, *Asia*, *Child*, and *Sachs*, where individual SCD methods exhibit clear domain dependence, while CaRe mitigates this solver-selection problem and delivers robust cross-domain performance. On *Arctic-Sea**, the relatively smaller gain can be attributed to its ground-truth structure containing many bidirectional relations and cycles, which departs from the DAG-oriented output space and makes our DAG recovery intrinsically harder. Notably, Uniform-CaRe already performs competitively among non-LLM baselines, and our CaRe further improves over it across benchmarks, showing that context-aware expert reweighting provides additional benefits beyond uniform ensembling. We provide a more detailed ablation of this effect in Section 4.2.

4.2 Method Analysis

Ablation Study 1. We consider two main ablations: 1) We first examine whether LLM-guided expert reweighting is essential to the performance of CaRe, and replace the LLM estimated weights (w/ LLM) with uniform weights (w/o LLM). This variant corresponds to Uniform-CaRe in Table 1, and any performance gap between w/ LLM and w/o LLM directly reflects the contribution of expert weights estimation; 2) We then investigate whether the LLM is better used as a meta-referee than as a direct causal expert, where the (Causal Expert) variant uses the same meta context as CaRe but asks the LLM to infer the full causal graph directly. As shown in Table 2, replacing LLM-guided expert weights with uniform weights consistently degrades CaRe's performance across all datasets, leading to a relative drop of **41.9%** in RNB and **12.4%** in AUC. This indicates that simply ensembling heterogeneous SCD experts is insufficient: different experts should be trusted differently depending on the relation type, dataset characteristics, and expert-data compatibility. LLM-guided reweighting provides this relation-specific calibration, enabling CaRe to resolve expert disagreements more effectively than uniform ensembling. Moreover, replacing the LLM's role from meta-referee to direct causal expert also weakens overall performance, leading to a relative drop of **35.7%** in RNB and **15.2%** in AUC. This supports our central design choice: LLMs are more reliable when used to calibrate data-driven experts than when asked to directly infer causal graphs. Notably, the direct causal expert achieves perfect performance on *Asia*. We regard this as a potential benchmark-leakage signal, since *Asia* is a canonical causal discovery benchmark whose graph structure is likely to appear in public materials. Direct LLM querying may therefore recover memorized benchmark structure rather than infer the graph from the provided context. In contrast, CaRe restricts the LLM to expert reweighting, mitigating the leakage risk while keeping the final graph grounded in SCD outputs.

Table 2: Ablations of CaRe. Comparing LLM-guided expert reweighting (w/ LLM) with uniform weights (w/o LLM), and comparing different LLM roles, where the LLM acts either as a meta-referee for expert aggregation or as a direct causal expert.

LLM Abl.	Weight Assignment				LLM Role			
Replace	**w/ LLM**		w/o LLM		**Meta-referee**		Causal Expert	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
Alarm	**0.049**	**0.952**	0.147	0.865	**0.049**	**0.952**	0.117	0.919
Asia	**0.093**	**0.918**	0.281	0.732	0.093	0.918	**0.000**	**1.000**
Child	**0.140**	**0.864**	0.260	0.755	**0.140**	**0.864**	0.665	0.466
Insurance	**0.207**	**0.793**	0.300	0.683	**0.207**	**0.793**	0.356	0.695
Sachs	**0.169**	**0.838**	0.456	0.691	**0.169**	**0.838**	0.441	0.533
Alzheimer*	**0.125**	**0.900**	0.219	0.798	**0.125**	**0.900**	0.156	0.862
Arctic-Sea*	**0.427**	**0.597**	0.469	0.549	**0.427**	0.597	0.448	**0.607**
Sangiovese*	**0.373**	**0.652**	0.412	0.612	**0.373**	**0.652**	0.471	0.528
Performance	**100%**	**100%**	−41.9%	−12.4%	**100%**	**100%**	−35.7%	−15.2%

Table 3: Ablations of CaRe. Comparing the full LLM prompt with variants that remove meta context (w/o Context), expert opinions (w/o Opinions), or both (w/o Both), measuring their impact on LLM-guided expert reweighting.

Prompt Abl.	Prompt Content							
Content	**Full**		w/o Context		w/o Opinions		w/o Both	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
Alarm	**0.049**	**0.952**	0.114	0.887	0.147	0.865	0.147	0.865
Asia	**0.093**	**0.918**	0.344	0.657	0.281	0.732	0.281	0.732
Child	**0.140**	**0.864**	0.255	0.759	0.260	0.755	0.260	0.755
Insurance	**0.207**	**0.793**	0.305	0.674	0.300	0.683	0.300	0.683
Sachs	**0.169**	**0.838**	0.419	0.665	0.456	0.691	0.456	0.691
Alzheimer*	**0.125**	**0.900**	0.219	0.798	0.219	0.798	0.219	0.798
Arctic-Sea*	**0.427**	**0.597**	0.500	0.500	0.469	0.549	0.469	0.549
Sangiovese*	**0.373**	**0.652**	0.412	0.601	0.412	0.612	0.412	0.612
Performance	**100%**	**100%**	−41.7%	−14.8%	−41.9%	−12.4%	−41.9%	−12.4%

Ablation Study 2. We further investigate whether the information provided to the LLM is necessary for reliable expert reweighting. Specifically, we ablate two key prompt components: 1) *Meta Context* and 2) *Expert Opinions*. As shown in Table 3, removing the meta context substantially weakens CaRe, dropping the performance of **41.7%** in RNB and **14.8%** in AUC. This suggests that variable semantics, dataset characteristics, and relation-type information are important for the LLM to assess

expert-data compatibility and produce accurate weights. Note that removing expert opinions also causes clear degradation, decreasing RNB by **41.9%** and AUC by **12.4%** on average, implying that the LLM needs structured expert-specific evidence to distinguish which experts should be trusted more for each relation. Interestingly, removing Expert Opinions leads to almost the same results as w/o LLM in Table 2. This indicates that when the LLM is given only variable and dataset information but no concrete expert arguments, it does not arbitrarily favor particular experts based on names or prior knowledge. Instead, it falls back to conservative uniform weighting. Therefore, both meta context and expert opinions are essential: the former provides the background needed, while the latter supplies the expert-level evidence required for meaningful relation-specific reweighting.

Weight-Performance Alignment. To further examine whether the LLM reweighting is meaningful, we visualize the distribution of LLM-assigned expert weights and compare them with empirical expert performance on the *Insurance* dataset. As shown in Figure 3, experts with stronger AUC performance generally receive higher average weights: GES obtains the largest mean weight (7.63) and strong AUC, while BOSS, FCI, PC, and GRaSP receive moderate-to-high weights. In contrast, LiNGAM shows the weakest AUC and is clearly down-weighted, with a mean weight of only 2.36. The heatmap further shows high alignment scores for most experts, suggesting that the LLM does not assign weights arbitrarily but calibrates expert credibility in a way that is consistent with their empirical reliability. This provides additional evidence that CaRe's improvement comes from relation-specific expert calibration rather than from naive ensemble averaging.

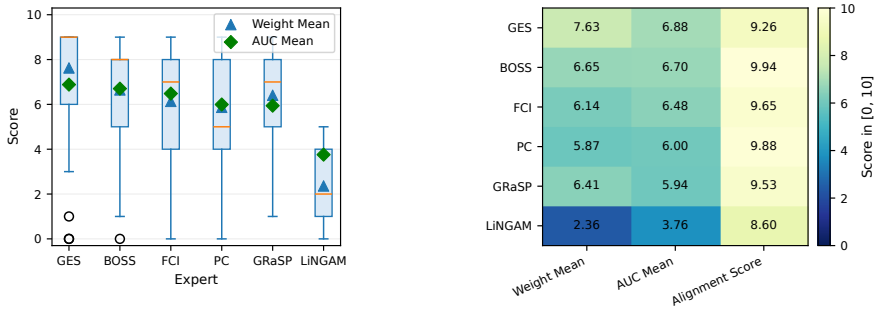

Figure 3: Weight-performance alignment on the Insurance dataset. *Left:* distribution of LLM-assigned expert weights across all queries, with triangles denoting mean weights and diamonds denoting expert mean AUC scores rescaled to $[0, 10]$ over four different sample sizes. *Right:* heatmap of mean LLM weight, mean AUC, and alignment score for each expert. The LLM assigns larger weights to stronger experts, indicating that CaRe's reweighting is aligned with empirical expert reliability.

4.3 Additional Experiments

We provide additional results and analyses in Appendix D, particularly examining benchmark-leakage concerns and the robustness of CaRe. We also include evaluations of query cost, hierarchical design, margin sensitivity, expert reliance rate, and final DAG completion strategies.

5 Conclusion

In this paper, we studied how to reconcile conflicting outputs from heterogeneous statistical causal discovery methods and discovered a more reliable role for LLMs in causal discovery beyond direct causal prediction. We proposed **Ca**usal **Re**feree (**CaRe**), a hierarchical causal discovery ensemble framework that aggregates relation beliefs through linear opinion pooling and invokes an LLM to reweight experts only for disputed relations. By progressively integrating skeletons, v-structures, and edge orientations, CaRe constructs a causal graph while combining the reliability of SCD methods with domain knowledge. Extensive experiments on both synthetic and real-world benchmarks show that CaRe achieves the strongest overall performance against a wide range of causal discovery methods, highlighting the effectiveness of positioning LLMs as meta-referees for causal discovery. Future improvements may include: 1) improving context and opinion construction quality, and 2) exploring more adaptive LLM invocation to better trade off performance and time cost.

AI USE STATEMENT

We use AI tools to assist with writing and language polishing, including improving clarity, refining wording, and checking the consistency of reported results and descriptions. These checks are fully based on existing experimental results. We have reviewed the AI-assisted content and take full responsibility for this work.

ETHICS STATEMENT

This work focuses on methodological research in causal discovery using synthetic data and public benchmark datasets. The proposed framework and the reported evaluations are intended for research use. We are not aware of any ethical concerns arising from this study.

REPRODUCIBILITY STATEMENT

We describe the proposed framework in Section 3. The appendix provides theoretical analyses, pseudocode, dataset descriptions, baselines, implementation details, and LLM prompt templates to facilitate reproducibility. The supplementary material provides source code and reproduction instructions for CaRe, together with all methods' corresponding scripts and execution commands. We also provide a direct download link for all well-prepared datasets. We will publicly release the source code once the paper is accepted.

REFERENCES

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pAoqRlTBtY.

Bryan Andrews, Joseph Ramsey, Ruben Sanchez Romero, Jazmin Camchong, and Erich Kummerfeld. Fast scalable and accurate discovery of DAGs using the best order score search and grow shrink trees. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=80g3Yqlo1a.

Babak Aslani and Shima Mohebbi. Ensemble framework for causality learning with heterogeneous directed acyclic graphs through the lens of optimization. *Computers & Operations Research*, 152: 106148, 2023. ISSN 0305-0548. doi: https://doi.org/10.1016/j.cor.2023.106148. URL https://www.sciencedirect.com/science/article/pii/S0305054823000126.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach, 2013. URL https://arxiv.org/abs/1301.6695.

Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Pei Guo, Yiyi Huang, and Jianwu Wang. Scalable and flexible two-phase ensemble algorithms for causality discovery. *Big Data Research*, 26:100252, 2021. ISSN 2214-5796. doi: https://doi.org/10.1016/j.bdr.2021.100252. URL https://www.sciencedirect.com/science/article/pii/S2214579621000691.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems*, 21, 2008.

Yiyi Huang, Matthäus Kleindessner, Alexey Munishkin, Debvrat Varshney, Pei Guo, and Jianwu Wang. Benchmarking of data-driven causality discovery approaches in the interactions of arctic sea ice and atmosphere. *Frontiers in big Data*, 4:642182, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Thomas Jiralerspong, Xiaoyin Chen, Yash More, Vedant Shah, and Yoshua Bengio. Efficient causal graph discovery using large language models. In *ICLR 2024 Workshop: How Far Are We From AGI*, 2024. URL https://openreview.net/forum?id=5RBUTx75yr.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Wai-Yin Lam, Bryan Andrews, and Joseph Ramsey. Greedy relaxations of the sparsest permutation algorithm. In James Cussens and Kun Zhang (eds.), *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1052–1062. PMLR, 01–05 Aug 2022. URL https://proceedings.mlr.press/v180/lam22a.html.

S. L. Lauritzen and D. J. Spiegelhalter. *Local computations with probabilities on graphical structures and their application to expert systems*, pp. 415–448. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990. ISBN 1558601252.

Zihao Li and Fabrizio Russo. Leveraging large language models for causal discovery: a constraint-based, argumentation-driven approach, 2026. URL https://arxiv.org/abs/2602.16481.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*, 2022. URL https://openreview.net/forum?id=LQQoJGw8JD1.

