CauSciBench: Can LLMs Automate Causal Inference in Real-World Scientific Research?

Sawal Acharya[1,2]* Terry Jingchen Zhang[2,3,4]* Andrew Kim[2] Rahul Babu Shrestha[2] Xianlin Sun[5] Pepijn Cobben[2,3,4] Maximilian Mordig[3,6] Jacob Tae Emmerson[2] Anahita Haghighat[2] Furkan Danisman[2] Yuen Chen[7] Clijo Jose[8] Andrei Ioan Muresanu[2,4] Justin Cui[2] Jiarui Liu[2,9] Yahang Qi[2] Punya Syon Pandey[2,4] Yinya Huang[3] Bernhard Schölkopf[6,10] Zhijing Jin[2,4,6]

[1]Stanford University [2]Jinesis Lab, University of Toronto & Vector Institute [3]ETH Zürich [4]EuroSafeAI [5]HKU [6]Max Planck Institute for Intelligent Systems, Tübingen [7]UIUC [8]IP Paris [9]CMU [10]ELLIS Institute Tübingen

Introduction & Motivation

- Causal inference is central to empirical research across domains such as social science, public health, economics, and biomedicine.
- Ground truth causal effects are not directly observable, and appropriate causal analysis requires careful consideration of the scenario, the data-generating process, and the assumptions underlying suitable methods.
- The need for both domain knowledge and technical understanding has opened the possibility of using LLMs to automate causal inference.
- We introduce **CauSciBench**, a benchmark for evaluating whether LLMs can perform *end-to-end causal inference*: selecting methods and variables, estimating causal effects, and interpreting results in realistic scientific settings.

CauSciBench

- The benchmark consists of 300+ queries drawn from (1) real-world publications, (2) QRData-CI (textbook examples and seminal datasets), and (3) synthetic scenarios.
- Provides standardized annotations for evaluating the full causal inference pipeline: method selection, variable selection, effect estimation, and interpretation.
- Covers **foundational causal inference methods** used in empirical research and evaluates **7 frontier LLMs** under **4 prompting strategies**.

Sample Annotation

Sample Query Based on Card (1993)

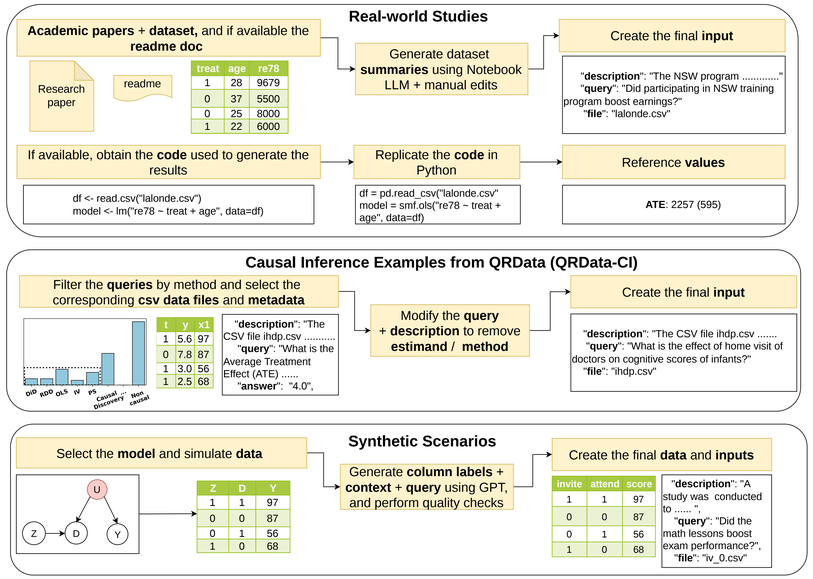
Paper Source: Using geographic variation in college proximity to estimate the return to schooling.
Description: The National Longitudinal Survey of Young Men (NLSYM) was conducted to collect data on demographics, education, and employment outcomes. Participants were tracked over time to study long-term patterns. The dataset used here comes from the 1976 wave of the survey. Variables in the dataset:

