CauSciBench: Assessing LLM Causal Reasoning for Scientific Research

Anonymous Author(s)
Affiliation
Address
email

Abstract

While large language models (LLMs) are increasingly integrated into scientific research, their capability to perform causal inference remains under-evaluated. Existing benchmarks either focus narrowly on method execution or provide open-ended tasks lacking precision in defining causal estimands, methodological choices, and variable selection. We introduce CauSciBench, a comprehensive benchmark combining expert-curated problems from published research with diverse synthetic scenarios and textbook examples. Our benchmark spans both potential outcomes and Pearl's structural causal model frameworks, enabling systematic evaluation of LLM causal reasoning in scientific contexts. By leveraging temporal publication structure, CauSciBench also provides a foundation for detecting data contamination through questions based on papers published before and after LLM knowledge cutoff dates.

1 Introduction

Causal inference is fundamental to scientific discovery, enabling researchers to establish cause-and-effect relationships across social science [12], public health [7], and biomedicine [16]. LLM integration into scientific workflows creates opportunities to democratize sophisticated causal analysis. Recent LLM-powered agents show promise for automating causal inference procedures [10, 25], potentially accelerating research across disciplines [15].

Evaluating LLM causal inference capabilities presents unique challenges. Causal inference deals with unobservable counterfactual outcomes [11], requiring sophisticated methodological frameworks and identification strategies. Current approaches typically assume users can appropriately select methods and specify problems [17, 3]. Whether LLMs demonstrate genuine causal reasoning or sophisticated pattern matching remains an open question [15].

Existing benchmarks address different aspects but leave gaps. Text-based approaches evaluate commonsense causal understanding [22, 20, 14, 4] or formal reasoning within Pearl's SCM framework [13, 5]. Implementation-focused benchmarks like QRData [17] assess method execution on tabular data but not problem formulation from natural language descriptions. General data analysis benchmarks such as BLADE [9] and DiscoveryBench [19] provide open-ended tasks without causal inference specificity.

CauSciBench bridges these gaps through systematic evaluation across the complete analysis pipeline. Our benchmark provides fine-grained assessment from problem formulation and variable selection to method choice, estimation, and interpretation. We make three key contributions: (1) 100 expert-curated problems from published research across economics, epidemiology, political science, and public health capturing authentic methodological complexity, (2) controlled synthetic evaluation framework with known causal structures enabling systematic assessment of identification strategies,

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

and (3) dual-purpose evaluation framework serving both capability assessment and data contamination detection through temporal publication structure.

2 Problem Formulation

Our goal is assessing LLMs' ability to generate answers to causal queries through sound causal analysis involving: (i) framing the causal estimation problem by selecting appropriate treatment and outcome variables and the correct estimand, (ii) assessing whether the estimand can be identified from the provided dataset, (iii) formulating and implementing the correct statistical model, and (iv) extracting and interpreting the causal effect.

Each benchmark instance consists of five core components: **Data** (experimental or observational input), **Dataset Description** (information about data collection, variable definitions, and background context), **Query** (causal question involving the effect of one variable on another), **Causal Inference Method and Effect Estimate** (expert-validated method and corresponding effect providing ground truth), and **Model Variables** (key variables including treatment, outcome, confounders, and method-specific variables).

3 Dataset Collection: CauSciBench

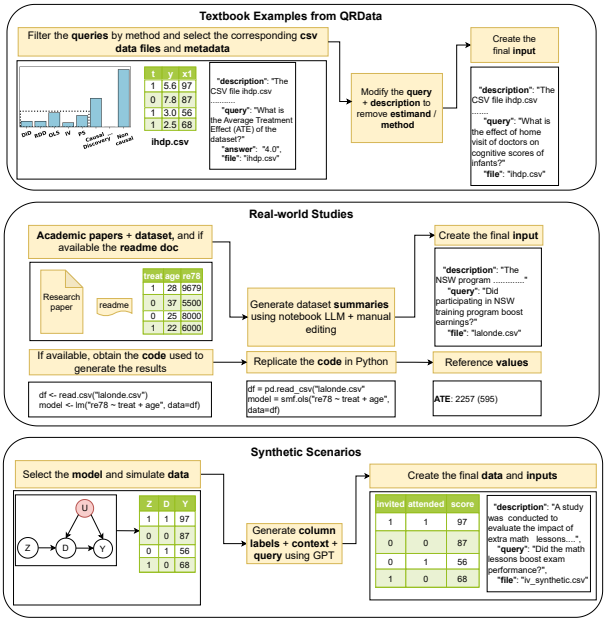


Figure 1: Dataset creation process for **QRData, Real-world Studies, and Synthetic Data**