Stephanie Long, Alexandre Piché, Valentina Zantedeschi, Tibor Schuster, and Alexandre Drouin. Causal discovery with language models as imperfect experts. In *ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling*, 2023. URL https://openreview.net/forum?id=RXlvYZAE49.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pp. 403–410, 1995.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Marco Scutari and Jean-Baptiste Denis. *Bayesian Networks with Examples in R*. Chapman and Hall, Boca Raton, 2nd edition, 2021. ISBN 978-0367366513.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006. URL http://jmlr.org/papers/v7/shimizu06a.html.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.

Peter Spirtes, Christopher Meek, and Thomas Richardson. Causal inference in the presence of latent variables and selection bias. UAI'95, pp. 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1558603859.

Gideon Stein, Maha Shadaydeh, Jan Blunk, Niklas Penzel, and Joachim Denzler. Causalrivers - scaling up benchmarking of causal discovery for real-world time-series. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=wmV4cIbgl6.

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, 2025. URL https://arxiv.org/abs/2402.01454.

Vy Vo, Haoxuan Li, and Mingming Gong. Causal discovery in the wild: A voting-theoretic ensemble approach. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=WtbPaWO8lH.

Ru Wang and Jie Peng. Learning directed acyclic graphs via bootstrap aggregating, 2014. URL https://arxiv.org/abs/1406.2098.

Moritz Willig, Matej Zečević, Devendra Singh Dhami, and Kristian Kersting. Can foundation models talk causality?, 2022. URL https://arxiv.org/abs/2206.10591.

Kun Zhang and Aapo Hyvärinen. On the identifiability of the post-nonlinear causal model. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 647–655, 2009.

Qiqi Zhang, Yingwei Zhang, Yanhui Ding, Yiqiang Chen, Shuchao Song, and Shuang Wu. A weighted ensemble causal discovery method for effective connectivity estimation. In *2023 IEEE Smart World Congress (SWC)*, pp. 1–8, 2023. doi: 10.1109/SWC57546.2023.10448964.

Yuzhe Zhang, Yipeng Zhang, Yidong Gan, Lina Yao, and Chen Wang. Causal graph discovery with retrieval-augmented generation based large language models, 2024. URL https://arxiv.org/abs/2402.15301.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2025. URL https://arxiv.org/abs/2303.18223.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. Causal-learn: Causal discovery in python, 2023. URL https://arxiv.org/abs/2307.16405.

A Additional Theoretical Analyses

We provide concise guarantees for the components of CaRe that are directly used in Section 3: bootstrap-based linear pooling, margin-aware LLM invocation, LLM-guided reweighting, candidate coverage, threshold-based layer consistency, and final DAG validity. The statements below focus on the causally sufficient DAG/CPDAG setting. PAG-valued experts can still be used empirically after applying a deterministic preprocessing map to the relation families below, but the guarantees are stated for the resulting skeleton, v-structure, and edge-orientation relations.

Relation families. Fix an arbitrary total order on the variable set V; we write $a < b$ only to avoid duplicate unordered relations. Define

$$\mathcal{R}_{\mathrm{sk}} := \big\{\{a,b\} : a,b \in V,\ a<b\big\},$$

$$\mathcal{R}_{\mathrm{vs}} := \big\{(a,b,c) : a,b,c \in V \text{ are distinct},\ a<c\big\},$$

and, for each unordered pair $\{a,b\}$ with $a<b$,

$$\mathcal{C}_{ab} := \big\{a\to b,\ b\to a,\ a-b\big\}, \qquad \mathcal{R}_{\mathrm{eg}} := \bigcup_{a<b} \mathcal{C}_{ab}.$$

Here sk, vs, and eg match the notation in Eq. (1). For a graph G, $T_{\mathrm{sk}}(G)$ extracts adjacencies, $T_{\mathrm{vs}}(G)$ extracts unshielded colliders $a\to b\leftarrow c$, and $T_{\mathrm{eg}}(G)$ extracts, for each represented adjacent pair, one of the three states in $\mathcal{C}_{ab}$ whenever the pair is directed or undirected in G. If the pair is absent, no state in $\mathcal{C}_{ab}$ is returned. For threshold-based relations, namely $\kappa \in \{\mathrm{sk},\mathrm{vs}\}$, the LLM-reweighted coefficients are relation-specific as in Eq. (4). For edge orientation, the comparison among $a\to b$, $b\to a$, and $a-b$ is well-defined when the same pair-level weights are used for all states in $\mathcal{C}_{ab}$. Thus, in eg layer, the notation $\alpha_i'(r)$ should be read as a pair-level vector $\alpha_i'(a,b)$ shared by all $r \in \mathcal{C}_{ab}$.

A.1 Concentration of Pooled Support

For each expert e_i, relation $r \in \mathcal{R}_\kappa$, and resampled dataset $X^{(b)}$, define

$$Z_{i,b}(r) := \mathbf{1}\big\{r \in T_\kappa(\widehat{G}_i^{(b)})\big\}, \qquad \hat{p}_i(r) = \frac{1}{B}\sum_{b=1}^{B} Z_{i,b}(r).$$

Let $\mathcal{F}_0$ contain all information used to choose a normalized weight vector $\alpha(r) = (\alpha_1(r),\dots,\alpha_m(r))$, excluding the bootstrap randomness used in the support estimate below. Conditional on $\mathcal{F}_0$, assume that the B resampled datasets and any algorithmic randomness are independent across b. For fixed realized weights $\alpha_i(r) \ge 0$ with $\sum_i \alpha_i(r) = 1$, define

$$S_b(r) := \sum_{i=1}^{m} \alpha_i(r) Z_{i,b}(r) \in [0,1], \qquad \hat{p}_\alpha(r) := \frac{1}{B}\sum_{b=1}^{B} S_b(r),$$

and

$$q_\alpha(r) := \mathbb{E}\big[S_b(r) \mid \mathcal{F}_0\big].$$

Proposition A.1 (Concentration of pooled support and selective LLM invocation)**.** *Under the conditional independence assumption above, for every $\eta > 0$,*

$$\Pr\Big(\big|\hat{p}_\alpha(r) - q_\alpha(r)\big| \ge \eta \;\Big|\; \mathcal{F}_0\Big) \le 2\exp(-2B\eta^2).$$

Consequently, for the threshold decision

$$\hat{y}_\alpha(r) := \mathbf{1}\{\hat{p}_\alpha(r) \ge \tau_\kappa\}, \qquad y_\alpha^\star(r) := \mathbf{1}\{q_\alpha(r) \ge \tau_\kappa\},$$

we have

$$\Pr\Big(\hat{y}_\alpha(r) \ne y_\alpha^\star(r) \;\Big|\; \mathcal{F}_0\Big) \le 2\exp\Big(-2B\big|q_\alpha(r) - \tau_\kappa\big|^2\Big).$$

In particular, if $|q_\alpha(r) - \tau_\kappa| \ge \epsilon$, then the thresholding error is at most $2e^{-2B\epsilon^2}$.

Proof. Conditional on $\mathcal{F}_0$, the weights are fixed and $S_1(r), \ldots, S_B(r)$ are independent random variables supported on $[0,1]$ with mean $q_\alpha(r)$. Hoeffding's inequality (Hoeffding, 1963) gives the concentration bound. A thresholding error can occur only when $\hat{p}_\alpha(r)$ crosses τ_κ while $q_\alpha(r)$ remains on the other side, so

$$\{\hat{y}_\alpha(r) \neq y^\star_\alpha(r)\} \subseteq \Big\{ |\hat{p}_\alpha(r) - q_\alpha(r)| \geq |q_\alpha(r) - \tau_\kappa| \Big\}.$$

Applying the first bound with $\eta = |q_\alpha(r) - \tau_\kappa|$ proves the thresholding bound. □

Intuition. Once the expert weights are fixed, the pooled support is an average of bounded bootstrap quantities, so it concentrates as B grows. Relations whose mean support is far from the threshold are therefore already stable; the relations most worth sending to the LLM are exactly the near-boundary cases selected by Eq. (3).

Remark on LLM-reweighted scores. The proposition applies directly to the base score in Eq. (2). It also applies to LLM-reweighted scores when the LLM weights are computed from information independent of the bootstrap randomness used in $\hat{p}_\alpha(r)$ (e.g., from fixed metadata or an independent pilot split). If the same bootstrap indicators are used both to choose the LLM weights and to estimate the refined score, the concentration statement should be read as the guarantee for the pre-reweighting score that triggers LLM invocation, while the refined score is an adaptive reweighted estimate.

A.2 Grounding and Candidate Coverage

Lemma A.2 (Convex-hull property of linear reweighting)**.** *For any threshold-layer relation $r \in \mathcal{R}_\kappa$ with $\kappa \in \{\mathrm{sk}, \mathrm{vs}\}$, the final pooled score $\tilde{p}(r)$ in Eq. (6) satisfies*

$$\min_{1 \leq i \leq m} \hat{p}_i(r) \leq \tilde{p}(r) \leq \max_{1 \leq i \leq m} \hat{p}_i(r).$$

For edge orientation, the same statement holds for each state $s \in \mathcal{C}_{ab}$ when the shared pair-level weights $\alpha'_i(a,b)$ are used.

Proof. Both the base score $\hat{p}(r) = \sum_i \alpha_i \hat{p}_i(r)$ and the reweighted score $\hat{p}'(r) = \sum_i \alpha'_i(r) \hat{p}_i(r)$ are convex combinations of $\hat{p}_1(r), \ldots, \hat{p}_m(r)$, because the weights are nonnegative and sum to one. Any convex combination lies between the minimum and maximum of its entries. The edge-orientation statement follows by the same argument state-wise for $s \in \mathcal{C}_{ab}$. □

Intuition. The LLM can change how much CaRe trusts each expert, but it cannot create a support score outside the range supplied by the experts. Thus the LLM acts as a meta-referee for expert credibility rather than as a direct generator of causal relations.

Proposition A.3 (Candidate-coverage ceiling)**.** *Let*

$$\mathcal{A}_{\mathrm{sk}} := \bigcup_{i=1}^{m} \widehat{E}_{i,\mathrm{sk}}$$

be the skeleton candidate set in Eq. (7). Given the consensus skeleton $\widehat{G}_{\mathrm{sk}} = (V, \widehat{E}_{\mathrm{sk}})$, let

$$\mathcal{A}_{\mathrm{vs}}(\widehat{G}_{\mathrm{sk}}) := \Big\{ (a,b,c) \in \bigcup_{i=1}^{m} \widehat{\mathcal{V}}_i : \{a,b\}, \{b,c\} \in \widehat{E}_{\mathrm{sk}},\ \{a,c\} \notin \widehat{E}_{\mathrm{sk}} \Big\}$$

be the v-structure candidates that satisfy the skeleton-validity condition in Section 3.3. Then

$$\widehat{E}_{\mathrm{sk}} \subseteq \mathcal{A}_{\mathrm{sk}}, \qquad \widehat{\mathcal{V}} \subseteq \mathcal{A}_{\mathrm{vs}}(\widehat{G}_{\mathrm{sk}}).$$

Moreover, if $\{a,b\} \notin \widehat{E}_{\mathrm{sk}}$, no later edge-orientation step can introduce either $a \to b$ or $b \to a$ into the final graph.

Proof. Eqs. (7) and (8), together with the stated v-structure validity filter, only threshold candidates already proposed by the expert library. Later orientation steps operate on the remaining undirected edges of $\widehat{G}_{\mathrm{vs}}$, whose adjacencies come from $\widehat{E}_{\mathrm{sk}}$. Therefore an adjacency absent from $\widehat{E}_{\mathrm{sk}}$ cannot be introduced later. □

Intuition. CaRe is an ensemble-and-calibration method, not a relation generator. If a true relation is never proposed by any expert, LLM reweighting has no candidate support on which to act. This is why full-recovery guarantees require a candidate-coverage assumption.

Remark on Meek-rule closure. The Meek-rule applied during edge orientation does not introduce *new adjacencies*: it only propagates compelled orientations along edges already present in the current PDAG (Meek, 1995). Combined with the cycle and conflict roll-back in Section 3.3, this ensures that any adjacency excluded from $\widehat{E}_{\mathrm{sk}}$ cannot be reintroduced by later orientation steps. The selected v-structures are introduced only during the v-structure layer and remain within $\widehat{\mathcal{V}} \subseteq \mathcal{A}_{\mathrm{vs}}(\widehat{G}_{\mathrm{sk}})$.

A.3 Consistency of Threshold-based Layers

Fix $\kappa \in \{\mathrm{sk}, \mathrm{vs}\}$. Let $Y_\kappa(r) \in \{0,1\}$ denote whether relation $r \in \mathcal{R}_\kappa$ is present in the target graph layer. Let $\mathcal{A}_\kappa \subseteq \mathcal{R}_\kappa$ be the candidate set evaluated by the corresponding CaRe layer. For $\kappa = \mathrm{vs}$, this candidate set is understood after the skeleton-validity filtering in Proposition A.3.

