NEURO-SYMBOLIC RULE DISCOVERY: EMPOWERING LLMS WITH CAUSALITY FOR VEHICLE DIAGNOSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Defining Boolean logic for vehicle fault detection of error patterns (EPs) is a manual, error-prone bottleneck process in automotive safety. Standard LLMs struggle to automate this task, as they prioritize semantic plausibility over logical necessity. We propose **CAREP**, a framework that empowers LLMs with causal discovery to extract strict diagnostic rules from noisy high-dimensional event sequences. Instead of relying on semantic correlations, CAREP provides the LLM with a grounded set of causal drivers (excitatory) and constraints (inhibitory). This enables the automated synthesis of accurate, human-readable rules alongside reasoning traces. On a real-world dataset of 29,100 unique codes, CAREP achieves superior rule reconstruction accuracy compared to standard RAG baselines.

1 INTRODUCTION

Modern intelligent vehicles generate asynchronous sequences of discrete events known as *Diagnostic Trouble Codes* (DTCs) Pirasteh et al. (2019). To monitor fleet health, automotive manufacturers define *Error Patterns* (EPs)—strict Boolean formulas that characterize critical system failures Math et al. (2025a). For instance, a rule might be defined as $ep_1 = dtc_1 \wedge dtc_2 \wedge (\neg dtc_3 \vee dtc_4)$, triggering only when specific DTCs combinations are present or not in a sequence. These rules are central to vehicle safety but are currently handcrafted by domain experts—a manual, error-prone process that cannot scale to the increasing complexity of modern architectures.

Automating this discovery requires bridging two distinct fields: causal discovery and semantic reasoning. While EPs are fundamentally causal, traditional causal discovery methods Gong et al. (2024) struggle with high-dimensional event sequences (thousands of unique event types Math et al. (2025b)) and often fail to distinguish between *excitatory* (positive) and *inhibitory* (negative) effects required for Boolean logic. Conversely, Large Language Models (LLMs) offer powerful semantic processing OpenAI (2023), in-context learning capabilities Dong et al. (2024) and an actionable interpretability that causal graphs alone cannot provide, especially in high-dimensional settings. However, as noted by Mirzadeh et al. (2025), LLMs struggle to infer strict logical constraints from noisy data, often hallucinating spurious correlations when the context window is large.

To resolve this dichotomy, we introduce **CAREP** (Causal Automated Reasoning for Error Patterns), a framework that grounds the semantic reasoning of LLMs in structural causal evidence. CAREP employs a *causal discovery tool* which estimates a summary causal graph Assaad et al. (2022) with signed causal indicators (excitatory vs. inhibitory, co-occurrences) from a batch of event sequences. This sparse "causal knowledge" is then enriched by semantic information (e.g., model ranges, descriptions), before being passed to a diagnostic agent. By constraining the LLM's generation with causal evidence, the agent synthesizes accurate Boolean rules accompanied by transparent reasoning traces Yao et al. (2023). Evaluated on a massive real-world dataset comprising 29,100 unique DTCs and 474 ground-truth rules, CAREP significantly outperforms unconstrained LLM baselines, demonstrating that empowering generative models with causal structure is a prerequisite for reliable automation in safety-critical domains.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

Data Setting. We consider a dataset $\mathcal{D} = \{(S^{(i)}, \mathbf{m}^{(i)}, \mathbf{y}^{(i)})\}_{i=1}^{m}$ consisting of m i.i.d. vehicle records. Each record comprises: (1) an asynchronous sequence of events $S^{(i)} = \{(t_1, x_1), \ldots, (t_{L_i}, x_{L_i})\}$ where x_t belongs to a finite vocabulary of discrete DTCs $\mathcal{X}$; (2) static vehicle metadata $\mathbf{m}^{(i)} \in \mathcal{M}$ (e.g., model range); and (3) a multi-hot label vector $\mathbf{y}^{(i)} \in \{0,1\}^K$ indicating the presence of K Error Patterns (EPs). We assume access to an auxiliary knowledge base $\mathcal{K}$ containing natural language descriptions for all DTCs, vehicle attributes, and EPs: $\mathcal{K} = \{(z, \text{desc}(z)) | z \in \mathcal{X} \cup \mathcal{M} \cup \mathcal{Y}\}$.