Figure 1 details our comprehensive dataset creation process across all three sources. Table 1 positions CauSciBench relative to existing benchmarks, highlighting our unique combination of end-to-end analysis, intermediate evaluation, data semantic comprehension, and synthetic scenarios. Figure 4 shows the distribution of causal inference methods across our three dataset collections, demonstrating methodological diversity essential for comprehensive evaluation.

Source 1: Research Paper Curation We compile papers from economics, criminology, public health policy, and political science, creating comprehensive summaries capturing key dataset information including variable descriptions, data collection procedures, and research purpose. We formulate causal queries by systematically examining empirical methodology and conclusions from causal effects, selecting methods authors cite to justify findings and choosing the most expressive model

[1]102 is used as the number of queries for QRData, as only 102 of 411 is causal.

Benchmark	End-to-End Analysis	Intermediate Analysis	Comprehend Data Semantics	Synthetic Scenarios	Answer Format	Task Sources	# Queries
corr2cause [14]	✗	✗	✓	✓	Freeform QA	10 Publications	207,972
CLadder [13]	✗	✓	✗	✓	Freeform QA	9 Publications	10,112
QRData [18]	✗	✗	✓	✗	Freeform QA	195 Publications	102[1]
DiscoveryBench [19]	✓	✗	✓	✓	Freeform QA	27 Publications	239
BLADE [9]	✓	✓	✓	✗	Analysis Code	31 Publications	12
CauSciBench	✓	✓	✓	✓	Point Estimate	52 Publications	305

Table 1: Comparison of CauSciBench against existing benchmark datasets.

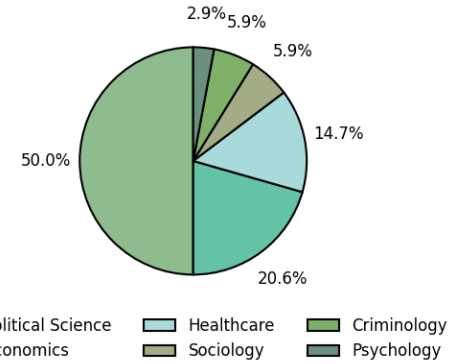


Figure 2: Distribution of paper domains in Real-world publications

specifications for completeness. This curation process ensures methodological rigor reflecting real-world research standards while providing authentic complexity requiring navigation of confounders, identification validity assessment, and results interpretation within disciplinary contexts.

Source 2: Automated Synthesis We automatically synthesize datasets by randomly selecting true causal effects τ in range $(1, 10)$ with continuous covariates drawn from normal distributions and binary covariates from binomial distributions. For randomized trials: $Y = \alpha + X\vec{\theta} + \tau T + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$, $\vec{\theta} \sim \mathcal{N}(u, kI)$, and α is the intercept. We use GPT-4o to synthesize diverse contexts for each synthetic dataset, creating plausible scenarios explaining data collection with comprehensive dataset metadata including headings and descriptions. This approach improves dataset diversity while testing model performance consistency in high-fidelity scenarios mirroring real-world research contexts.

Source 3: Refined QRData Since QRData tasks specify inference methods or estimands and our focus is end-to-end causal inference including automatic method and variable selection, we systematically modify queries by removing explicit references to estimation techniques. For example, "What is the Average Treatment Effect (ATE) of the dataset?" becomes "What is the effect of home visits by doctors on cognitive scores of infants?" We retain original dataset descriptions and numerical causal effect estimates, restricting evaluation to queries with numerical answers to enable precise quantitative assessment.

4 Experimental Setup

We investigate two prompting strategies: Direct prompting provides comprehensive dataset information with causal questions, testing implicit expertise for methodological choices without intermediate reasoning steps. Chain of Thought (CoT) maintains the same input but breaks down the workflow:

variable selection → identification → statistical estimation model → implementation. Models first identify treatment, outcome, and confounding variables with justifications, then pick estimands and corresponding methods while explaining identification assumption satisfaction.