For sample size n, write the final threshold-layer score as

$$\tilde{p}^{(n)}(r) = \sum_{i=1}^{m} \tilde{\alpha}_i^{(n)}(r)\hat{p}_i^{(n)}(r),$$

where $\tilde{\alpha}_i^{(n)}(r) = \alpha_i$ for non-disputed relations and $\tilde{\alpha}_i^{(n)}(r) = \alpha_i'(r)$ for disputed relations.

Theorem A.4 (Consistency under candidate coverage and score separation)**.** *Assume that m and $|V|$ are fixed. Suppose candidate coverage holds:*

$$\{r \in \mathcal{R}_\kappa : Y_\kappa(r) = 1\} \subseteq \mathcal{A}_\kappa.$$

For every $r \in \mathcal{A}_\kappa$ and expert $i \in \{1, \dots, m\}$, assume

$$\hat{p}_i^{(n)}(r) \xrightarrow{p} \theta_i(r) \in [0,1], \qquad \tilde{\alpha}_i^{(n)}(r) \xrightarrow{p} \tilde{\alpha}_i^\star(r),$$

where $\tilde{\alpha}_i^\star(r) \geq 0$ and $\sum_{i=1}^m \tilde{\alpha}_i^\star(r) = 1$. Define

$$\tilde{\theta}(r) := \sum_{i=1}^{m} \tilde{\alpha}_i^\star(r)\theta_i(r).$$

If, for every $r \in \mathcal{A}_\kappa$, there exists $\gamma_r > 0$ such that

$$\big(\tilde{\theta}(r) - \tau_\kappa\big)\big(2Y_\kappa(r) - 1\big) \geq \gamma_r, \tag{10}$$

then the CaRe decision

$$\widehat{Y}_\kappa^{(n)}(r) := \begin{cases} \mathbf{1}\{\tilde{p}^{(n)}(r) \geq \tau_\kappa\}, & r \in \mathcal{A}_\kappa, \\ 0, & r \notin \mathcal{A}_\kappa, \end{cases}$$

recovers the whole threshold-based layer consistently:

$$\Pr\Big(\exists r \in \mathcal{R}_\kappa : \widehat{Y}_\kappa^{(n)}(r) \neq Y_\kappa(r)\Big) \to 0.$$

Proof. For each fixed $r \in \mathcal{A}_\kappa$, the boundedness $0 \leq \hat{p}_i^{(n)}(r) \leq 1$ and the assumed convergence of $\hat{p}_i^{(n)}(r)$ and $\tilde{\alpha}_i^{(n)}(r)$ imply

$$\tilde{\alpha}_i^{(n)}(r)\hat{p}_i^{(n)}(r) \xrightarrow{p} \tilde{\alpha}_i^\star(r)\theta_i(r).$$

Since m is fixed, summing over experts gives $\tilde{p}^{(n)}(r) \xrightarrow{p} \tilde{\theta}(r)$. If $Y_\kappa(r) = 1$, Eq. (10) gives $\tilde{\theta}(r) \geq \tau_\kappa + \gamma_r$, so

$$\Pr\big(\widehat{Y}_\kappa^{(n)}(r) = 0\big) \leq \Pr\big(|\tilde{p}^{(n)}(r) - \tilde{\theta}(r)| \geq \gamma_r\big) \to 0.$$

If $Y_\kappa(r) = 0$, then $\tilde{\theta}(r) \leq \tau_\kappa - \gamma_r$, and the false-positive probability is bounded in the same way. Thus each candidate relation is classified correctly with probability tending to one. Relations outside $\mathcal{A}_\kappa$ are always rejected and are true negatives by candidate coverage. Because $\mathcal{R}_\kappa$ is finite, a union bound proves the layer-wise claim. □

Intuition. The theorem separates two requirements. First, the expert library must cover the true positive relations. Second, after pooling and optional LLM reweighting, the limiting score must fall on the correct side of the threshold with a nonzero margin. Under these two conditions, sampling noise eventually becomes too small to flip the threshold decision.

Remark on local orientation consistency. The edge layer uses an argmax rather than a threshold rule. For a remaining pair $\{a, b\}$, suppose the three final state scores $\tilde{p}_{ab}^{(n)}(s)$, $s \in \mathcal{C}_{ab}$, converge in probability to limits $\mu_{ab}(s)$. If a directed state $s^\star \in \{a \to b, b \to a\}$ has a strict limiting gap,

$$\mu_{ab}(s^\star) > \max_{s \in \mathcal{C}_{ab} \setminus \{s^\star\}} \mu_{ab}(s),$$

then the local strict-maximum rule selects $s^\star$ with probability tending to one. This is the same convergence argument as above: once all three estimated scores are close to their limits, the state with a strict limiting maximum remains the empirical maximum. This statement is local; the final accepted orientation can still be affected by later conflict checks, directed-cycle checks, and Meek-rule closure.

A.4 Validity of the Final DAG

Proposition A.5 (DAG validity under safe orientation and completion)**.** *Let $P_0 := \widehat{G}_{\mathrm{vs}}$ be the partially directed graph produced after the skeleton and v-structure layers, and assume that the directed part of P_0 is acyclic. Suppose each subsequent tentative orientation, together with the Meek-rule closure applied after it (Meek, 1995), is accepted only if the directed part of the resulting partially directed graph remains acyclic and does not contradict an already oriented edge. If all remaining undirected edges are oriented in a way that preserves acyclicity, then the output $\widehat{G}$ is a DAG. Such a completion always exists: take any topological order of the current directed part and orient every remaining undirected edge from the earlier variable to the later variable in that order.*

Proof. The directed part of P_0 is acyclic by assumption. Each accepted orientation step is required to preserve acyclicity after Meek-rule closure, so induction over accepted steps shows that every intermediate directed part is acyclic. For the final completion, the current directed part has a topological order. If every remaining undirected edge is oriented from the earlier endpoint to the later endpoint in this order, then every directed edge in the completed graph points forward in the same order. A directed cycle would require at least one edge to point backward, which is impossible. Hence the completed graph is fully directed and acyclic, i.e., a DAG. □

Intuition. This is an algorithmic safety guarantee. Local evidence determines which directions are attempted first, while conflict and cycle checks prevent invalid orientations. If some edges remain unresolved, a topological-order completion can always orient them without introducing a cycle.

B Reproducibility of CaRe

We provide implementation-oriented details of CaRe to facilitate reproducibility. We first present the full algorithmic procedure, and then describe the source code availability.

B.1 Pseudocode Details

We provide the pseudocode for CaRe in Algorithm 1, which illustrates the complete coarse-to-fine ensemble procedure over skeletons, v-structures, and edge orientations. Note that the pseudocode follows the notation in Section 3: $\hat{p}_i(r)$ denotes the bootstrap support score of expert e_i, $\hat{p}(r)$ denotes the uniformly pooled score, and $\tilde{p}(r)$ denotes the final score after optional LLM-guided reweighting. For edge orientation, the LLM produces pair-level expert weights for an unordered pair $\{a, b\}$, which are shared across the candidate states in $\mathcal{C}_{ab} = \{a \to b, b \to a, a - b\}$.

B.2 Source Code Release

We provide the full source code in the `supplementary material`, including the implementation of CaRe, configuration files, prompt templates, and scripts for reproducing the experimental results. We will publicly release the code repository once the paper is accepted.

Algorithm 1 Procedure of the CaRe Framework

Input: data X, experts $\{e_i\}_{i=1}^m$, thresholds $\tau_{\text{sk}}, \tau_{\text{vs}}$, margin coef. δ, bootstrap count B
Output: final DAG estimate $\widehat{G}$
Run each expert e_i on B resampled datasets and compute support scores $\hat{p}_i(r)$ for all r
Initialize $\alpha_i \leftarrow 1/m$ for all i
// Step 1: Skeleton Ensemble ($\kappa = \text{sk}$)
Initialize $\widehat{E}_{\text{sk}} \leftarrow \emptyset$ and extract skeleton edge set $\widehat{E}_{i,\text{sk}}$ from each expert
for each candidate skeleton relation $r = \{a, b\} \in \bigcup_{i=1}^m \widehat{E}_{i,\text{sk}}$ **do**
  $\hat{p}(r) \leftarrow \sum_i \alpha_i \hat{p}_i(r)$
  if $|\hat{p}(r) - \tau_{\text{sk}}| \leq \delta\, \tau_{\text{sk}}$ **then**
    Query LLM $\rightarrow w_i(r)$; compute $\alpha_i'(r)$ via Eq. (4); $\tilde{p}(r) \leftarrow \hat{p}'(r)$
  else
    $\tilde{p}(r) \leftarrow \hat{p}(r)$
  if $\tilde{p}(r) \geq \tau_{\text{sk}}$ **then**
    Add r to $\widehat{E}_{\text{sk}}$
Set $\widehat{G}_{\text{sk}} \leftarrow (V, \widehat{E}_{\text{sk}})$
// Step 2: V-structure ensemble ($\kappa = \text{vs}$)
Initialize $\widehat{\mathcal{V}} \leftarrow \emptyset$ and extract valid v-structures $\widehat{\mathcal{V}}_i$ under $\widehat{G}_{\text{sk}}$ from each expert e_i
for each candidate v-structure $r = (a, b, c) \in \bigcup_{i=1}^m \widehat{\mathcal{V}}_i$ **do**
  $\hat{p}(r) \leftarrow \sum_i \alpha_i \hat{p}_i(r)$
  if $|\hat{p}(r) - \tau_{\text{vs}}| \leq \delta\, \tau_{\text{vs}}$ **then**
    Query LLM $\rightarrow w_i(r)$; compute $\alpha_i'(r)$ via Eq. (4); $\tilde{p}(r) \leftarrow \hat{p}'(r)$
  else
    $\tilde{p}(r) \leftarrow \hat{p}(r)$
  if $\tilde{p}(r) \geq \tau_{\text{vs}}$ **then**
    Add r to $\widehat{\mathcal{V}}$
Sort $\widehat{\mathcal{V}}$ by descending $\tilde{p}(r)$ and initialize $\widehat{G}_{\text{vs}} \leftarrow \widehat{G}_{\text{sk}}$
for each v-structure $r = (a \rightarrow b \leftarrow c) \in \widehat{\mathcal{V}}$ in sorted order **do**
  if insertion is non-conflicting and acyclicity-preserving **then**
    Insert $a \rightarrow b \leftarrow c$ into $\widehat{G}_{\text{vs}}$
// Step 3: Edge-orientation ensemble ($\kappa = \text{eg}$)
Initialize resolved orientation set $\mathcal{O} \leftarrow \emptyset$
for each undirected pair $\{a, b\}$ in $\widehat{G}_{\text{vs}}$ **do**
  For $r \in \mathcal{C}_{ab} = \{a \rightarrow b,\ b \rightarrow a,\ a - b\}$, $\hat{p}(r) \leftarrow \sum_i \alpha_i\, \hat{p}_i(r)$ ($\hat{p}_i(r) = 0$ if $\{a, b\} \notin \widehat{E}_{i,\text{sk}}$)
  if one directed candidate in $\{a \rightarrow b, b \rightarrow a\}$ is the strict maximum over $\mathcal{C}_{ab}$ **then**
    Add the corresponding tentative orientation to $\mathcal{O}$ and set $\tilde{p}(r) \leftarrow \hat{p}(r)$ for all $r \in \mathcal{C}_{ab}$
  else
    Query LLM $\rightarrow$ pair-level $w_i(r)$; compute $\alpha_i'(r)$; recompute $\tilde{p}(r) \leftarrow \hat{p}'(r)$ for all $r \in \mathcal{C}_{ab}$
    if one directed candidate is now the strict maximum over $\mathcal{C}_{ab}$ **then**
      Add the corresponding tentative orientation to $\mathcal{O}$
Sort $\mathcal{O}$ by descending evidence strength $\pi(a, b) = \sum_{r \in \mathcal{C}_{ab}} \tilde{p}(r)$
for each tentative orientation in $\mathcal{O}$ **do**
  if the orientation is non-conflicting and acyclicity-preserving **then**
    Insert the orientation into $\widehat{G}_{\text{vs}}$ and apply Meek-rule closure
// Step 4: DAG completion
Complete any remaining undirected edges using either deterministic topological extension or LLM-assisted acyclicity-preserving completion
Set $\widehat{G} \leftarrow$ the completed DAG
return $\widehat{G}$

C Experimental Details

C.1 Dataset Details

We evaluate CaRe on eight causal discovery benchmarks, covering both synthetic and real-world scenarios across various domains, such as healthcare, epidemiology, biology, agriculture, and disease. Specifically, we use five synthetic datasets: **Alarm**, **Asia**, **Child**, **Insurance**, and **Sachs** (Lauritzen & Spiegelhalter, 1990), and three real-world datasets: **Alzheimer** (Abdulaal et al., 2024), **Arctic Sea** (Huang et al., 2021), and **Sangiovese** (Kıcıman et al., 2024). For the synthetic benchmarks, their ground-truth causal graphs and conditional probability tables are obtained from the `bnlearn` repository, and observational datasets are generated by sampling the corresponding network distributions with sample sizes $n \in \{500, 1000, 5000, 10000\}$ (Scutari & Denis, 2021). For real-world datasets, they are used as fixed-size real-world or domain-derived benchmarks with ground-truth graphs provided by the corresponding benchmark sources. Table 4 summarizes all used datasets.