Symbolic Discovery Objective. We posit that for each EP k, there exists a **global** ground-truth Boolean function $f_k^* : 2^{\mathcal{X}} \to \{0,1\}$ (e.g., $x_a \wedge (x_b \vee \neg x_c)$) that maps the set of unique DTCs present in a sequence to the EP status, independent of the metadata $\mathbf{m}$. Our goal is to approximate f_k^* (composed strictly of atoms from $\mathcal{X}$) with a candidate rule $\hat{f}_k$.

2.2 CAUSAL DISCOVERY IN SINGLE STREAMS

The causal discovery follows two steps (1) We identify if a specific DTC (event) x_t is a cause of an EP (outcome) y by measuring their Conditional Mutual Information (CMI) conditioned on a past trajectory $x_{<t}$ for all sequential steps t and provide a local directed acyclic graph (DAG) per sequence as $\mathcal{G}^{(i)}$ (2) We aggregate these local evidences based on the edge frequencies to form a global consensus graph per outcome y_j as $\mathcal{G}_y^*$ with causal indicators. The pipeline structure can be seen in Fig. 3 (Appendix A.2.2).

Learned Generative Priors. Estimating causal dependencies directly from raw, high-dimensional logs is statistically fragile due to data sparsity and scalability to thousands of different event types $\mathcal{X}$. To overcome this, we leverage two autoregressive Transformers Tf_x and Tf_y Math et al. (2025a) pretrained on a massive auxiliary corpus $\mathcal{D}_{pre}$ drawn from the same joint distribution $P(X, Y)$ as $\mathcal{D}$. They model the probability of the next event X_t (DTCs) and the next outcomes Y (error patterns status y) given the history $x_{<t} = \{x_1, \cdots, x_{t-1}\}$:

$$P_\theta(X_t | S_{<t}) = \text{Softmax}(\text{Tf}_x(x_{<t})) \quad P_\theta(Y | S_{\leq t}) = \text{Sigmoid}(\text{Tf}_y(x_{\leq t})) \tag{1}$$

These priors provide a parallelized density estimation, allowing us to estimate entropic measures that would be computationally intractable on the finite observed dataset $\mathcal{D}$ at inference.

Quantifying Causal Strength. Event apparitions leading to outcomes $\boldsymbol{y}^{(i)}$ are modeled using a sequential Bayesian Network (Def. 1, Fig. A.1.1, Appendix A.1.1). Specifically, we would like to access the additional information event X_t provides about the outcome Y_j when we already know the past trajectory of events $x_{<t}$. Formally, this equates to the Information Gain Quinlan (1986):

$$I_G(Y, X_t | x_{<t}) \triangleq D_{KL}(P(Y | x_t, x_{<t}) || P(Y | x_{<t})) = H(Y | x_{<t}) - H(Y | x_t, x_{<t}) \tag{2}$$

By taking the expected value over past trajectories $x_{<t}$ and event x_t, this is equivalent to the CMI:

$$I(Y, X_t | X_{<t}) \triangleq \mathbb{E}_{x_t, x_{<t}}[I_G(Y, x_t | x_{<t})] = H(Y | X_{<t}) - H(Y | X_t, X_{<t}) \tag{3}$$

Hence, we can access conditional independence (Def. 2) of $X_t \perp Y | X_{<t}$ to construct the local causal graph $\mathcal{G}^{(i)}$ (Math et al. (2025b) Theorem 1).

Determining Inhibitory or Excitatory Effects. While CMI quantifies the strength of the relationship, Boolean rule synthesis requires distinguishing between *excitatory* (AND) and *inhibitory* (NOT) factors. We infer the sign of the relationship by comparing the conditional probabilities using the autoregressive models (equation 1) and infer the Eells measure Eells (1991):

$$\mathcal{C}_{x \to y} = \mathbb{E}_{x_{<t}} \left[P(Y | X_{<t}) - P(Y | X_t, X_{<t}) \right] \tag{4}$$

If $\mathcal{C}_{x \to y} > 0$, x is a positive cause (Excitatory); if $\mathcal{C}_{x \to y} < 0$, x is a negative cause (Inhibitory).