For causal effect estimation, we use `DoWhy` [24, 1] and `statsmodels` [23] libraries with GPT-4o-mini, GPT-4.1, and OpenAI-o3 as backbone LLMs [2, 21]. We evaluate using Method Selection Accuracy (MSA): $\text{MSA} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}[\hat{m}_i = m_i] \times 100$ and Mean Relative Error (MRE): $\text{MRE} = \frac{1}{N}\sum_{i=1}^{N} \min\left(\frac{|\hat{\tau}_i - \tau_i|}{|\tau_i|}, 1\right) \times 100\%$.

5 Results and Discussion

Performance results are shown in Table 2. Real-world causal estimation proves challenging with relative errors consistently exceeding 50% for real-world data, reflecting inherent messiness of real datasets lacking preprocessing. Larger models show superior performance as scaling effects apply with models (4.1 and o3) consistently outperforming smaller 4o-mini across both metrics. CoT prompting shows conditional effectiveness but does not universally improve performance over direct prompting, aligning with previous findings that CoT can degrade performance on implementation-oriented quantitative reasoning tasks.

Table 3 demonstrates that methodological misselection directly amplifies estimation errors as wrong method choices consistently produce substantially higher mean relative errors across nearly all evaluation contexts. This performance degradation is particularly acute in real-world datasets, underscoring how methodological sophistication becomes increasingly critical as data complexity approaches realistic conditions.

Detailed analysis in the appendix (see Table 3 and Figures 3a-3b) reveals systematic failure modes. Models systematically default to OLS estimation with pronounced bias toward Ordinary Least Squares selection across all model variants, creating algorithmic anchoring that overwhelms nuanced methodological considerations. Methodological misselection directly amplifies estimation errors as wrong method choices consistently produce substantially higher mean relative errors. Implementation failures persist even with correct methodological reasoning due to inappropriate variable selection, model misspecification, or algorithmic implementation mistakes.

		Method Accuracy (↑)			**Mean Rel. Error (↓)**		
Dataset	**Prompt**	**4o-mini**	**4.1**	**o3**	**4o-mini**	**4.1**	**o3**
Real	Basic	34.57	47.78	**71.76**	71.45	58.43	53.82
	CoT	40.23	55.56	67.74	62.62	53.59	**53.02**
Synthetic	Basic	15.38	59.43	72.41	22.58	**6.16**	6.30
	CoT	24.56	**77.14**	69.23	17.25	10.99	17.24
Textbook	Basic	60.00	64.10	69.23	42.03	40.05	46.41
	CoT	53.85	**71.79**	66.67	41.29	33.68	**30.59**

Table 2: Performance comparison across datasets and prompting methods.

6 Conclusion

CauSciBench establishes a comprehensive framework for evaluating causal inference capabilities in large language models, revealing critical limitations requiring attention before these systems can reliably support scientific research. Current LLMs exhibit systematic biases toward methodological oversimplification with concerning defaults to OLS estimation regardless of identification requirements, while struggling with implementation precision even when methodological reasoning proves sound. The substantial performance gap between synthetic and real-world scenarios suggests that advancing LLM causal reasoning requires developing more robust frameworks for handling observational complexity, improving methodological selection algorithms beyond pattern matching, and bridging the execution gap between theoretical understanding and practical implementation to democratize sophisticated causal analysis across scientific disciplines.

References

[1] Patrick Blöbaum, Peter Götz, Kailash Budhathoki, Atalanti A. Mastakouri, and Dominik Janzing. Dowhy-gcm: An extension of dowhy for causal inference in graphical causal models. *Journal of Machine Learning Research*, 25(147):1–7, 2024.

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[3] Qiang Chen, Tianyang Han, Jin Li, Ye Luo, Yuxiao Wu, Xiaowei Zhang, and Tuo Zhou. Can ai master econometrics? evidence from econometrics ai agent on expert-level tasks, 2025.

[4] Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. Causal evaluation of language models, 2024.

[5] Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. Causal evaluation of language models, 2024.

[6] Alan S. Gerber, Donald P. Green, and Christopher W. Larimer. Social pressure and voter turnout: Evidence from a large-scale field experiment. *The American Political Science Review*, 102(1):33–48, 2008.