Table 4: Dataset details. The number of variables (#Variables) refers to the variable set used in our experiments. Sample size refers to the number of data samples used for making causal discovery. Synthetic datasets are generated from benchmark Bayesian networks with four sample sizes of $n \in \{500, 1000, 5000, 10000\}$, while real-world datasets use fixed sample sizes and benchmark-provided ground-truth graphs. Type indicates whether the dataset is synthetic or real-world.

Dataset	#Variables	Sample Size	Type	Graph Source	Domain
Alarm	37	$\{500, 1000, 5000, 10000\}$	Synthetic	`bnlearn`	Healthcare
Asia	8	$\{500, 1000, 5000, 10000\}$	Synthetic	`bnlearn`	Epidemiology
Child	20	$\{500, 1000, 5000, 10000\}$	Synthetic	`bnlearn`	Child
Insurance	27	$\{500, 1000, 5000, 10000\}$	Synthetic	`bnlearn`	Insurance
Sachs	11	$\{500, 1000, 5000, 10000\}$	Synthetic	`bnlearn`	Biology
Alzheimer	9	1000	Real-world	Abdulaal et al. (2024)	Disease
Arctic-Sea	12	468	Real-world	Huang et al. (2021)	Climate
Sangiovese	14	700	Real-world	Kıcıman et al. (2024)	Agriculture

Alarm. Alarm is a medical monitoring Bayesian network designed for reasoning about patient physiology, anesthesia, and clinical monitoring variables (Lauritzen & Spiegelhalter, 1990; Scutari & Denis, 2021). It includes respiratory, cardiovascular, ventilator, and measurement-related variables, making it a medium-scale healthcare benchmark with multiple interacting physiological subsystems.

Asia. Asia is a medical diagnosis Bayesian network involving travel history, tuberculosis, smoking, lung cancer, bronchitis, X-ray results, and dyspnoea (Lauritzen & Spiegelhalter, 1990; Scutari & Denis, 2021). It is a canonical benchmark for testing disease, symptom, and diagnostic-test relations.

Child. Child is a neonatal medicine Bayesian network describing congenital heart disease and related clinical observations (Lauritzen & Spiegelhalter, 1990; Scutari & Denis, 2021). It contains variables related to birth asphyxia, disease category, cardiac mixing, duct flow, lung flow, hypoxia, chest X-ray findings, grunting, and clinical reports, providing a moderately sized clinical benchmark.

Insurance. Insurance is an automobile insurance Bayesian network that models driver, vehicle, accident, and claim-related variables (Lauritzen & Spiegelhalter, 1990; Scutari & Denis, 2021). It includes factors such as age, socioeconomic status, risk aversion, driving history, vehicle properties, accident severity, theft, medical cost, and property cost, representing a risk-modelling domain.

Sachs. Sachs is a biological signalling benchmark describing interactions among protein and phospholipid signalling readouts (Lauritzen & Spiegelhalter, 1990; Scutari & Denis, 2021). The variables include Akt, Erk, Jnk, Mek, P38, PIP2, PIP3, PKA, PKC, Plcg, and Raf. In our experiments, we use the Bayesian-network version with conditional probability tables as a synthetic benchmark for cellular signalling structure discovery.

Alzheimer. Alzheimer is a biomedical causal discovery benchmark for Alzheimer's disease, constructed from demographic, genetic, neuroimaging, amyloid, tau, and cognitive variables (Abdulaal et al., 2024). The processed variable set used in our experiments includes APOE4, Sex, Age, Education, AV45, P-tau, Brain Volume, Ventricular Volume, and MOCA Score.

Arctic-Sea. Arctic-Sea is a climate-science benchmark describing interactions between Arctic sea ice and atmospheric variables (Huang et al., 2021). It contains variables related to surface heat flux, shortwave and longwave radiation, sea-level pressure, precipitation, humidity, near-surface winds, sea ice, cloud cover, cloud water, and geopotential height. This dataset is challenging because its ground-truth graph contains many bidirectional relations and feedback cycles.

Sangiovese. Sangiovese is an agricultural benchmark based on a conditional linear Gaussian network for modelling Tuscan Sangiovese grape quality (Kıcıman et al., 2024). The variables used in our experiments include vine growth variables, vegetation indices, chlorophyll measurements, grape weight, pruning wood weight, acidity, potassium, sugar content, pH, anthocyanins, and polyphenols.

C.2 Baseline Details

We compare CaRe with three groups of causal discovery baselines: statistical causal discovery (SCD) methods, ensemble-based methods, and LLM-based methods. The individual SCD baselines also serve as the expert library integrated by CaRe. Detailed descriptions of these baselines are as follows:

PC (Spirtes et al., 2000): The Peter-Clark (PC) algorithm is a classical constraint-based causal discovery method. It first estimates the graph skeleton using conditional independence tests, and then orients edges by identifying v-structures and applying graphical orientation rules. PC assumes causal sufficiency, faithfulness, and reliable conditional independence testing, and outputs a CPDAG.

FCI (Spirtes et al., 1995): Fast Causal Inference (FCI) is a constraint-based causal discovery method designed for settings with possible latent confounders and selection bias. It allows more expressive endpoint marks and outputs a PAG representing causal relations that may involve hidden variables.

GES (Chickering, 2002): Greedy Equivalence Search (GES) is a score-based causal discovery method that searches over Markov equivalence classes of DAGs. It optimizes a decomposable score through forward and backward steps, and outputs an equivalence-class representation, typically a CPDAG.

BOSS (Andrews et al., 2023): Best Order Score Search (BOSS) is a permutation-based causal discovery method that searches over variable orderings and constructs causal graphs using a decomposable score. It provides a strong order-search baseline and outputs a PDAG of the learned structure.

GRaSP (Lam et al., 2022): Greedy Relaxations of the Sparsest Permutation (GRaSP) is a permutation-based causal discovery method that searches for sparse causal structures via greedy relaxations over variable orderings. It serves as a complementary order-based baseline to BOSS and outputs a PDAG.

ICA-LiNGAM (Shimizu et al., 2006): Independent Component Analysis Linear Non-Gaussian Acyclic Model (ICA-LiNGAM) is a functional causal model that exploits linearity, non-Gaussianity, acyclicity, and independent noise assumptions to identify directed causal relations. It outputs a DAG.

DAGBag (Wang & Peng, 2014): DAGBag is an ensemble-based causal discovery method based on bootstrap aggregation. It learns multiple DAGs from bootstrap samples and aggregates them into a consensus graph, aiming to reduce estimation variance and improve structural robustness. We include DAGBag as an external ensemble-based causal discovery baseline.

Uniform-CaRe: Uniform-CaRe is a controlled non-LLM variant of our framework. It uses the same expert library, bootstrap support estimation, hierarchical aggregation pipeline, and graph-validity checks as CaRe, but replaces LLM-guided relation-specific expert weights with uniform weights. Thus, Uniform-CaRe serves as a strong controlled ensemble baseline for isolating the contribution of context-aware LLM-guided expert reweighting.

LLM-Greedy (Long et al., 2023): LLM-Greedy is an LLM-based causal discovery baseline, where an LLM provides orientation knowledge for unresolved causal relations beyond Markov equivalence classes. In our setting, we use it as a direct LLM orientation baseline that greedily orients edges in a given skeleton. We follow the public implementation released at `Causal-disco-LLM-imperfect-experts`.

LLM-BFS (Jiralerspong et al., 2024): LLM-BFS is an LLM-based causal discovery baseline that reduces exhaustive pairwise querying by using a breadth-first search strategy over candidate causal relations. We follow the public implementation released at `causal-llm-bfs`.

For all LLM-involved methods, including CaRe and the LLM-based baselines, we use the same LLM backbone and the same variable information whenever applicable to ensure a fair comparison. Further details, including package versions and inference settings, are provided in Appendix C.3.

C.3 Implementation Details

We implement CaRe entirely in Python 3.10, with all individual SCD baselines directly adopted from the `causal-learn` library (Zheng et al., 2023), which is a comprehensive and fair platform for statistical causal discovery. All experiments reported in this paper are conducted on a virtual machine equipped with an AMD EPYC 7T83 64-core CPU. For fairness, we use a fixed set of hyperparameters across datasets within the same data regime and do not tune hyperparameters separately for individual datasets or baselines. Unless otherwise specified, we use bootstrap count $B = 10$, bootstrap sampling ratio 0.75, margin coefficient $\delta = 0.5$, and threshold coefficient 0.5 for all experiments.

For synthetic datasets sampled from benchmark Bayesian networks, we treat the data as discrete and use the `Chi-square` conditional independence test for constraint-based methods and the Bayesian Dirichlet equivalent uniform score (`local_score_BDeu`) for score-based and permutation-based methods. For real-world datasets, we treat the data as continuous and use Fisher's Z test for conditional independence testing, with Bayesian Information Criterion score (`local_score_BIC`) as the score function. All methods are run under the same preprocessing, graph-conversion, and evaluation pipeline. Since different causal discovery algorithms may output DAGs, CPDAGs, or PAGs, we convert their outputs into a common adjacency representation before computing metrics.

C.4 Evaluation Metrics

We evaluate causal discovery performance using two metrics: **RNB** (Ratio between Normalized Hamming Distance and Baseline Hamming Distance) and **AUC** (Area Under the Precision-Recall Curve), which respectively measure structural accuracy relative to baseline normalization and edge-level recovery quality, where lower RNB and higher AUC indicate better performance. Before computing metrics, each predicted graph is converted into a directed adjacency matrix representation.

Let $G = (V, E)$ denote the ground-truth graph and $\widehat{G}$ denote the predicted graph, with $r = |V|$. Let $A, \widehat{A} \in \{0, 1\}^{r \times r}$ be their adjacency matrices, where $A_{ij} = 1$ indicates a directed edge $i \to j$ in G.

Normalized Hamming Distance (NHD): NHD measures the number of edges that appear in one graph but not the other, normalized by the total number of possible directed entries. Following the directed adjacency representation, we define

$$\mathrm{NHD}(G, \widehat{G}) = \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} \mathbf{1}\{A_{ij} \neq \widehat{A}_{ij}\}. \tag{11}$$

In our implementation, anti-causal mistakes are counted twice. For example, predicting $j \to i$ when the ground truth is $i \to j$ contributes two adjacency-matrix errors.

Baseline Hamming Distance (BHD): The baseline graph is defined as a graph containing the same number of directed edges as G, but all of them are incorrect. Let

$$e = \sum_{i=1}^{r} \sum_{j=1}^{r} A_{ij} \tag{12}$$

denote the number of directed edges in the ground-truth graph. We define the BHD as:

$$\mathrm{BHD}(G) = \frac{2e}{r^2}. \tag{13}$$

Since anti-causal mistakes are counted twice, the corresponding baseline number of mistakes is $2e$.

Ratio between NHD and BHD (RNB): We follow the previous work (Kıcıman et al., 2024) and report the RNB as:

$$\mathrm{RNB}(G, \widehat{G}) = \frac{\mathrm{NHD}(G, \widehat{G})}{\mathrm{BHD}(G)} = \frac{\sum_{i=1}^{r} \sum_{j=1}^{r} \mathbf{1}\{A_{ij} \neq \widehat{A}_{ij}\}}{2e}. \tag{14}$$

A lower RNB means that the causal discovery algorithm outperforms the floor baseline by a larger margin. In particular, $\mathrm{RNB} < 1$ indicates that the predicted graph is closer to the ground truth than a graph with the same number of edges but a completely incorrect structure.

Area Under the Precision-Recall Curve (AUC): We also report the area under the precision-recall curve. Given the ground-truth adjacency matrix A and the predicted adjacency matrix $\widehat{A}$, precision and recall are computed as

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{15}$$

where TP, FP, and FN denote true positives, false positives, and false negatives in the predicted directed adjacency matrix. The area under the precision-recall curve is then computed as

$$\mathrm{AUC}(G, \widehat{G}) = \int_0^1 \mathrm{Precision}(\rho)\, d\rho, \tag{16}$$

where ρ denotes recall along the precision-recall curve. In our implementation, AUC is computed using `cdt.metrics.precision_recall` after converting both graphs into adjacency matrices.