Point-wise Mutual Information. To distinguish between AND and OR operators, we compute the empirical Point-wise Mutual Information (PMI) between pairs of causal candidates x_a, x_b. Let $\hat{p}(x \mid y)$ denote the empirical probability of event x occurring in a sequence where EP y is active, and $\hat{p}(x_a, x_b \mid y)$ their joint probability over the support set m_y. The conditional PMI is defined as:

$$\text{PMI}(x_a, x_b \mid y) = \log \frac{\hat{p}(x_a, x_b \mid y)}{\hat{p}(x_a \mid y)\hat{p}(x_b \mid y)} \tag{5}$$

A strong positive PMI indicates that x_a and x_b co-occur more frequently than chance, strongly suggesting a conjunctive relationship ($x_a \wedge x_b$). Conversely, a strong negative PMI indicates mutual exclusivity, suggesting they are independent triggers linked by a disjunction ($x_a \vee x_b$).

2.2.1 Global Consensus via Adaptive Aggregation

While single-stream causal discovery Math & Lienhart (2026b) yields local graphs $\mathcal{G}^{(i)}$, reliable rule synthesis requires a global consensus graph $\mathcal{G}^*$ that is robust to statistical noise. We aggregate local evidence by computing the empirical edge frequency $\hat{\pi}_{X \to Y} = \frac{1}{m_y} \sum_{i=1}^{m_y} \mathbb{1}[(X \to Y) \in \mathcal{G}^{(i)}]$ across all sequences where outcome Y is active, with m_y being the support of y. However, as real-world data exhibits a severe long-tail distribution: a fixed threshold for edge inclusion would either purge valid links for rare EPs (low recall) or admit spurious noise for common EPs (low precision).

To resolve this, we employ an adaptive thresholding strategy. We define a label-specific threshold $\tau_y(m_y)$ as a logistic decay function (equation 6) of the sample support m_y. This mechanism acts as a data-driven regularizer: it enforces strict inclusion criteria ($\tau \approx \tau_{max}$) for rare patterns to reject variance-induced noise, while relaxing criteria ($\tau \approx \tau_{min}$) for common patterns.

Finally, we compile the **Causal Knowledge** per label $\mathcal{K}_{causal}(y) = (\mathcal{V}_{skel}, \mathcal{E}_{pmi})$ with:

$$\mathcal{V}_{skel}(y) = \left\{ (x, \tilde{\mathcal{C}}_{x \to y}) \mid x \in \mathcal{X}, : \hat{\pi}_{x \to y} > \tau_y(m_y) \right\}, \mathcal{E}_{pmi}(y) = \{(x_i, x_j, \text{PMI}_{i,j}) \mid x_i, x_j \in \mathcal{V}_{skel}\}$$

where $\tilde{\mathcal{C}}_{x \to y}$ denotes the averaged Eells measure (excitatory/inhibitory) and $\mathcal{E}_{pmi}$ represents the set of pairwise co-occurrences (triplets). This composite structure provides both the necessary components (the nodes) and the logic (NOT AND/OR) required for rule synthesis.

2.3 Logical Orchestration and Semantic Grounding

To bridge the gap between the extracted causal knowledge and the target Boolean formula $\hat{f}_k$, we employ a hierarchical reasoning protocol involving a **Diagnostic Agent**. This agent serves as the rule synthesizer by implementing a ReAct paradigm Yao et al. (2023) and queries the necessary tools as:

1. **Candidate Anchoring:** The agent first analyzes the causal knowledge $\mathcal{K}_{causal}$ with the *Causes* and their indicators, such as the *CMI* and Eells measure $\mathcal{C}_{X \to Y}$ to identify the "dominant" failure modes, distinguishing primary triggers from secondary symptoms.
2. **Structural Inference:** Logical operators between the candidates are then inferred from the *PMI* to determine connectivity: high positive PMI suggests conjunctions (AND), while strong negative PMI between candidates suggests mutually exclusive disjunctions (OR). Simultaneously, negative $\mathcal{C}_{X \to Y}$ values are explicitly mapped to inhibitory constraints.
3. **Semantic & Empirical Verification:** Finally, the agent cross-references these structural hypotheses with a Retrieval-Augmented Generation (RAG) pipeline. We utilize a pretrained embedding model (Amazon Titan V2) to encode the knowledge base $\mathcal{K}$ into a dense vector space (grouping functionally similar codes, and known EP rules, vehicle metadata) and validate the logic against the observed sequences to ensure the rule holds.