[7] Thomas A. Glass, Steven N. Goodman, Miguel A. Hernán, and Jonathan M. Samet. Causal inference in public health. *Annual review of public health*, 34:61–75, March 2013.

[8] Rebecca Goldstein and Hye Young You. Cities as lobbyists. *American Journal of Political Science*, 61(4):864–876, 2017.

[9] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A. Merrill, Jeffrey Heer, and Tim Althoff. Blade: Benchmarking language model agents for data-driven science, 2024.

[10] Kairong Han, Kun Kuang, Ziyu Zhao, Junjian Ye, and Fei Wu. Causal agent based on large language model, 2024.

[11] Paul W. Holland. Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, 1986.

[12] Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[13] Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[14] Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*, 2024.

[15] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. Featured Certification.

[16] Samantha Kleinberg and George Hripcsak. Methodological review: A review of causal inference for biomedical informatics. *J. of Biomedical Informatics*, 44(6):1102–1112, December 2011.

[17] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are LLMs capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9215–9235, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[18] Xiao Liu, Zirui Wu, Xueqing Wu, Pan Lu, Kai-Wei Chang, and Yansong Feng. Are llms capable of data-based statistical and causal reasoning? benchmarking advanced quantitative reasoning with data, 2024.

[19] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models, 2024.

[20] Allen Nie, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori Hashimoto, and Tobias Gerstenberg. Moca: measuring human-language model alignment on causal and moral judgment tasks. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

[21] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan

Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[22] Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. CRAB: Assessing the strength of causal relationships between real-world events. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore, December 2023. Association for Computational Linguistics.

[23] Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

[24] Amit Sharma and Emre Kiciman. Dowhy: An end-to-end library for causal inference. *arXiv preprint arXiv:2011.04216*, 2020.

[25] Xinyue Wang, Kun Zhou, Wenyi Wu, Har Simrat Singh, Fang Nan, Songyao Jin, Aryan Philip, Saloni Patnaik, Hou Zhu, Shivam Singh, Parjanya Prashant, Qian Shen, and Biwei Huang. Causal-copilot: An autonomous causal analysis agent, 2025.

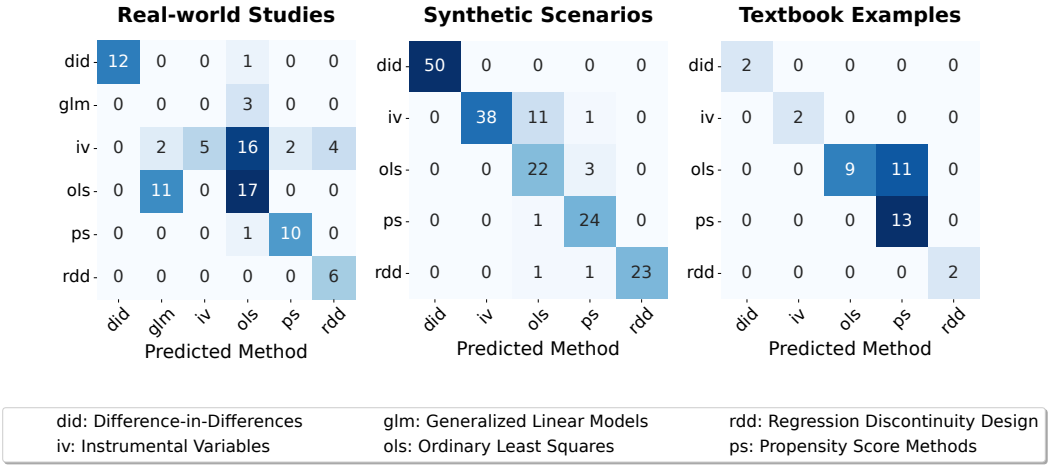
A Limitations

Our work has several limitations. The expert-curated subset requires extensive manual curation creating scalability constraints and potential annotation inconsistencies. Results are based on pass@1 evaluation to balance budgetary constraints with broad model coverage, though pass@k would strengthen findings generalizability. Our benchmark focuses primarily on potential outcomes framework with limited Pearl's structural causal model coverage. Synthetic data generation may not fully capture real-world dataset complexity including missing data patterns, measurement error, and domain-specific confounding structures. The binary treatment focus excludes multi-valued and continuous treatment scenarios while emphasis on tabular data overlooks emerging applications to text, images, and high-dimensional data.