D Additional Experiments and Analyses

This section provides additional results and analyses that complement the main experimental results in Section 4. We first examine benchmark-leakage concerns and the robustness of CaRe through benchmark-identity masking, evaluation on a newly generated causal graph, expert removal, and error bars across repeated runs. We then analyze query cost, hierarchical design, margin sensitivity, expert reliance rates, and final DAG-completion strategies. Finally, we discuss the limitations of CaRe.

D.1 Benchmark-Leakage Analysis

We note that restricting the LLM to expert reweighting does not completely rule out the benchmark memorization problem. We therefore conduct two additional LLM's knowledge contamination-controlled evaluations to verify that CaRe's gains are not driven by the LLM recalling the benchmark identity and memorized graph structures.

Benchmark-identity masking. For all five `bnlearn` datasets, which are particularly susceptible to benchmark-leakage concerns, we rerun the full CaRe pipeline under a masked setting in which dataset names are removed, variable names are anonymized with randomly assigned identifiers, variable ordering is randomly permuted, and variable descriptions are independently paraphrased to remove benchmark wording. We retain the semantic meanings of the variables, allowing the LLM to continue using general domain knowledge.

Table 5: Benchmark-identity masking results of CaRe, averaged over four sample sizes of $n \in \{500, 1000, 5000, 10000\}$. Best Expert reports the best individual SCD result for each dataset and metric separately. The best values are in bold.

Dataset	Alarm		Asia		Child		Insurance		Sachs	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
Best Expert	0.139	0.865	0.219	0.806	0.305	0.772	0.303	0.688	0.309	0.761
Uniform-CaRe	0.147	0.865	0.281	0.732	0.260	0.755	0.300	0.683	0.456	0.691
Masked CaRe	0.065	0.935	**0.093**	**0.918**	0.170	0.835	0.231	0.767	0.184	0.825
Original CaRe	**0.049**	**0.952**	**0.093**	**0.918**	**0.140**	**0.864**	**0.207**	**0.793**	**0.169**	**0.838**

Table 5 shows that benchmark-identity masking (`Masked CaRe`) causes only a modest increase in RNB and decrease in AUC over original performance, with Masked CaRe continuing to outperform both Uniform-CaRe and the best individual expert on both metrics across all datasets. Therefore, the main gains persist after benchmark identity and wording are removed, indicating that the strong performance of CaRe is not due to ground-truth memorization but to leveraging domain knowledge.

CauseNet-grounded synthetic data generation. We also follow the CauseNet-grounded synthetic evaluation protocol (Li & Russo, 2026) and construct an entirely new SCM whose graph topology is not taken from any published benchmark while retaining meaningful real-world variable semantics. The synthetic graph is a 10-node, 15-edge Erdős–Rényi (ER) DAG scaffold from the psychosocial and family-health domain. We evaluate our method on this unseen graph.

Table 6: Results on a newly generated CauseNet-grounded ER DAG with 10 nodes and 15 edges, averaged over four sample sizes of $n \in \{500, 1000, 5000, 10000\}$. The best values are in bold.

Method	CauseNet (ER DAG)	
Metric	RNB (↓)	AUC (↑)
BOSS	0.208	0.801
FCI	0.384	0.675
GES	0.442	0.593
GRaSP	0.275	0.757
ICA-LiNGAM	0.433	0.532
PC	0.592	0.413
Uniform-CaRe	0.225	0.783
CaRe (Ours)	**0.154**	**0.871**

The results in Table 6 demonstrate the best performance of CaRe on both metrics, outperforming both the strongest SCD expert and Uniform-CaRe. These results show that CaRe remains effective on a previously unseen, semantically meaningful graph topology, providing additional evidence that its gains do not require recalling the ground-truth structures of benchmarks.

D.2 Robustness to Expert Removal

ICA-LiNGAM assumes continuous, non-Gaussian data, whereas the `bnlearn` datasets are discrete and low-cardinality. This assumption mismatch may explain part of ICA-LiNGAM's downweighting in Figure 3. To test whether CaRe's improvement is not merely caused by identifying this obvious assumption mismatch, we conduct a Remove-LiNGAM experiment, where both Uniform-CaRe and CaRe use the same reduced expert library without ICA-LiNGAM (w/o IL).

Table 7: Robustness to removing ICA-LiNGAM (w/o IL), averaged over sample sizes $n \in \{500, 1000, 5000, 10000\}$. Uniform-CaRe and CaRe use the same reduced expert library. Best Expert reports the best remaining individual SCD result for each dataset and metric separately. The best values are in bold.

Dataset	Alarm		Asia		Child		Insurance		Sachs	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
Best Expert	0.139	0.865	0.219	0.806	0.305	0.772	0.303	0.688	0.309	0.761
Uniform-CaRe (w/o IL)	0.131	0.872	0.359	0.650	0.260	0.755	0.295	0.687	0.485	0.667
CaRe (w/o IL)	**0.054**	**0.947**	**0.109**	**0.904**	**0.180**	**0.825**	**0.245**	**0.749**	**0.199**	**0.811**

As shown in Table 7, without ICA-LiNGAM, CaRe still substantially outperforms Uniform-CaRe and the best individual expert on both metrics across all datasets. These results indicate that CaRe's gains are not solely driven by identifying and downweighting ICA-LiNGAM's assumption mismatch.

D.3 Error Bars

We evaluate the robustness of CaRe with respect to LLM stochasticity on the five synthetic benchmark datasets. Since a full repeated-run evaluation over all SCD experts and bootstrap resamples is computationally expensive, we fix the expert outputs and bootstrap-based support scores, and repeat the LLM-guided reweighting stage using the same LLM backbone with temperature 1.0 to assess the stability of LLM's outputs. Table 8 reports the main values together with standard deviations across five repeated LLM runs, demonstrating that CaRe yields consistently stable results.

Table 8: Error bars of CaRe with respect to LLM stochasticity on synthetic benchmark datasets. The expert outputs and bootstrap-based support scores are fixed, while the LLM-guided reweighting is repeated five times with temperature 1.0. Results are reported as central value ± standard deviation.

Dataset	Alarm		Asia		Child		Insurance		Sachs	
Sample Size	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
500	0.065±0.011	0.937±0.012	0.125±0.032	0.891±0.036	0.140±0.014	0.863±0.016	0.250±0.014	0.748±0.016	0.324±0.041	0.680±0.043
1000	0.076±0.013	0.924±0.013	0.125±0.094	0.891±0.099	0.100±0.000	0.902±0.000	0.202±0.010	0.802±0.011	0.176±0.000	0.836±0.000
5000	0.033±0.000	0.968±0.000	0.062±0.031	0.945±0.028	0.240±0.028	0.768±0.028	0.231±0.053	0.765±0.058	0.118±0.059	0.891±0.054
10000	0.022±0.010	0.979±0.011	0.062±0.031	0.945±0.028	0.080±0.000	0.922±0.000	0.144±0.000	0.858±0.000	0.059±0.030	0.945±0.027
Avg	0.049±0.008	0.952±0.008	0.093±0.016	0.918±0.020	0.140±0.011	0.864±0.011	0.207±0.019	0.793±0.021	0.169±0.015	0.838±0.017

D.4 LLM Query Cost

We report the number of LLM queries used by CaRe and compare it with exhaustive LLM querying. For a dataset with $r = |V|$ variables, exhaustive pairwise querying for causal discovery requires

$$Q_{\text{full}}(r) = \binom{r}{2} = \frac{r(r-1)}{2} \tag{17}$$

queries if each unordered variable pair is queried once. This is a conservative reference because querying ordered pairs separately would require $r(r-1)$ queries. Given the number of LLM queries used by CaRe, denoted by Q_{CaRe}, we define the query reduction ratio as

$$\text{Reduction} = 1 - \frac{Q_{\text{CaRe}}}{Q_{\text{full}}}. \tag{18}$$

For synthetic datasets, we report the average query count over the four sample sizes. For real-world datasets, we report the fixed benchmark query count. As shown in Table 9, CaRe substantially reduces the number of LLM queries compared with exhaustive pairwise querying. Under the per-dataset averaged view reported in the table, CaRe uses 237 LLM queries in total across all benchmarks, compared with 1483 exhaustive pairwise queries, corresponding to an overall reduction of **84.0%**. The reduction is especially pronounced on larger benchmark networks such as *Alarm*, *Child*, and *Insurance*. The reduction is smaller on *Arctic-Sea** and *Sangiovese**, where the ground-truth structures are relatively dense or contain non-DAG relations, leading to more disputed cases. Nevertheless, CaRe still requires fewer queries than exhaustive pairwise LLM-based causal discovery.

Table 9: LLM query cost comparison. $|V|$ denotes the number of variables, and $|E|$ denotes the number of ground-truth adjacent edges. Q_{full} denotes the number of LLM queries over all possible pairs of variable, i.e., $\binom{|V|}{2}$ exhaustive pairwise queries. Q_{CaRe} denotes the total number of LLM queries used by our method. For synthetic datasets, CaRe query counts are averaged over sample sizes $n \in \{500, 1000, 5000, 10000\}$; real-world datasets are marked with $*$.

Dataset	Alarm	Asia	Child	Insurance	Sachs	Alzheimer*	Arctic-Sea*	Sangiovese*
$\|V\|$	37	8	20	27	11	9	12	14
$\|E\|$	46	8	25	52	17	16	30	51
Q_{full}	666	28	190	351	55	36	66	91
Q_{CaRe} **(Ours)**	**26.3**	**8.5**	**26.5**	**35.3**	**23.5**	**7.0**	**47.0**	**63.0**
Reduction	**96.1%**	**69.6%**	**86.1%**	**90.0%**	**57.3%**	**80.6%**	**28.8%**	**30.8%**

D.5 HIERARCHY ABLATION

The skeleton-v-structure-edge hierarchy is motivated by both the requirements of heterogeneous expert aggregation and the structural dependencies of causal graph identification. Different experts may have different reliability at different graph levels, and our expert library contains heterogeneous DAG, CPDAG, and PAG outputs. Decomposing these graphs into common relation families provides a unified aggregation interface, allowing expert opinions to be comparable across structural levels.

Note that directly aggregating causal relations can mix adjacency and orientation errors in a single decision. By first resolving whether an adjacency exists, then identifying v-structures, and finally orienting the remaining edges, CaRe narrows the candidate space and prevents later decisions from contradicting relations established at earlier stages. Under causal sufficiency and faithfulness, v-structures provide identifiable directional information that is absent from the skeleton. This ordering is also well established in classical methods such as PC (Spirtes et al., 2000), which first recover adjacencies, then identify v-structures, and finally orient the remaining edges using Meek rules.

To further examine the contribution of our hierarchy design, we compare three variants: **flat**, which directly aggregates edges without separating graph levels; **w/o vs**, which aggregates the skeleton and then directly resolves edge orientations; and **full**, which uses the complete skeleton-v-structure-edge hierarchy. We evaluate these variants for both Uniform-CaRe and CaRe.

Table 10: Hierarchy ablation results, averaged over sample sizes $n \in \{500, 1000, 5000, 10000\}$. The flat variant directly aggregates edges without separating graph levels; w/o vs removes the v-structure ensemble layer; full uses the original CaRe pipeline. The best values are in bold.

Dataset	Alarm		Asia		Child		Insurance		Sachs	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
Uniform-CaRe (flat)	0.465	0.663	0.453	0.595	0.545	0.428	0.495	0.546	0.588	0.208
Uniform-CaRe (w/o vs)	0.522	0.490	0.469	0.525	0.250	0.755	0.457	0.504	0.412	0.726
Uniform-CaRe (full)	0.147	0.865	0.281	0.732	0.260	0.755	0.300	0.683	0.456	0.691
CaRe (flat)	0.457	0.697	0.453	0.595	0.425	0.609	0.478	0.582	0.478	0.515
CaRe (w/o vs)	0.511	0.482	0.422	0.569	0.280	0.726	0.469	0.490	0.191	0.822
CaRe (full)	**0.049**	**0.952**	**0.093**	**0.918**	**0.140**	**0.864**	**0.207**	**0.793**	**0.169**	**0.838**

Table 10 reports the results averaged over four sample sizes. The full CaRe hierarchy achieves the best performance on both metrics across all five datasets compared with flat aggregation and the variant without the v-structure ensemble layer. These results support the effectiveness of our coarse-to-fine construction and the explicit treatment of v-structures in causal graph aggregation.

D.6 MARGIN SENSITIVITY

The main purpose of the margin-aware disputed set is to reduce unnecessary LLM calls by focusing intervention on uncertain relations. We conduct a margin sweep on *Alarm* at $n = 5000$ and compare it with reweighting all candidate relations.