3 Experiments

Dataset & Protocol. We validate CAREP on a large-scale industrial fleet dataset comprising $m = 300,000$ sequences with on average 100 ± 35 DTCs per sequence defined over a vocabulary of $|\mathcal{X}| = 29,100$ unique DTCs and $|\mathcal{Y}| = 474$ target Error Patterns (EPs). We randomly mask 20 known EPs (with support $n \geq 100$) and task the system with rediscovering their Boolean formulas solely from the event sequences and descriptions. We report results averaged across 5 random folds. A complete evaluation protocol and data description can be found respectively in Appendix A.3 and A.3.4. We control that the generative prior Tf_y achieved at least 0.9 of F1-score on the same test-set.

Baselines. We compare CAREP with (Claude 3.7 Sonnet) against LLMs: **Claude 3.5 Sonnet**, **Claude 3.7 Sonnet**, **GPT-4.1/4.1-mini**. All baselines operate in a **semantic-only RAG** setting: they receive the observations $\{S^{(i)}, \boldsymbol{m}^{(i)}, y^{(i)}\}$ but lack the Causal Knowledge (only step 3. Semantic & Empirical Validation). To quantify the specific contribution of the Causal Module, we also include an ablation **CAREP w/o Causal**, which removes the causal indicators $(\mathcal{C}_{x\to y}, \mathcal{E}_{pmi})$ and a comparison with the causal discovery pipeline alone (Appendix A.2.2). We give as input a json format (Appendix 4) in the prompt and ask the LLM to generate 5 plausible rules.

Evaluation Metrics. We define two rigorous metrics (Appendix A.3.5) to evaluate the synthesized rules (1) Structural (Atom Retrieval): Measures if the rule contains the correct variables, ignoring logic. We treat the rule as a bag-of-atoms and compute classification metrics against the ground truth set. (2) Semantic (Truth-Table Consistency): Measures logical equivalence. We generate the truth tables $\mathcal{T}(f^*)$ and $\mathcal{T}(\hat{f})$ by enumerating all 2^k binary assignments of the union of atoms. We report classification metrics of the output bit-vector using the 5 plausible rules to form top-k metrics.

3.1 Results: Mitigating Hallucination via Causality

Structural Precision (The "Hallucination" Gap). Standard LLMs struggle to isolate the correct root causes from the high-dimensional vocabulary. GPT-4o-mini achieves a negligible *Precision@1* of 0.10, and even Claude 3.7 peaks at 0.25. This indicates that the models retrieve semantically plausible but statistically irrelevant codes. In contrast, CAREP achieves a **Precision@1 of 0.78**, a 3× improvement. By filtering the search space through the causal knowledge, CAREP eliminates spurious correlations before the LLM even begins reasoning.

Semantic Validity (The "Logic" Gap). Beyond finding the right codes, CAREP correctly infers the Boolean rule. On Semantic *Recall@1*, CAREP scores **0.70**, compared to 0.25 for the best baseline. The ablation study reveals that removing the causal indicators drops performance significantly, proving that the LLM cannot infer the correct operators (AND vs. OR, Excitatory vs. Inhibitory) from text descriptions or causes alone.