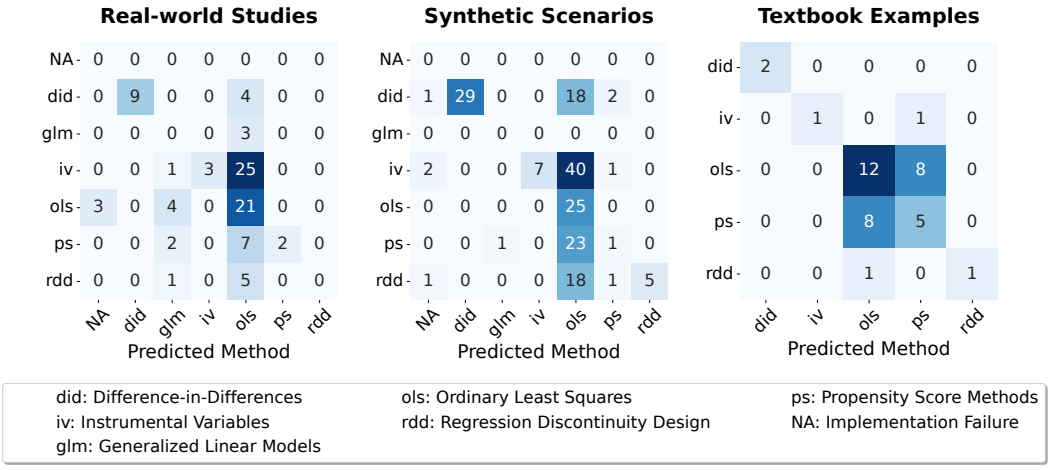
B Detailed Results and Failure Analysis

Model	**Real** (C / W)	**Synth** (C / W)	**Textbook** (C / W)
4o-mini	51.56 / 70.48	13.83 / 19.48	40.09 / 43.33
GPT-4.1	43.31 / 67.40	10.66 / 13.92	42.27 / 11.81
o3	44.51 / 71.27	14.67 / 33.75	35.35 / 15.34

Table 3: Relative error of causal effect estimation: **Correct (C)** vs. **Wrong (W)** method selection across LLMs and datasets for **CoT** prompting



(a) **GPT-4.1:** Confusion matrix for method selection across the three datasets



(b) **GPT-4o-mini:** Confusion matrix for method selection across the three datasets

Figure 3: Confusion matrix for method selection under **CoT-based baseline** with (a) GPT-4.1; (b) GPT-4o-mini.

C Dataset Curation Process

The dataset curation process of our work follows a three-stage methodology, designed to ensure high quality benchmarks through rigorous, expert-curated papers.

- **Paper Selection** focuses on finding articles from diverse fields such as healthcare and economics that utilize established estimation methods including OLS, DiD, RDD, IV, and propensity score methods. The selection criteria emphasized reproducibility and dataset complexity, where we prioritize papers with simpler and explicit approach to causal estimation to work with current LLM's preprocessing limitations. Furthermore, as we go through replication process in future steps, we exclude papers that do not include a publicly accessible dataset with adequate data sharing licensing.
- **Core Information Extraction** follows paper selection, focusing on extracting the core information that causal scientists require for a causal analysis, including treatment variables, outcomes, and non-causal natural language queries to avoid any methodological hints. Multiple questions per paper are permitted when the controls or outcomes differ meaningfully, maximizing the scientific value, while preventing analytical redundancy.
- **Quality Filtering** implements multi-layered expert inspection throughout the entire curation process. All curated datasets undergo replication verification, where experts replicate the estimation process in Python, and exclude all papers that fail to reproduce the original estimates within 10% error in around 50 lines of code. This process validates that the estimates in the paper are truly replicable with the given dataset and methods, so that should LLM fails to replicate the results, the cause lies in the LLM's approach, and not the dataset or the paper's approach.