- lwage: log of wages
- educ: years of education
- exper: years of work experience
- black: 1 if Black, 0 otherwise
- south: 1 if lives in a southern state, 0 otherwise
- married: 1 if married, 0 otherwise
- smsa: 1 if living in a metropolitan area, 0 otherwise
- nearc4: 1 if there is a four-year college in the county, 0 otherwise

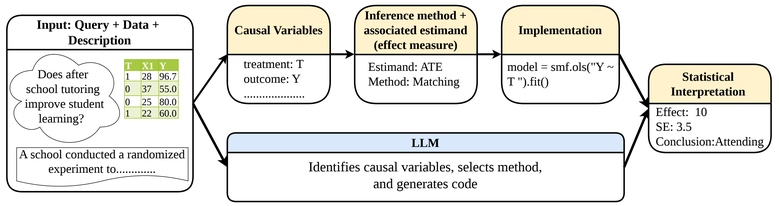
Query: What is the effect of education on earnings?
Answer: 0.132
Standard Error: 0.049
Method: IV (Instrumental Variable)
Instrument Variable: nearc4
Dataset File: card_geographic.csv
Reference in Paper: Table 4
Field: Economics

Figure 1. Sample benchmark instance with color-coded treatment variable, outcome variable, and control covariates, along with other attributes including the method (IV), the dataset filename, and the corresponding reference in the source paper.

Benchmark Construction



Experimental Setup



Models Evaluated: GPT-4o-mini, GPT-4o, GPT-5-mini, GPT-4.1, OpenAI-o3, Gemini-2.5-Flash-Lite, Qwen3-Next-80B-Inst, Grok-4-Fast.

Prompting Strategies: Direct, Chain of Thought (CoT), Program of Thought (PoT), ReAct.

Evaluation Metrics: Method Correctness (F1 score), Effect Accuracy (effect within 5% of reference value), Variable Selection Accuracy for treatment, outcome, and controls.

Results and Findings

- Performance on real data is lower than on other dataset types. Real data involves more variables, increased preprocessing due to the presence of non-numeric values, and more complex research designs.

Strategy	Real	QRData-CI	Synthetic
Direct	60.58	85.44	88.99
CoT	53.75	88.39	91.60
PoT	49.64	71.79	80.86
ReAct	51.84	82.49	65.78
Avg.	54.00	82.03	81.81

Table 1. Method Correctness (F1 score) of GPT-5-mini across 4 prompting strategies and 3 datasets.

- The main bottlenecks are method selection and control variable selection.

Model	Method C. (F1 Score)	Variable Selection Accuracy		
		Treatment	Outcome	Control
GPT-4o-mini	17.91	83.90	92.62	48.65
GPT-4o	46.50	77.66	95.00	49.07
GPT-5-mini	53.95	78.65	94.78	55.31
OpenAI-o3	64.56	82.95	94.67	51.86

Table 2. Variable Selection Accuracy and Method Correctness (F1 score) of OpenAI models on the Real dataset.

- Models frequently default to OLS as the baseline method. This bias is more pronounced in smaller models such as GPT-4o-mini.
- o3 models show poor code completion rates under reasoning-heavy strategies such as ReAct and CoT, falling below 50%.

Key Takeaways

- Empirical causal inference involves two components: *reasoning* to build the causal model (selecting the method, treatment, outcome, and controls) and *computation* to implement it correctly. Both are required for end-to-end success.
- Current LLM-powered tools tend to address only one of these. Achieving reliable end-to-end causal inference requires equal attention to both components.
- Evaluations on real-world data are essential. While seminal and textbook datasets are clean and structured, real-world data are noisy, require extensive preprocessing, and better reflect the challenges faced in practice.

References

D. Card, "Using geographic variation in college proximity to estimate the return to schooling," National Bureau of Economic Research, Working Paper 4483, October 1993.