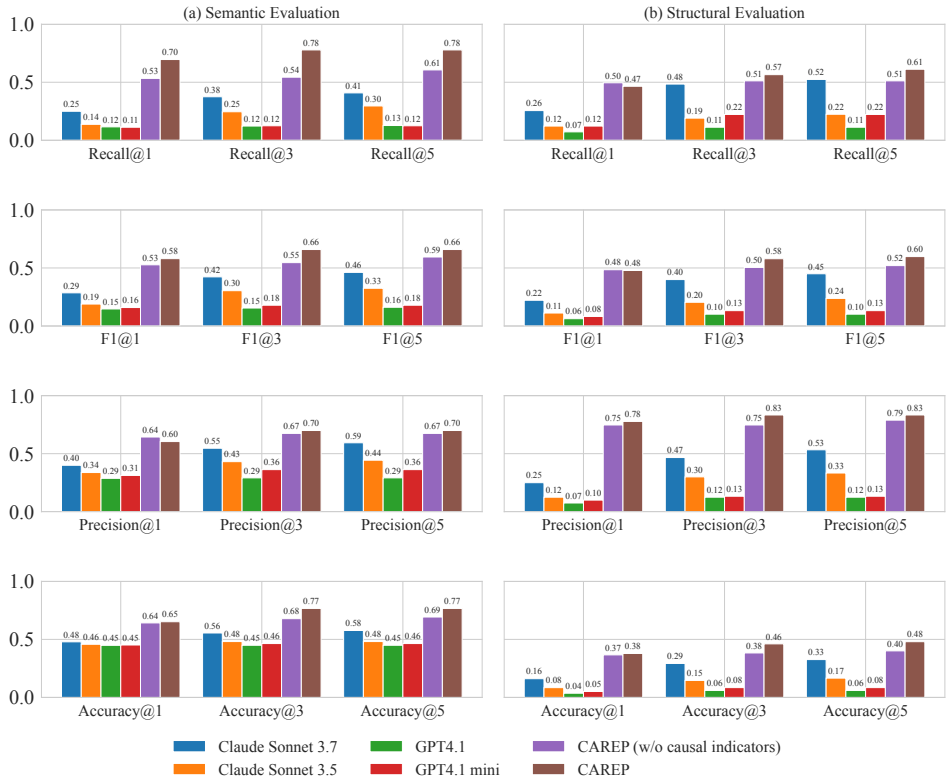

Figure 1: **Neuro-Symbolic Performance.** (a) Semantic Evaluation: Truth-table agreement of the synthesized logic. (b) Structural Evaluation: Precision/Recall of the extracted DTC variables. CAREP significantly outperforms unconstrained LLMs baseline.

REFERENCES

Charles K. Assaad, Emilie Devijver, and Eric Gaussier. Survey and evaluation of causal discovery methods for time series. J. Artif. Int. Res., 73, May 2022. ISSN 1076-9757. doi: 10.1613/jair.1.13428. URL https://doi.org/10.1613/jair.1.13428.

T.M. Cover. Elements of Information Theory. Wiley series in telecommunications and signal processing. Wiley-India, 1999. ISBN 9788126508143. URL https://books.google.de/books?id=3yGJrqyanyYC.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.64. URL https://aclanthology.org/2024.emnlp-main.64/.

Ellery Eells. Probabilistic Causality. Cambridge Studies in Probability, Induction and Decision Theory. Cambridge University Press, 1991.

Chang Gong, Chuzhe Zhang, Di Yao, Jingping Bi, Wenbin Li, and YongJun Xu. Causal discovery from temporal data: An overview and new perspectives. ACM Comput. Surv., 57(4), December 2024. ISSN 0360-0300. doi: 10.1145/3705297. URL https://doi.org/10.1145/3705297.

Hugo Math and Rainer Lienhart. Towards practical multi-label causal discovery in high-dimensional event sequences via one-shot graph aggregation. In NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling, 2025. URL https://openreview.net/forum?id=1HZfpuDVeW.

Hugo Math and Rainer Lienhart. Context-informed sequence classification: A multimodal approach to vehicle diagnostics. In 1st ICLR Workshop on Time Series in the Age of Large Models, 2026a. URL https://openreview.net/forum?id=G4iAE9xOpb.

Hugo Math and Rainer Lienhart. Scalable sample-level causal discovery in event sequences via autoregressive density estimation, 2026b. URL https://arxiv.org/abs/2602.01135.

Hugo Math, Rainer Lienhart, and Robin Schön. Harnessing event sensory data for error pattern prediction in vehicles: A language model approach. Proceedings of the AAAI Conference on Artificial Intelligence, 39(18):19423–19431, 4 2025a. doi: 10.1609/aaai.v39i18.34138. URL https://ojs.aaai.org/index.php/AAAI/article/view/34138.

Hugo Math, Robin Schön, and Rainer Lienhart. One-shot multi-label causal discovery in high-dimensional event sequences. In NeurIPS 2025 Workshop on CauScien: Uncovering Causality in Science, 2025b. URL https://openreview.net/forum?id=z7NT8vGWC2.

Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. GSM-symbolic: Understanding the limitations of mathematical reasoning in large language models. In The Thirteenth International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=AjXkRZIvjB.

OpenAI. Gpt-4 technical report. ArXiv, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Judea Pearl. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988. ISBN 1558604790.

Parivash Pirasteh, Slawomir Nowaczyk, Sepideh Pashami, Magnus Löwenadler, Klas Thunberg, Henrik Ydreskog, and Peter Berck. Interactive feature extraction for diagnostic trouble codes in predictive maintenance: A case study from automotive domain. In Proceedings of the Workshop on Interactive Data Mining, WIDM'19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362962. doi: 10.1145/3304079.3310288. URL https://doi.org/10.1145/3304079.3310288.

J. R. Quinlan. Induction of decision trees. Machine Learning, 1:81–106, 1986.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In ICLR. OpenReview.net, 2023. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2023.html#YaoZYDSN023.

Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. Knowledge and Data Engineering, IEEE Transactions on, 26:1819–1837, 08 2014. doi: 10.1109/TKDE.2013.39.

A Appendix

A.1 Notations & Definitions

We use capital letters (e.g., X) to denote random variables, $P(X)$ the probability distribution of X, $P(X = x) = p(x)$ the probability of the realisation x for the random variable X, and bold capital letters (e.g., $\boldsymbol{X}$) for sets of variables. Let $\boldsymbol{U}$ denote the set of all (discrete) random variables. We define the event set $\boldsymbol{X} = \{X_1, \ldots, X_n\} \subset \boldsymbol{U}$, and the label set $\boldsymbol{Y} = \{Y_1, \ldots, Y_n\} \subset \boldsymbol{U}$.

Definition 1 (Bayesian Network)**.** *Pearl (1988) Let P denote the joint distribution over a variable set $\boldsymbol{U}$ of a directed acyclic graph (DAG) $\mathcal{G}$. The triplet $< \boldsymbol{U}, \mathcal{G}, P >$ constitutes a BN if the triplet $< \boldsymbol{U}, \mathcal{G}, P >$ satisfies the Markov condition: every random variable is independent of its non-descendant variables given its parents in $\mathcal{G}$. Each node $X_t \in \boldsymbol{U}$ represents a random variable. The directed edge $(X_t \to X_j)$ encodes a probabilistic dependence. The joint probability distribution can be factorized $P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_t | X_1, \cdots, X_{i-1})$. If a variable does not depend on all of its predecessors, we can write: $P(X_t | X_1, \cdots, X_{i-1}) = P(X_t | par(X_t))$ with 'par' the parents of node X_t such that: $par(X_t) = \{X_1, \cdots, X_{i-1}\}$.*

Definition 2 (Conditional Independence)**.** *Variables X and Y are said to be conditionally independent given a variable set $\boldsymbol{Z}$, if $P(X, Y | \boldsymbol{Z}) = P(X | \boldsymbol{Z}) P(Y | \boldsymbol{Z})$, denoted as $X \perp Y | \boldsymbol{Z}$. Inversely, $X \not\perp Y | \boldsymbol{Z}$ denotes the conditional dependence. Using the conditional mutual information Cover (1999) to measure the independence relationship, this implies that $\boldsymbol{I}(X, Y | \boldsymbol{Z}) = 0 \Leftrightarrow X \perp Y | \boldsymbol{Z}$.*

A.1.1 Data Generating Process

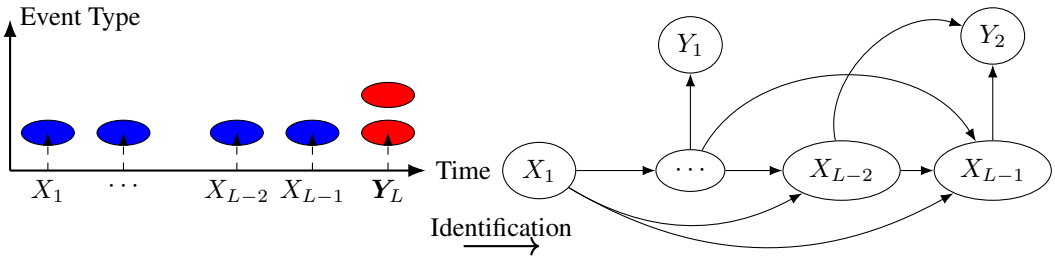

Figure 2: **An example of an instance time causal graph** $\mathcal{G}^{(i)}$ extracted from a multi-label event sequence leading to an outcome.