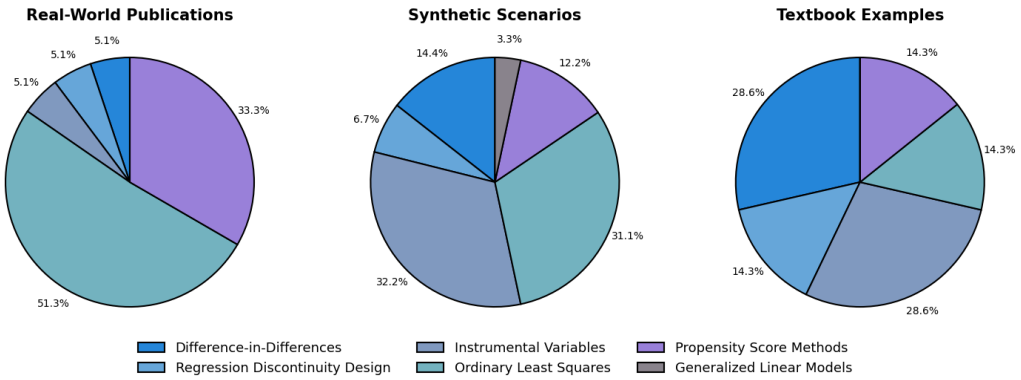


Figure 4: Distribution of estimation methods across the three dataset collections

D Inference Method Selection

For each article, we select the appropriate causal inference method for each query through a systematic approach:

- **Method Identification:** The LLMs are provided with the natural language causal query that does not provide any clues regarding the inference method used, alongside the dataset and a brief summary of the dataset. Reflecting on the provided context, the LLM suggests an inference method that it finds to be the most appropriate for this causal query. In doing so, some models tend to show a bias towards OLS as shown in Table 2, but we discard those suggestions to emphasize on more sophisticated inference methods.
- **Covariate Completeness Check:** Similarly, we prioritize answers associated with the model that uses all, if not, as many of the valid covariates to promote the highest degree of replication. If the specification omits certain variables, we verify that the found effect remains consistent within a small relative margin with the full covariate model to maintain the accuracy of the answer.

E Sample Questions From Each Pillar

Real-World Publications
Source: Cities as Lobbyists [8]
Domain: Economics
Causal Question: How much does the money spent on lobbying increase the number of earmarks received?
Method: Instrumental Variables **Treatment:** `ln_citylobby` (log of city lobbying spending) **Instrument:** `direct_flight_dc` (1=direct flight to DC in 2007, 0=otherwise) **Outcome:** `ln_earmark` (log of total earmarks 2008-2009) **Controls:** `state`, `pop_e`, `land_e`, `water_e`, `senior_e`, `student_e`, `ethnic_e`, `mincome_e`, `unemp_e`, `poverty_e`, `gini_e`, `city_propertytaxshare_e`, `city_intgovrevenueshare_e`, `city_airexp_e`, `houdem_e`, `ln_countylobby` **Data:** Cities with population over 25,000, 2007-2009 panel
Synthetic Dataset
Source: Cardiovascular Rehabilitation Program Effectiveness Study
Domain: Healthcare
Causal Question: Does the new rehabilitation program help patients with cardiovascular diseases recover faster?
Method: Regression Discontinuity Design **Treatment:** `treatment_received` (1=new program, 0=standard care) **Running Variable:** `income_level` (threshold at 12 for eligibility) **Outcome:** `recovery_time` (days to recovery) **Controls:** `patient_age`, `health_index`, `smoking_status`, `obesity_status` **Data:** Regional health department evaluation study
Textbook Examples
Source: Effect of Cigarette Taxation on Consumption [17]
Domain: Healthcare, Political Science
Causal Question: Did Proposition 99 help reduce cigarette sales?
Method: Difference-in-Differences **Treatment:** `california` (1=CA with Prop 99, 0=other states) **Time:** `after_treatment` (1=post-1988, 0=pre-1988) **Outcome:** `cigsale` (total cigarette sales) **Controls:** `state`, `year`, `lnincome`, `beer`, `age15to24`, `retprice` **Data:** 39 US states, 1970-2000 panel

Table 4: Sample questions from each source pillar with the information regarding the paper that the LLM uses as context.

