# [TITLE]

TECHNICAL REPORT

**Causal Copilot**

October 28, 2024

**ABSTRACT**

[ABSTRACT]

***Keywords*** Causal Discovery, Large Language Model, [ALGO], [DATASET]

## 1 Introduction

[INTRO$_INFO$]

## 2 Background Knowledge

### 2.1 Detailed Explanation about the Variables

[BACKGROUND$_INFO1$]

### 2.2 Possible Causal Relations among these Variables

[BACKGROUND$_INFO2$]

Figure 1: Possible Causal Relation Graph

## 3 Dataset Descriptions and EDA

The following is a preview of our original dataset.

Table 1: Dataset Preview
[DATA$_PREVIEW$]

### 3.1 Data Properties

We employ several statistical methods to identify data properties.

The shape of the data, data types, and missing values are assessed directly from the dataframe. Linearity is evaluated using Ramseys RESET test, followed by the Benjamini & Yekutieli procedure for multiple test correction. Gaussian noise is assessed through the Shapiro-Wilk test, also applying the Benjamini & Yekutieli procedure for multiple test correction. Time-Series and Heterogeneity are derived from user queries.

Properties of the dataset we analyzed are listed below.

Table 2: Data Properties
[DATA$_PROP_TABLE$]

### 3.2 Distribution Analysis

The following figure shows distributions of different variables. The orange dash line represents the mean, and the black line represents the median. Variables are categorized into three types according to their distribution characteristics.

Figure 2: Distribution Plots of Variables

[DIST$_INFO$]

### 3.3 Correlation Analysis

Figure 3: Correlation Heatmap of Variables

[CORR$_INFO$]

## 4 Discovery Procedure

[DISCOVER$_PROCESS$]

## 5 Results Summary

(a) True Graph (b) Initial Graph (c) Revised Graph

Figure 4: Graphs Comparision of [ALGO]

The above are result graphs produced by our algorithm. The initial graph is the graph in the first attempt, and the revised graph is the one pruned with LLM suggestion.

### 5.1 Graph Reliability Analysis

Figure 5: Reliability Graph

Based on the confidence probability heatmap and background knowledge, we can analyze the reliability of our graph.

[RELIABILITY$_ANALYSIS$]