A.2 Causal Discovery Tool

A.2.1 Pipeline

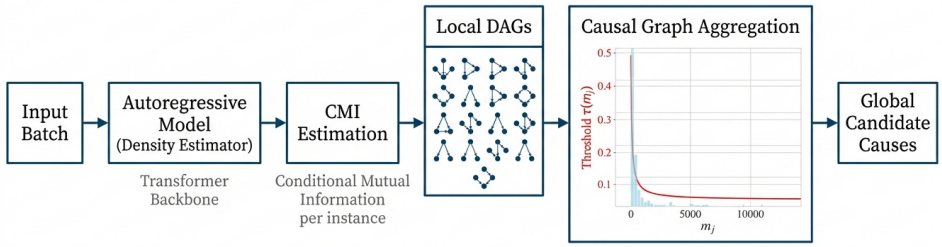

Figure 3: **Causal Discovery from a Batch of Sequences**. The causal discovery follows two steps. (1) We identify if a specific DTC (event) x_t is a cause of EP (outcome) y by measuring their Conditional Mutual Information (CMI) conditioned on a past trajectory $x_{<t}$ for all sequential steps t and provide a local directed acyclic graph (DAG) per sequence as $\mathcal{G}^{(i)}$ Math et al. (2025b) (2) We aggregate these local evidences to form a global consensus graph Math & Lienhart (2025) per outcome y_j as $\mathcal{G}_y^*$ with causal indicators that will be feed to the LLMs.

A.2.2 Ablation

Method	**Precision** ↑	**Recall** ↑	**F1** ↑	**Time (min)** ↓
Phase 1: Local Causal Discovery				
Instance Graph $\mathcal{G}^{(i)}$	55 ± 1.42	31 ± 0.82	40 ± 1.03	**11.7**
Phase 2: Global Aggregation				
Consensus Graph $\mathcal{G}_y^*$	61 ± 1.5	46 ± 1.7	46 ± 1.2	11.8
Phase 3: Diagnostic Agent (End-to-End)				
CAREP	$\mathbf{83 \pm 1.1}$	$\mathbf{60 \pm 1.3}$	$\mathbf{61 \pm 1.1}$	12

Table 1: **Ablation of causal discovery performance across the pipeline stages**: (1) Local Discovery, (2) Global Aggregation, and (3) The final Diagnostic Agent (CAREP). The baselines are evaluated using the structural (does the rule contain the correct variables, ignoring the logic). 'Time' represents cumulative runtime. By adding semantic information (metadata) and LLM's reasoning, CAREP adds a significant prediction margin in selecting the correct *causes* of an outcome (EP).

A.2.3 Adaptive Thresholding

$$\tau_j(m_j) = (\tau_{\max} - \tau_{\min}) \cdot \frac{1}{1 + e^{\alpha(\log m_j - \log m_0)}} + \tau_{\min} \tag{6}$$

where $\tau_{\max} = 0.9, \tau_{\min} = 0.05$, m_0 is set to the median of all label supports and α is set inversely proportional to the log-interquartile range of supports: $\alpha = \frac{2 \log 3}{\log q_{75} - \log q_{25}}$. This ensures the transition adapts to the specific shape of the long-tail distribution.

A.3 Evaluation

A.3.1 Settings

We used a $g4dn.12xlarge$ instance from AWS Sagemaker to run comparisons. It contains 48 vCPUs and 4 NVIDIA T4 GPUs. We used a combination of F1-Score, Precision, and Recall with different averaging Zhang & Zhou (2014) to perform comparisons.

A.3.2 PRETRAINING OF GENERATIVE PRIORS

The language models Tf_x and Tf_y were used with respectively 90M and 15M parameters Math et al. (2025a); Math & Lienhart (2026a), they didn't see the test set during training. We ensure that they reach at least 0.9 F1 Score macro on the provided test-set (unobserved during training) to provide accurate conditional mutual information estimation during the local causal discovery phase.