F Prompt Templates

In this section, we present the templates for two of the baseline prompting strategies: **Direct Prompt** and **Chain of Thoughts (CoT) prompt**

Direct Prompt

You are an expert in statistics and causal reasoning. You will answer a causal question on a tabular dataset.
The dataset is located at `{self.dataset_path}`.
The dataset has the following description:

`{self.dataset_description}`

To help you understand it, here is the result of `df.describe()`:

`{df_info}`

Here are the columns and their types:

`{columns_and_types}`

Here are the first 5 rows of the dataset:

`{df.head()}`

If there are fewer than 10 columns, here is the result of `df.cov()`:

```
{(df.cov(numeric_only=True) if len(df.columns) < 10
else "Too many columns to compute covariance")}
```

Finally, here is the output of `df.isnull().sum(axis=0)`:

`{nan_per_column}`

The causal question I would like you to answer is:

`{self.query}`

Here are some example methods; you can choose one from them:

- IPW (Inverse Probability Weighting); choose the right estimand (ATE/ATT/ATC)
- matching_treatment_to_control; choose the ATT
- linear_regression with control variables; output the coefficient of the variable of interest
- instrumental_variable; output the coefficient of the variable of interest
- matching; choose the right estimand (ATE/ATT/ATC)
- difference_in_differences ; output the coefficient of the variable of interest
- regression_discontinuity_design; output the coefficient
- linear_regression / difference_in_means; output the coefficient / the difference in means
- generalized_linear_models (GLM); output the coefficient of the variable of interest

`{method_explanation}`
Using the descriptions and information from the dataset, implement Python code to answer the causal question. Remember, the dataset is located at `{self.dataset_path}`. If you need to preprocess the data, please do so in the code. The following libraries are available to you: `dowhy`, `pandas`, `numpy`, `scipy`, `scikit-learn`, and `statsmodels`. Use the methods from the libraries as best as you can.
Do not code yourself what is already implemented in the libraries. Do not create random data. Make sure it outputs the quantitative value in the comments of the example method. The code you output will be executed, and you will receive the output. Please make sure to output only one block of code, and make sure the code prints the result you are looking for at the end.
Everything between your first code block: ```` ```python ```` and ```` ``` ```` will be executed. If there is an error, you will have several attempts to correct the code.

Chain of Thoughts Prompt

You are an expert in causal inference. You will use a chain-of-thought approach to answer a causal question on a tabular dataset.
The dataset is located at `{self.dataset_path}`.
The dataset has the following description:

`{self.dataset_description}`

To help you understand it, here are the columns and their types:

`{columns_and_types}`

Here are the first 5 rows of the dataset:

`{df.head()}`

If there are fewer than 10 columns, here is the result of df.cov():

```
{(df.cov(numeric_only=True) if len(df.columns)
< 10 else "Too many columns to compute covariance")}
```

Here is the output of df.isnull().sum(axis=0):

`{nan_per_column}`

The causal question I would like you to answer is:

`{self.query}`

Let us approach this problem step by step.
Step 1. First, go through the dataset description and the columns and their types. Then, identify the treatment variable, the outcome variable, and the potential confounders. Explain your reasoning for choosing these variables.
Step 2. What would be the right estimand to consider for this problem? Then, choose the most appropriate method that can be used to estimate the causal effect. Here are some methods. You can choose one from them:

- IPW (Inverse Probability Weighting); choose the right estimand (ATE/ATT/ATC)
- matching_treatment_to_control; choose the ATT
- linear_regression with control variables; output the coefficient of the variable of interest
- instrumental_variable; output the coefficient of the variable of interest
- matching; choose the right estimand (ATE/ATT/ATC)
- difference_in_differences ; output the coefficient of the variable of interest
- regression_discontinuity_design; output the coefficient
- linear_regression / difference_in_means; output the coefficient / the difference in means
- generalized_linear_models (GLM); output the coefficient of the variable of interest

Explain why you chose this method, and how the data and its description support your choice. This means you should explain why the identification assumptions of the method are satisfied.