Table 11: Margin sensitivity on Alarm at $n = 5000$. The best RNB and AUC values are in bold.

Strategy	Margin Coef. δ	LLM Queries	RNB (↓)	AUC (↑)
No LLM Reweighting	–	0	0.120	0.890
Margin-aware	0.25	10	0.076	0.924
Margin-aware (default)	0.5	21	**0.033**	**0.968**
Margin-aware	0.75	56	0.043	0.957
All Candidate Relations	1.0	105	**0.033**	**0.968**

As shown in Table 11, the default margin-aware strategy with $\delta = 0.5$ achieves the same best performance as reweighting all candidate relations while requiring substantially fewer LLM calls: 21 queries compared with 105 queries. Increasing the query budget does not consistently improve performance in this setting. These results support focusing LLM-guided reweighting on disputed relations to reduce unnecessary queries.

D.7 Expert Reliance Analysis

To examine how CaRe uses complementary experts, we quantify expert reliance rates using each expert's mean normalized pooling coefficient. These values describe soft aggregation weights rather than the frequency of selecting a single expert. Table 12 reports the reliance rates for both the full expert library and the reduced library without ICA-LiNGAM.

Table 12: Expert reliance rates (%), measured by mean normalized pooling coefficients. We report results for the full expert library and the reduced library without ICA-LiNGAM. The largest value in each column is in bold. Percentages may not sum to 100% due to rounding.

Expert Library (%)	Full				Without ICA-LiNGAM			
Expert	Skeleton	V-structure	Edge	Overall	Skeleton	V-structure	Edge	Overall
BOSS	18.44	18.08	16.27	17.76	20.80	20.90	17.88	19.45
FCI	17.08	16.90	18.50	17.33	17.08	19.76	21.39	19.83
GES	**20.98**	**21.12**	18.05	**20.35**	**24.60**	**24.37**	21.09	**22.89**
GRaSP	17.36	18.27	15.41	17.33	19.34	21.10	16.89	18.66
ICA-LiNGAM	8.74	8.68	11.30	9.32	–	–	–	–
PC	17.40	16.96	**20.48**	17.92	18.18	13.87	**22.76**	19.17

The reliance rates remain broadly distributed across experts while varying across graph levels: GES receives relatively greater weight for skeleton and v-structure aggregation, whereas PC and FCI receive increased weight for edge orientation. In the full expert library, GES receives 20.98% and 21.12% at the skeleton and v-structure levels, respectively, while approximately 79% of the pooled weight remains distributed across the other experts. Even after removing ICA-LiNGAM, the weights remain broadly distributed and the allocation changes across graph levels. These results show that CaRe does not collapse to a single globally dominant solver, but performs level-dependent soft aggregation over complementary experts.

D.8 Last-resort LLM Completion vs Deterministic DAG Extension

We further examine whether the final DAG-completion strategy substantially affects the performance of CaRe. In the main experiments, after hierarchical relation aggregation, confidence-ordered orientation, and Meek-rule closure, a small number of edges may still remain undirected. Our default implementation uses the LLM only as a last-resort completion aid, where the proposed orientation is accepted only if it preserves acyclicity and does not conflict with previously compelled orientations.

As a deterministic alternative, we replace this last-resort LLM completion with a deterministic DAG extension strategy. Specifically, after all resolved orientations have been inserted, we consider the remaining undirected edges in a fixed order and orient each edge according to a topologically compatible direction whenever it preserves acyclicity. This produces a DAG without LLM calls.

Table 13: Results between last-resort LLM-assisted completion and deterministic DAG extension.

Sample Size	Alarm				Asia				Child				Insurance				Sachs			
	LLM (Ours)		DAG		**LLM (Ours)**		DAG		**LLM (Ours)**		DAG		**LLM (Ours)**		DAG		**LLM (Ours)**		DAG	
Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
500	0.065	0.937	0.065	0.937	0.125	0.891	0.125	0.891	0.140	0.863	0.140	0.863	0.250	0.748	0.260	0.737	0.324	0.680	0.382	0.619
1000	0.076	0.924	0.098	0.902	0.125	0.891	0.188	0.819	0.100	0.902	0.100	0.902	0.202	0.802	0.288	0.700	0.176	0.836	0.176	0.836
5000	0.033	0.968	0.033	0.968	0.062	0.945	0.062	0.945	0.240	0.768	0.200	0.806	0.231	0.765	0.240	0.754	0.118	0.891	0.176	0.836
10000	0.022	0.979	0.022	0.979	0.062	0.945	0.062	0.945	0.080	0.922	0.160	0.845	0.144	0.858	0.135	0.868	0.059	0.945	0.142	0.859
Avg	**0.049**	**0.952**	0.054	0.947	**0.093**	**0.918**	0.109	0.900	**0.140**	**0.864**	0.150	0.854	**0.207**	**0.793**	0.231	0.765	**0.169**	**0.838**	0.219	0.788

Table 13 compares the two completion strategies on the five synthetic benchmark datasets across all sample sizes. Overall, the two strategies yield similar performance trends. The gap is relatively small, and deterministic completion even matches or slightly improves the LLM variant in several individual settings, such as *Child* at $n = 5000$ and *Insurance* at $n = 10000$. These results suggest that the main gains of CaRe are not primarily driven by the final completion step, but by the preceding hierarchical expert aggregation and LLM-guided expert reweighting. The last-resort LLM completion mainly provides a lightweight way to resolve residual orientations when expert evidence remains insufficient.

D.9 Limitation Analysis

Despite its strong empirical performance, CaRe has several limitations. First, LLM-guided reweighting depends on prompt quality: ambiguous variable names, incomplete feature descriptions, or poorly summarized dataset profiles can bias the LLM's judgment and lead to bad weight estimation. Second, CaRe reduces LLM query cost but still requires running experts over bootstraps, so its runtime depends on expert cost. Third, CaRe is limited by expert coverage as the LLM only reweights existing SCD outputs rather than proposing new relations, and relations missed by all experts cannot be recovered. By highlighting limitations, we aim to encourage further research to refine our method.

E Prompt and Response Examples

E.1 Prompt Templates

We provide prompt templates used by CaRe for LLM-guided expert reweighting (Figures 4–8). All prompts share the same system instruction and JSON response format, while the user prompt is specialized to the current graph level: skeleton, v-structure, or edge orientation. For readability, we show the templates with placeholders such as [Dataset], [Variable Description], [Dataset Profile], and [Expert Opinions]. In our implementation, these placeholders are instantiated with dataset names, variable descriptions, dataset statistics, expert information, and relation-specific expert opinions.

System Prompt:

You are a Causal Knowledge Expert with a sub-specialist interest in Causal Discovery. Your job is to read and analyze the provided context, then complete the required task accurately. Think carefully, then respond **ONLY** in the following JSON format:

```
{"analysis":{"[Expert Short Name]":"[brief analysis for this
expert, 1 sentence for Expert-Dataset Match, 1 sentence for Causal
Plausibility, and 1 sentence for Performance at the current graph
level.]","...":"..."},
"answer":[{"name":"[Expert Short Name]","weight":[integer in
0--10],"reason":"[brief justification, 1--2 sentences]"},{"...":"..."}]}
```

Replace the instruction values with your analysis and your answers, such that each expert's **Short Name** must appear exactly once in the `name` field.

Figure 4: System prompt and JSON response format for CaRe.

Shared Prompt Fields: Dataset Profile and Expert Opinions

Dataset Profile:

For discrete datasets:

Dataset: `[VALUE]` Samples, `[VALUE]` Features, Discrete Variables (`[TEXT]` Sample Size). Sample-to-feature ratio: `[VALUE]` (`[TEXT]`).
Cardinality: Min=`[VALUE]`, Max=`[VALUE]`, Median=`[VALUE]` (Type: `[TEXT]`).
Concentration: `[TEXT]` (Mean Mode Frequency `[VALUE]`%).
Dependence (Cramer's V): Mean V=`[VALUE]`, Max V=`[VALUE]`; Strong Associations are `[TEXT]`.
Entropy: `[TEXT]` (Mean normalized feature entropy).
Noise: `[TEXT]`.

For continuous datasets:

Dataset: `[VALUE]` Samples, `[VALUE]` Features, Continuous Variables (`[TEXT]` Sample Size). Sample-to-feature ratio: `[VALUE]` (`[TEXT]`).
Scaling: `[TEXT]` (std ratio `[VALUE]`x).
Distributions: `[TEXT]` Skewness, `[TEXT]` Tails, Features are `[TEXT]`.
Outliers: `[TEXT]` (MAD-based).
Dependence: Mean $|r|$=`[VALUE]`, Max $|r|$=`[VALUE]`; Multicollinearity `[TEXT]` Exists.
Nonlinearity: `[TEXT]` Signal (Pearson–Spearman Gap).
Noise: `[TEXT]`.

Expert Opinions:

Expert 1:
Short Name: `[EXPERT]`
Full Name: `[TEXT]`
Type: `[TEXT]`
Output: `[TEXT]`
Description: `[TEXT]`
Information: `[TEXT]`
Opinion: `[AGREE/DISAGREE]` for sk and vs queries, or `[OPTION A/B/C/D]` for eg queries.

Expert 2: ...

Expert N: ...

Figure 5: Shared dataset-profile and expert-opinion fields inserted into relation-specific prompts.

User Prompt: Skeleton Expert Reweighting

Consider the following **CONTEXT**:
Multiple Causal Discovery experts are recovering causal relationships among variables in the `[NAME]` dataset. At the **Skeleton** level, experts provide their opinions on whether the following two variables should have a **direct graph-level adjacency** in the causal graph skeleton (without considering the causal direction).

Variables and their descriptions:
(1) `[VARIABLE_1]: [DESCRIPTION_1]`
(2) `[VARIABLE_2]: [DESCRIPTION_2]`

Dataset Profile:
`[DATASET_PROFILE]`

Expert Opinions:
Out of `[VALUE]` experts, `[VALUE]` experts support a **direct graph-level adjacency** between `[VARIABLE_1]` and `[VARIABLE_2]. [EXPERT_OPINIONS]`

Given the following **TASK**:
(1) Assess whether each expert's opinion plausibly fits the dataset characteristics.
(2) Evaluate whether a **direct graph-level adjacency** is plausible between `[VARIABLE_1]` and `[VARIABLE_2]`, given variable semantics, domain constraints, and possible mediation by other observed variables.
(3) Finally, based on (1), (2), and the expert agreement pattern, assign a reliability weight in $[0, 1, 2, 3, \ldots, 10]$ to each expert for **THIS** relation.
(4) The score is based on the following rubric:
- $[0, 1, 2, 3, 4]$ Expert–Dataset Match.
- $[0, 1, 2, 3]$ Direct-adjacency Plausibility under the dataset's graph abstraction.
- $[0, 1, 2, 3]$ Performance of Expert at the **Skeleton** level.

(5) Use extreme weights such as 10 or 0 only when the evidence is overwhelming; otherwise prefer calibrated non-extreme weights.

Your answer is:

Figure 6: Prompt template for skeleton-level LLM-guided expert reweighting.

User Prompt: V-structure Expert Reweighting

Consider the following **CONTEXT**:
Multiple Causal Discovery experts are recovering causal relationships among variables in the `[NAME]` dataset. At the **V-structure** level, experts provide their opinions on whether the following three variables form a **direct unshielded collider** of the pattern (1) $\rightarrow$ (2) $\leftarrow$ (3). This requires (1) and (3) to be plausible direct parents of (2), and also requires (1) and (3) not to have a direct adjacency under the dataset's causal abstraction.

Variables and their descriptions:
(1) `[VARIABLE_1]: [DESCRIPTION_1]`
(2) `[VARIABLE_2]: [DESCRIPTION_2]`
(3) `[VARIABLE_3]: [DESCRIPTION_3]`

Dataset Profile:
`[DATASET_PROFILE]`

Expert Opinions:
Out of `[VALUE]` experts, `[VALUE]` experts support the **direct unshielded collider** `[VARIABLE_1]` $\rightarrow$ `[VARIABLE_2]` $\leftarrow$ `[VARIABLE_3]. [EXPERT_OPINIONS]`

Given the following **TASK**:
(1) Assess whether each expert's opinion plausibly fits the dataset characteristics.
(2) Evaluate whether `[VARIABLE_1]` and `[VARIABLE_3]` are plausible **independent parents** of `[VARIABLE_2]` under an unshielded collider structure.
(3) Finally, based on (1), (2), and the expert agreement pattern, assign a reliability weight in $[0, 1, 2, 3, \ldots, 10]$ to each expert for **THIS** relation.
(4) The score is based on the following rubric:
- $[0, 1, 2, 3, 4]$ Expert–Dataset Match.
- $[0, 1, 2, 3]$ Unshielded-collider Plausibility, including direct-parent plausibility and parent non-adjacency.
- $[0, 1, 2, 3]$ Performance of Expert at the **V-structure** level.

