CausalReasoningBenchmark: A Real-World Benchmark for Disentangled Evaluation of Causal Identification and Estimation

Ayush Sawarni
Stanford University
Stanford, CA, USA
ayushsaw@stanford.edu

Jiyuan Tan
Stanford University
Stanford, CA, USA
jiyuantan@stanford.edu

Vasilis Syrgkanis
Stanford University
Stanford, CA, USA
vsyrgk@stanford.edu

Abstract

Many benchmarks for automated causal inference evaluate a system's performance based on a single numerical output, such as an Average Treatment Effect (ATE). This approach conflates two distinct steps in causal analysis: **identification**—formulating a valid research design under stated assumptions—and **estimation**—implementing that design numerically on finite data. We introduce **CausalReasoningBenchmark**, a benchmark of 173 queries across 132 real-world datasets, curated from 79 peer-reviewed research papers and three widely-used causal-inference textbooks. For each query a system must produce (i) a structured identification specification that names the strategy, the treatment, outcome, and control variables, and all design-specific elements, and (ii) a point estimate with a standard error. By scoring these two components separately, our benchmark enables granular diagnosis: it distinguishes failures in causal reasoning from errors in numerical execution. Baseline results with a state-of-the-art LLM show that, while the model correctly identifies the high-level strategy in 79% of cases, full identification-specification correctness drops to only 34%, revealing that the bottleneck lies in the nuanced details of research design rather than in computation. CausalReasoningBenchmark[1] is publicly available on Hugging Face and is designed to foster the development of more robust automated causal-inference systems.

1 Introduction

LLMs and LLM-based agents are increasingly used for causal reasoning over observational data. However, the evaluation of these systems often falls short of the rigor required for real-world applications. A common practice is to assess a model's performance based solely on a single numerical output—typically an effect estimate such as the Average Treatment Effect (ATE). This single-output evaluation is limited because it conflates two distinct steps that are central to any empirical causal analysis.

The first step is **identification**: a conceptual exercise in which the analyst determines whether a causal quantity of interest is recoverable from the available data, given a set of assumptions about the data-generating process. This requires specifying a valid research design—often called an *identification strategy*—such as an Instrumental Variable (IV) design, a Regression Discontinuity Design (RDD), or a Difference-in-Differences (DiD) design, and defining all of its necessary components (e.g., the instrument, the running variable and cutoff, or the time and group indices). The second step is **estimation**: a numerical exercise in which the identified strategy is implemented on a finite data sample to compute a point estimate of the causal effect and to quantify the uncertainty around that estimate.

Existing benchmarks typically collapse these two steps into a single score, making it impossible to diagnose the source of errors. Did the model fail because it chose an invalid identification strategy, or did it implement a valid strategy incorrectly? Furthermore, many benchmarks rely on synthetic or simplified data, which may not reflect the complexities of real-world empirical research: missing data, ambiguous variable definitions, and study-specific design choices.

To address these gaps, we introduce **CausalReasoningBenchmark**, a benchmark for evaluating automated causal reasoning systems. Our main contributions are:

(1) **A curated real-world benchmark.** We curate 173 queries over 132 unique datasets from 79 peer-reviewed research papers and three causal-inference textbooks.
(2) **A structured identification schema.** The benchmark requires agents to specify not only the design family, but also the estimand, treatment, outcome, controls, and design-specific fields required for IV, RDD, DiD, conditional-exogeneity, and RCT designs.
(3) **Disentangled evaluation of identification and estimation** We provide gold identification specifications, reference estimation scripts, and an evaluator that separately scores research-design correctness and numerical estimation accuracy.
(4) **Standardized estimation scripts.** We provide gold-standard estimation code for every query, allowing failures in identification to be isolated from failures in implementation.

The rest of this paper is organized as follows. Section 2 reviews related benchmarks and agents. Section 3 motivates the separation of identification and estimation. Section 4 describes the CausalReasoningBenchmark dataset. Section 5 defines the evaluation task. Section 6 presents baseline results, including a qualitative analysis of identification errors. Sections 7–9 cover hosting, limitations, and conclusions; detailed strategy definitions, full metrics, prompt templates, a sample query, and the paper list are provided in the appendices.

2 Related Work

We situate CausalReasoningBenchmark relative to two lines of work: benchmarks for causal reasoning and LLM-based causal-inference agents.