`{method_explanation}`

Step 3. Next, we will plan the implementation. Before writing the code, describe your implementation process. This includes:

1. Describing the necessary preprocessing steps.
2. How we will select the variables to use in the model.

Step 4. Finally, reflecting on the previous steps, write Python code to answer the causal question: `{self.query}`. Feel free to preprocess the data. The following libraries are available to you: `dowhy`, `pandas`, `numpy`, `scipy`, `scikit-learn`, `statsmodels`. Use the methods from the libraries to implement the method you chose. Be careful about implementation. Recall that the dataset is located at `{self.dataset_path}`. Make sure your code describes key steps and outputs the final results, including:

1. The causal effect
2. The standard deviation
3. Whether the effect is significant or not

The code you write will be executed, and you will next analyze the output. To ease the process, please output one block of code, and make sure the code prints the key results and values.
Everything between your first code block: ‘‘‘`python` and ‘‘‘ will be executed. If there is an error, you will have several attempts to correct the code. Hence, if there is an error, please fix it and re-run.

G Annotation Details

For each article we curate the following information:

- **Paper Name:** Name of the study
- **Description:** The description about the dataset that includes the collection process, purpose, and brief explanation about the variable names
- **Query:** Causal question associated with the dataset
- **Answer:** Causal effect derived in the paper
- **Standard Error:** Standard error associated with the causal effect estimate
- **Significant:** Binary variable indicating if the effect is statistically significant
- **Method:** The causal inference method
- **Treatment:** The name of the treatment variable in the dataset
- **Outcome:** The name of the outcome variable in the dataset
- **Control Covariates:** The control variables / confounders used in the estimation model
- **Interaction Variable:** The name of the variable that interacts with the treatment. This is used for measuring heterogeneous treatment effects
- **Instrument:** The variable used as an instrument. If instrumental variable is not used, this is set to null
- **Running Variable:** The running variable for Regression Discontinuity Design (RDD). If RDD is not used, we set this to null
- **Temporal Variable:** The variable denoting the timing of treatments. This is used for difference-in-differences
- **State Variable:** The variable denoting the different participating entities. This is used for two way fixed effects versions of difference in difference
- **Multi-RCT Treatment Variable:** The treatment type of interest. This is used in RCTs with multiple treatments
- **Data File:** The name of the csv file containing the data
- **Reference:** Reference to the original paper, where the result is found
- **Publication Year:** The year the original study was published

G.1 Annotation Example

Below we provide a sample annotation for a query based on **(author?)** [6]

Annotation Example

- **Paper Name:** Social Pressure and Voter Turnout: Evidence from a Large-Scale Field Experiment
- **Description:** The randomized experiment aims to analyze the effect of different types of social pressures on voter behavior. A field experiment was conducted in Michigan ahead of the August 2006 primary election. Households were randomly assigned to a control group or one of four treatment groups: Civic Duty, Hawthorne, Self, Neighbors. Eleven days before the election, each treatment group received a different mailing:
  - Civic Duty: Emphasized the recipient's responsibility as a citizen to vote.
  - Hawthorne: Notified recipients that their voting behavior would be studied using public records, introducing mild social pressure.
  - Self: Listed the voting history of all registered voters in the household and noted that an updated chart would be mailed after the election.
  - Neighbors: Included both the household's and neighbors' voting records, implying public exposure of voting behavior.
  - Control Group: Received no mailing.

  Variables in the dataset:
  - sex: Participant's sex (male or female)
  - g2000, g2002, g2004: Voted in the 2000, 2002, and 2004 gubernatorial elections
  - p2000, p2002, p2004: Voted in the 2000, 2002, and 2004 primary elections
  - treatment: Assigned group (Civic Duty, Hawthorne, Neighbors, Self, or Control)
  - cluster: Cluster identifier for the unit
  - voted: Indicator for voting in the 2006 primary election
  - hh_id: Household ID
  - hh_size: Number of individuals in the household
  - yob: Year of birth of the participant
- **Query:** Does the Hawthorne scheme lead to an increase in voter turnout?
- **Answer:** 0.026
- **Standard Error:** 0.003
- **Significant:** Yes
- **Method:** OLS
- **Treatment:** treatment
- **Outcome:** voted
- **Control Covariates:** g2000, g2002, p2000, p2002, p2004
- **Interaction Variable:** null
- **Instrument:** null
- **Running Variable:** null
- **Temporal Variable:** null
- **State Variable:** null
- **Multi-RCT Treatment Variable:** Hawthorne
- **Data File:** voter_turnout_data.csv
- **Reference:** Table 3a
- **Publication Year:** 2008