A.3.3 INPUT FORMAT

```
"unknown EP": {
  "potential_causes": {
    "DTC3": {
      "Eells_mean": 0.70,
      "Eells_std": 0.13,
      "pmi": {
        "DTC345": -1.89
      }
    },
    "DTC345": {
      "Eells_mean": 0.59,
      "Eells_std": 0.12,
      "pmi": {
        "DTC3": -1.89
      }
    },
    "dtcs_samples": {
      "0X001 0X0008 0X0120 0X8900 ..",
      "0x6052 0X0204 0X0129 0X3410 ..",
      ...
     }
    }
}
```

Figure 4: **Example of inputs given to the LLMs for the Boolean rule generation**. The *potential causes* are output by the causal discovery pipeline thus the baselines LLMs don't see it. For instance, *DTC3* is a cause of *unknown EP* and has an Eells measure of 0.7, i.e., it increases the likelihood of observing the EP on average by 70%. Example of defective vehicles with such an error pattern are given in *dtcs samples*

A.3.4 DATASET DISTRIBUTION & COMPLEXITY

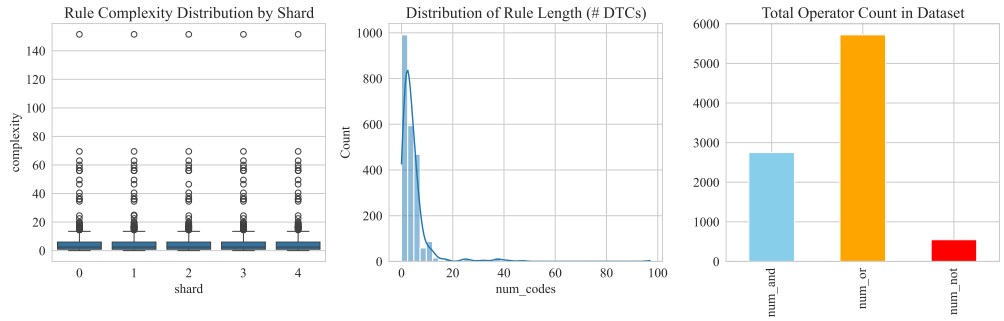

Figure 5: **Analysis of Ground Truth Error Patterns.** (Left) **Complexity Distribution by Shard:** The consistency of rule complexity scores across shards confirms a balanced data split strategy. (Middle) **Rule Length Histogram:** The distribution of Diagnostic Trouble Codes (DTCs) per rule reveals a long-tail nature, with some complex rules involving over 20 unique codes. (Right) **Operator Frequency:** The significant presence of logical NOT ($\neg$) and OR ($\vee$) operators highlights the non-trivial, inhibitory causal structures.

A.3.5 STRUCTURAL AND SEMANTIC EVALUATION

The task of identifying error pattern rules is complex and requires multiple levels of evaluation. Here, we distinguish between two evaluation types: (1) *structural*, where we compare the Boolean rules directly as a classification set, and (2) *semantic*, where we compare the corresponding truth-tables of the estimated Boolean expressions to the truth-table of the ground truth.

Specifically, for (1), we evaluate: *Are the DTCs in the estimated rules correct?* For this, we divide the estimated sets and ground truth using the Boolean operators as separators and perform a standard multi-label classification:

$$\text{dtc1 \& dtc2 \& !dtc5} \mid \text{dtc3} \implies [dtc1, dtc2, dtc3, dtc5]$$

For (2) we enumerate all possible assignments of the present Boolean variables and compute the truth table of the estimated rules and the ground truth to express: *Is the rule logically correct ?* We calculate the accuracy, precision, recall, f1 of the predicted value of the truth tables (e.g., Tab 2 in Appendix).

Table 2: Illustrative truth table for semantic evaluation. The accuracy using the predicted rule is 50%.

x_1	x_2	**Ground Truth** (x_1 & x_2)	**Predicted Rule** ($x_1 \mid x_2$)
0	0	0	0
0	1	0	1
1	0	0	1
1	1	1	1

We compute the top-3 and top-5 metrics by using the 5 estimations from the decreasing order of confidence *[HIGH, MEDIUM, LOW]* to provide a more concrete picture. We average the metrics over the number of labels (macro average).