(5) Use extreme weights such as 10 or 0 only when the evidence is overwhelming; otherwise prefer calibrated non-extreme weights.

Your answer is:

Figure 7: Prompt template for v-structure-level LLM-guided expert reweighting.

User Prompt: Edge Orientation Expert Reweighting

Consider the following **CONTEXT**:
Multiple Causal Discovery experts are recovering causal relationships among variables in the `[NAME]` dataset. At the **Edge Orientation** level, the adjacency between the following two variables has already been considered, and experts provide their opinions on the most plausible orientation of the following two variables.

Variables and their descriptions:
(1) `[VARIABLE_1]: [DESCRIPTION_1]`
(2) `[VARIABLE_2]: [DESCRIPTION_2]`

Edge Orientation Options:
– OPTION A: the expert supports that changing `[VARIABLE_1]` causes a change in `[VARIABLE_2]`.
– OPTION B: the expert supports that changing `[VARIABLE_2]` causes a change in `[VARIABLE_1]`.
– OPTION C: the expert supports an undirected adjacency between `[VARIABLE_1]` and `[VARIABLE_2]`, meaning that the expert supports adjacency but abstains from deciding the orientation.
– OPTION D: the expert does not provide usable evidence for this adjacency or orientation.

Dataset Profile:
`[DATASET_PROFILE]`

Expert Opinions:
Out of `[VALUE]` experts:
– `[VALUE]` experts support OPTION A (`[VARIABLE_1]` → `[VARIABLE_2]`).
– `[VALUE]` experts support OPTION B (`[VARIABLE_2]` → `[VARIABLE_1]`).
– `[VALUE]` experts support an undirected adjacency or abstain from orientation. `[EXPERT_OPINIONS]`

Given the following **TASK**:
(1) Assess whether each expert's opinion plausibly fits the dataset characteristics.
(2) Evaluate which experts provide reliable evidence for orienting this already-considered adjacency between `[VARIABLE_1]` and `[VARIABLE_2]`, given variable semantics and domain constraints.
(3) Pay special attention to whether each expert is suitable for the **Edge Orientation** level.
(4) Treat OPTION C as orientation abstention, not as directional evidence for either OPTION A or OPTION B.
(5) Finally, assign a reliability weight in $[0, 1, 2, 3, \ldots, 10]$ to each expert for **THIS** relation.
(6) The score is based on the following rubric:
– $[0, 1, 2, 3, 4]$ Expert–Dataset Match.
– $[0, 1, 2, 3]$ Directional Causal Plausibility.
– $[0, 1, 2, 3]$ Performance of Expert at the **Edge Orientation** level.
(7) Use extreme weights such as 10 or 0 only when the evidence is overwhelming; otherwise prefer calibrated non-extreme weights.

Your answer is:

Figure 8: Prompt template for edge-orientation-level LLM-guided expert reweighting.

E.2 Example LLM Response and Analysis

We provide two examples to show how CaRe uses the LLM as a meta-referee for expert reweighting.

Expert Reweighting for Skeleton. Figure 9 shows a skeleton-level example on the *Alarm* dataset for the candidate adjacency between `VENTLUNG` and `KINKEDTUBE`. Although four out of six experts support the adjacency, the LLM does not simply perform majority voting. Instead, it assigns high weights to BOSS, GES, GRaSP, and PC because their assumptions and inference mechanisms are compatible with the large, low-cardinality discrete dataset, and because their AGREE opinions are consistent with the direct mechanical relation between tube obstruction and effective lung ventilation. In contrast, LiNGAM is assigned a minimal weight because its continuous linear non-Gaussian assumptions are poorly matched to the discrete *Alarm* variables. FCI receives only a moderate weight: although its conditional-independence testing is type-compatible, its DISAGREE opinion conflicts with the relation-level causal plausibility. This example demonstrates how CaRe combines dataset compatibility and causal plausibility to produce calibrated relation-specific expert weights.

Expert Reweighting for Edge Orientation. Figure 10 presents an edge orientation example on the *Sachs* dataset for the already-considered adjacency between `PKA` and `Jnk`. Three experts support `PKA` → `Jnk`, while the remaining three abstain from orientation. The LLM assigns high weights to FCI and PC because their chi-square conditional-independence tests are type-compatible with the large discrete *Sachs* setting and their OPTION A opinions provide usable directional evidence. LiNGAM also supports the same direction, but receives only a moderate weight because its continuous linear

non-Gaussian assumptions are less compatible with the discretized low-cardinality measurements. Meanwhile, BOSS, GES, and GRaSP receive lower weights because their OPTION C outputs indicate orientation abstention rather than directional support. This example highlights that CaRe can distinguish experts even when they appear to support the same direction, and that undirected outputs are treated as orientation uncertainty rather than as evidence for either causal direction.

Example 1: Expert Reweighting for Skeleton on Alarm

Relation: `VENTLUNG – KINKEDTUBE`

Task: Assign relation-specific reliability weights to each expert for whether `VENTLUNG` and `KINKEDTUBE` should have a direct graph-level adjacency in the skeleton.

Variable Semantics:
`VENTLUNG`: effective ventilation status of the lungs through the airway and ventilator circuit.
`KINKEDTUBE`: kinked or obstructed endotracheal or breathing tube that restricts airflow.

Expert Opinions:
Out of 6 experts, 4 experts support a direct graph-level adjacency between `VENTLUNG` and `KINKEDTUBE`.

Expert	BOSS	FCI	GES	GRaSP	LiNGAM	PC
Opinion	AGREE	DISAGREE	AGREE	AGREE	DISAGREE	AGREE

LLM-assigned Weights:

Expert	BOSS	FCI	GES	GRaSP	LiNGAM	PC
Weight	9	4	9	8	1	9

LLM Analyses:
`BOSS`: BOSS is well matched to the large low-cardinality discrete ALARM data with a type-compatible BDeu score; its AGREE opinion is highly plausible because tube obstruction directly affects effective lung ventilation, though permutation/order equivalence makes its skeleton evidence slightly less definitive.
`FCI`: FCI is reasonably matched through chi-square CI tests on large discrete data, but its DISAGREE opinion conflicts with the direct mechanical relation between KINKEDTUBE and VENTLUNG and may reflect conservative conditioning or PAG extraction uncertainty.
`GES`: GES is strongly matched because BDeu scoring fits the discrete data and the sample size supports stable score comparisons; its AGREE opinion is strongly plausible and carries high skeleton-level reliability.
`GRaSP`: GRaSP is well matched to the large discrete setting with BDeu-based permutation search, and its AGREE opinion fits the direct mechanical effect from KINKEDTUBE to VENTLUNG, though ordering-based searches can admit near-equivalent sparse explanations.
`LiNGAM`: LiNGAM is poorly matched because the variables are discrete and low-cardinality rather than continuous linear non-Gaussian measurements; its DISAGREE opinion conflicts with the direct mechanism, making its skeleton output unreliable here.
`PC`: PC is strongly matched because the sample size is large and chi-square CI tests are appropriate for low-cardinality discrete variables; its AGREE opinion is plausible and supported by strong skeleton-level evidence.

Final Weight Justifications:
`BOSS (9)`: Type-compatible BDeu scoring and plausible AGREE evidence support a high weight, with slight caution from permutation/order uncertainty.
`FCI (4)`: Although methodologically reasonable for discrete data, its DISAGREE vote conflicts with a direct mechanical relation.
`GES (9)`: Strong discrete-data compatibility and strong skeleton reliability support its plausible AGREE vote.
`GRaSP (8)`: Good data compatibility and plausible AGREE evidence support a high weight, with moderate caution from permutation-based uncertainty.
`LiNGAM (1)`: Severe assumption mismatch with discrete ALARM variables and an implausible DISAGREE vote justify minimal reliability.
`PC (9)`: Type-compatible CI testing and strong skeleton-level reliability support its plausible AGREE vote.

Figure 9: LLM response for skeleton-level expert reweighting on the Alarm dataset.

Example 2: Expert Reweighting for Edge Orientation on Sachs

Relation: PKA → Jnk

Task: Assign relation-specific reliability weights to each expert for orienting the already-considered adjacency between PKA and Jnk.

Variable Semantics:

PKA: protein kinase A activity level, a cAMP-dependent kinase readout involved in broad intracellular signaling regulation.

Jnk: phosphorylated JNK or c-Jun N-terminal kinase activity level, a stress-activated MAP kinase readout involved in inflammatory and apoptotic signaling.

Expert Opinions:

Out of 6 experts, 3 experts support PKA → Jnk, 0 experts support Jnk → PKA, and 3 experts abstain from orientation. Here, OPTION A denotes PKA → Jnk, and OPTION C denotes orientation abstention.

Expert	BOSS	FCI	GES	GRaSP	LiNGAM	PC
Opinion	OPTION C	OPTION A	OPTION C	OPTION C	OPTION A	OPTION A

LLM-assigned Weights:

Expert	BOSS	FCI	GES	GRaSP	LiNGAM	PC
Weight	4	8	4	3	5	8

LLM Analyses:

BOSS: BOSS is well matched to the large low-cardinality discrete SACHS data with a type-compatible BDeu score, but OPTION C is orientation abstention and provides no usable directional evidence for either PKA → Jnk or the reverse direction.

FCI: FCI is well matched through chi-square CI tests on large discrete data and can be useful when hidden confounding is plausible in biological signaling; its OPTION A opinion supports the biologically plausible PKA → Jnk direction, though PAG endpoint uncertainty and circle-to-tail conversion limit certainty.

GES: GES is well matched because BDeu scoring fits the discrete variables and the sample size supports stable score comparisons, but its OPTION C output is an undirected CPDAG edge and therefore provides little directional evidence.

GRaSP: GRaSP is reasonably matched to the large discrete dataset with BDeu scoring, but it abstains from orientation and its edge directions are permutation-induced when present, making its evidence weak here.

LiNGAM: LiNGAM supports the biologically plausible PKA → Jnk direction, but it is poorly matched because the SACHS variables are discretized low-cardinality measurements rather than continuous linear non-Gaussian.

PC: PC is well matched because the sample size is large and chi-square CI tests are appropriate for low-cardinality discrete variables; its OPTION A opinion supports the biologically plausible direction, though causal sufficiency may be less secure in cellular signaling systems.

Final Weight Justifications:

BOSS (4): Good discrete-data compatibility, but OPTION C is only orientation abstention and gives weak edge-orientation evidence.

FCI (8): Type-compatible CI testing and support for the plausible PKA → Jnk direction justify a high weight, with caution from PAG endpoint uncertainty.

GES (4): Strong data compatibility, but its undirected CPDAG output indicates orientation uncertainty rather than directional support.

GRaSP (3): Reasonable dataset compatibility, but orientation abstention and permutation-based uncertainty provide little usable directional evidence.

LiNGAM (5): The proposed direction is biologically plausible, but assumption mismatch with discretized SACHS variables limits reliability.

PC (8): Type-compatible CI testing and plausible OPTION A evidence support a high weight, with caution about possible causal-sufficiency violations.

Figure 10: LLM response for edge orientation-level expert reweighting on the Sachs dataset.

F Full Results

Table 14 provides the complete causal discovery results on the synthetic benchmark datasets across sample sizes $n \in \{500, 1000, 5000, 10000\}$. We report per-sample size results for these datasets as only the synthetic benchmarks are evaluated under multiple sample-size regimes. The table provides

a fine-grained view of the averaged results reported in Table 1, showing that CaRe achieves strong overall performance across different sample sizes.

Table 14: Complete causal discovery results on synthetic datasets across different sample sizes. The best values are in **bold**, and the second-best are underlined.

Sample Size	Method	Alarm		Asia		Child		Insurance		Sachs	
	Metric	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)	RNB (↓)	AUC (↑)
500	BOSS	0.380	0.690	0.500	0.586	0.440	0.677	0.327	0.657	0.559	0.550
	FCI	0.337	0.659	0.375	0.615	0.440	0.574	0.413	0.541	0.441	0.601
	GES	0.207	0.791	0.188	0.834	0.440	0.677	0.327	0.654	0.647	0.363
	GRaSP	0.217	0.782	0.500	0.586	0.440	0.659	0.481	0.511	0.559	0.550
	ICA-LiNGAM	0.739	0.428	0.250	0.766	0.500	0.562	0.673	0.363	0.588	0.422
	PC	0.315	0.670	0.375	0.615	0.380	0.625	0.490	0.454	0.588	0.483
	Uniform CaRe	0.228	0.801	0.312	0.693	0.340	0.665	0.375	0.590	0.529	0.508
	CaRe (Ours)	**0.065**	**0.937**	**0.125**	**0.891**	**0.140**	**0.863**	**0.250**	**0.748**	**0.324**	**0.680**
1000	BOSS	0.337	0.727	0.625	0.605	0.440	0.659	0.404	0.625	0.500	0.725
	FCI	0.239	0.770	0.562	0.441	0.180	0.823	0.308	0.675	0.265	0.827
	GES	0.152	0.854	0.312	0.724	0.340	0.734	0.375	0.601	0.559	0.523
	GRaSP	0.217	0.801	0.500	0.586	0.400	0.692	0.433	0.550	0.529	0.587
	ICA-LiNGAM	0.446	0.605	0.625	0.414	0.680	0.415	0.558	0.411	0.618	0.442
	PC	0.239	0.760	0.562	0.441	0.180	0.819	0.327	0.652	0.618	0.472
	Uniform CaRe	0.141	0.863	0.438	0.567	0.160	0.841	0.337	0.642	0.294	0.757
	CaRe (Ours)	**0.076**	**0.924**	**0.125**	**0.891**	**0.100**	**0.902**	**0.202**	**0.802**	**0.176**	**0.836**
5000	BOSS	0.359	0.763	0.500	0.695	**0.240**	**0.838**	0.375	0.684	0.500	0.750
	FCI	0.087	0.915	0.375	0.648	0.400	0.659	0.356	0.648	0.235	0.823
	GES	0.098	0.907	0.188	0.834	**0.240**	**0.838**	0.327	0.669	0.500	0.750
	GRaSP	0.391	0.713	0.562	0.563	0.280	0.806	0.317	0.697	0.500	0.750
	ICA-LiNGAM	0.522	0.617	0.250	0.766	0.700	0.433	0.865	0.267	0.647	0.398
	PC	0.109	0.895	0.375	0.648	0.400	0.643	0.385	0.610	0.500	0.573
	Uniform CaRe	0.120	0.890	0.188	0.834	0.320	0.715	0.298	0.694	0.500	0.750
	CaRe (Ours)	**0.033**	**0.968**	**0.062**	**0.945**	**0.240**	0.768	**0.231**	**0.765**	**0.118**	**0.891**
10000	BOSS	0.446	0.674	0.500	0.695	0.260	0.815	0.298	0.715	0.500	0.750
	FCI	0.185	0.834	0.375	0.648	0.240	0.787	0.269	0.729	0.294	0.795
	GES	0.098	0.907	0.188	0.834	0.240	0.838	0.183	0.829	0.500	0.750
	GRaSP	0.402	0.653	1.000	0.255	0.400	0.726	0.490	0.619	0.500	0.750
	ICA-LiNGAM	0.598	0.561	0.438	0.567	0.860	0.350	0.558	0.462	0.529	0.508
	PC	0.098	0.907	0.375	0.648	0.260	0.763	0.308	0.682	0.500	0.573
	Uniform CaRe	0.098	0.907	0.188	0.834	0.220	0.800	0.192	0.807	0.500	0.750
	CaRe (Ours)	**0.022**	**0.979**	**0.062**	**0.945**	**0.080**	**0.922**	**0.144**	**0.858**	**0.059**	**0.945**
Avg	BOSS	0.381	0.713	0.531	0.645	0.345	0.747	0.351	0.670	0.515	0.694
	FCI	0.212	0.795	0.422	0.588	0.315	0.711	0.337	0.648	0.309	0.761
	GES	0.139	0.865	0.219	0.806	0.315	0.772	0.303	0.688	0.551	0.597
	GRaSP	0.307	0.737	0.641	0.497	0.380	0.721	0.430	0.594	0.522	0.659
	ICA-LiNGAM	0.576	0.553	0.391	0.628	0.685	0.440	0.663	0.376	0.596	0.443
	PC	0.190	0.808	0.422	0.588	0.305	0.712	0.378	0.600	0.551	0.525
	Uniform CaRe	0.147	0.865	0.281	0.732	0.260	0.755	0.300	0.683	0.456	0.691
	CaRe (Ours)	**0.049**	**0.952**	**0.093**	**0.918**	**0.140**	**0.864**	**0.207**	**0.793**	**0.169**	**0.838**

G Additional Visualizations

G.1 Candidate Coverage Visualization

Figure 11 visualizes the causal graphs recovered by different SCD experts on the Asia dataset, together with the final graph produced by CaRe and the ground-truth graph. This example illustrates the candidate-coverage limitation of CaRe. Since CaRe is designed as an ensemble and reweighting framework, it only aggregates and calibrates relations proposed by the experts, rather than allowing the LLM to freely introduce new causal relations. Consequently, if a true relation is missed by all experts, then it cannot be recovered by CaRe. In the Asia example, the ground-truth graph contains the edge `asia` → `tub`. However, none of the SCD experts shown in Figure 11 identifies this relation. As a result, this edge is absent from the expert candidate pool and is therefore also absent from the final CaRe graph. This behavior reflects a deliberate design trade-off: CaRe avoids LLM-generated hallucination edges and keeps the final graph grounded in data-driven SCD outputs, but its recall is bounded by the coverage of the expert's candidate set.

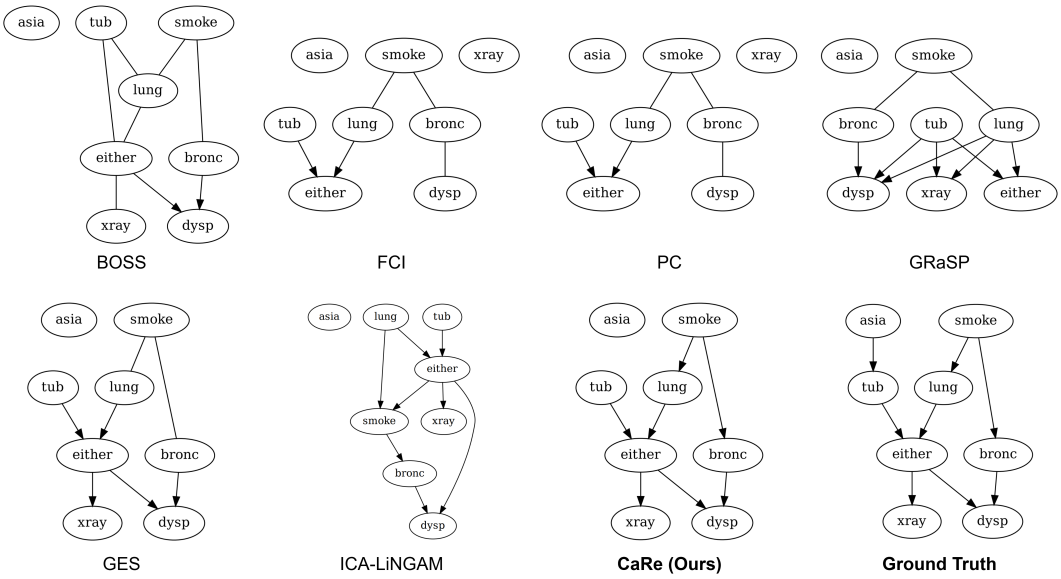

Figure 11: Candidate coverage visualization on the Asia dataset. The figure shows causal graphs recovered by individual SCD experts, the final graph produced by CaRe, and the ground-truth graph. Since the true edge `asia` → `tub` is missed by all experts, it is not included in the CaRe's graph.

G.2 Additional Weight-performance Alignment

We provide additional visualizations of the weight-performance alignment between LLM-assigned expert weights and empirical expert performance. Figures 12–16 further illustrate that CaRe's LLM-guided reweighting tends to assign higher weights to experts that are empirically more reliable.

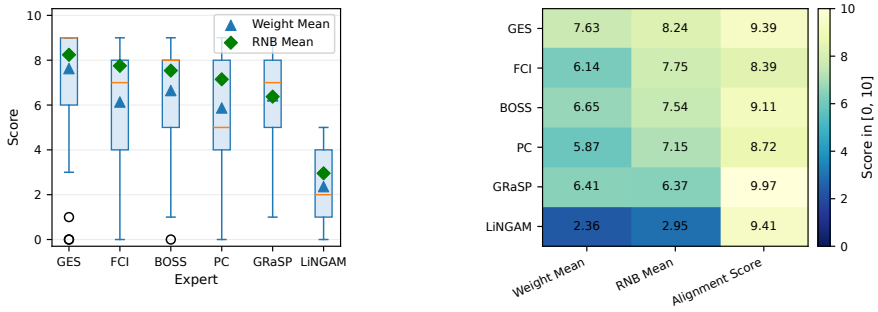

Figure 12: Weight-performance alignment on the Insurance dataset under the RNB metric.

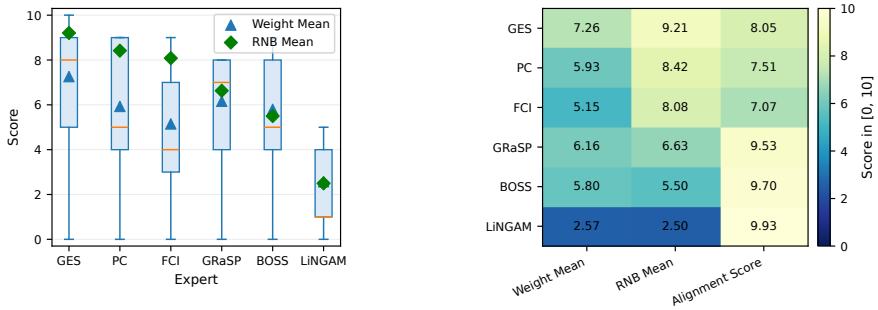

Figure 13: Weight-performance alignment on the Alarm dataset under the RNB metric.

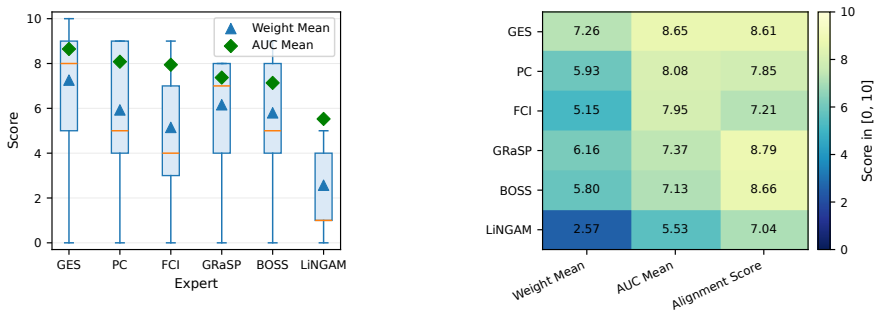

Figure 14: Weight-performance alignment on the Alarm dataset under the AUC metric.

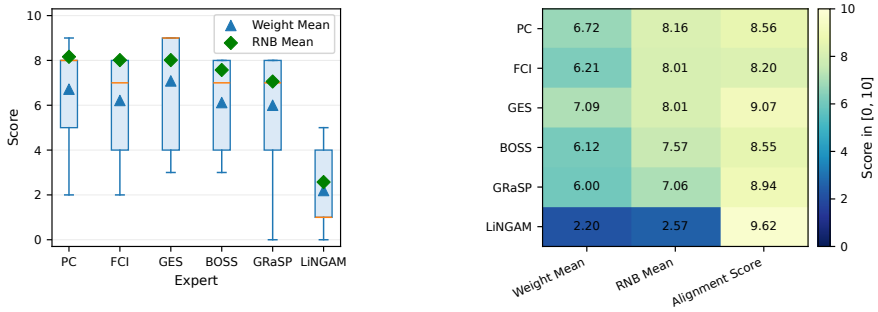

Figure 15: Weight-performance alignment on the Child dataset under the RNB metric.

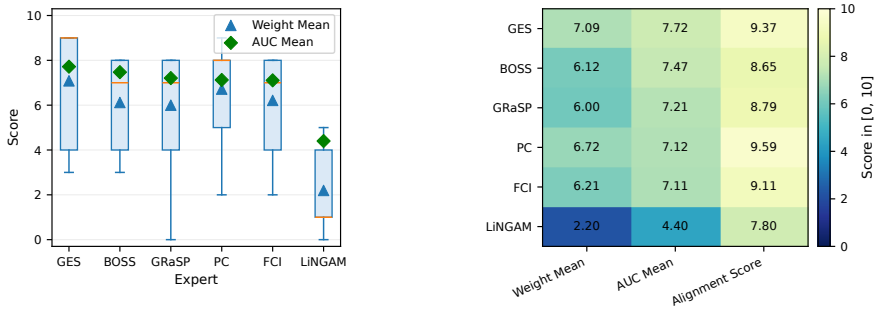

Figure 16: Weight-performance alignment on the Child dataset under the AUC metric.