[1] https://huggingface.co/datasets/syrgkanislab/CausalReasoningBenchmark

The authors used ChatGPT and Manus as research and writing assistants in preparing this manuscript. All interpretations, conclusions, and any errors remain solely the responsibility of the authors.

Benchmarks for Causal Reasoning. Liu et al. [65] introduce **QRData**, a benchmark of quantitative reasoning tasks over spreadsheet-style data, including some causal estimation problems. While QRData tests a broad range of data-analysis skills, it does not focus on the specific identification strategies used in observational studies and does not evaluate identification separately from estimation. Zhou et al. [91] present **CausalBench**, which covers causal graph identification, counterfactual reasoning, and statistical estimation from text and tables. CausalBench is valuable for testing general causal reasoning, but it does not require systems to produce a full identification specification for quasi-experimental designs. Lee et al. [61] build a benchmark by extracting validated, but not quantitative, cause–effect relations from economics and policy papers. Their dataset includes common designs such as IV, DiD, and RDD, but—like the others—it primarily evaluates the final effect estimate, making it difficult to distinguish between identification and estimation errors.

LLM-Based Causal-Inference Agents. Several agent-based systems have been developed to automate parts of the causal inference workflow. **CATE-B** [2] is an LLM co-pilot that constructs directed acyclic graphs (DAGs), selects adjustment sets, and suggests estimators. **ORCA** [12] connects LLMs to causal inference libraries (e.g., `DoWhy`) to load data, fit models, and summarize results. In the biomedical domain, **MRAgent** [89] automates Mendelian randomization by selecting instruments from the literature and analyzing GWAS datasets. These systems are typically evaluated on internal or synthetic tasks. A separate line of work focuses on causal reasoning over graphical models. Jin et al. [52] introduce **CLadder**, a benchmark for formal causal reasoning on synthetic graphs, testing aspects like identifying confounding bias. Jin et al. [51] propose **Corr2Cause**, which tasks models with inferring causal relationships from correlational statements. Sheth et al. [83] present **CausalGraph2LLM**, a large-scale benchmark with over 700k queries on diverse causal graphs. While these benchmarks are crucial for evaluating graph-based and counterfactual reasoning, they do not focus on the quasi-experimental designs common in applied empirical research.

CausalReasoningBenchmark provides a challenging, external evaluation suite derived from peer-reviewed research, with a focus on disentangling identification from estimation.

Table 5 summarizes the key differences between CausalReasoningBenchmark and the most closely related benchmarks.

3 Why Separate Identification from Estimation?

Separating identification from estimation mirrors the structure of empirical causal analysis. In applied research, identification is where the core intellectual contribution resides: it requires understanding the data-generating process, articulating the assumptions under which a causal quantity is recoverable, and specifying all the components of a valid research design. Estimation, by contrast, is largely a technical exercise: given a correctly specified design, the choice of estimator (e.g., two-stage least squares for IV, local polynomial regression for RDD, or a two-way fixed-effects model for DiD) is often well-understood and can even be automated.

This distinction has practical consequences for evaluation. Consider a model that correctly identifies an IV design and names the right instrument, treatment, and outcome, but makes a coding error in the two-stage least squares implementation. Under a single-score evaluation, this model would receive the same score as one that misidentifies the entire research design. By scoring identification and estimation separately, CausalReasoningBenchmark can distinguish between these two very different failure modes.

Moreover, the identification specification itself is a rich, structured object that can be evaluated along multiple dimensions. For example, a model might correctly identify the strategy (IV) and the instrument, but fail to exclude a post-treatment variable from the control set, a critical error that would bias the estimate. Our evaluation framework captures this level of detail, as described in Section 5.

4 The CausalReasoningBenchmark Dataset

CausalReasoningBenchmark is designed to evaluate an agent's ability to correctly specify and execute a causal analysis. It focuses on the canonical research designs used in observational studies: Instrumental Variables (IV), Regression Discontinuity (RD), Difference-in-Differences (DiD), Conditional Exogeneity (selection on observables), and Randomized Controlled Trials (RCT). The benchmark consists of 173 queries over 132 datasets (see Tables 1–2; source-group totals are in Appendix Table 6). Each query includes:

- A natural-language causal question.
- A dataset in CSV format.
- A metadata file describing the variables and providing study context.
- A gold-standard solution, including a detailed identification specification (as a JSON object) and a reference estimation script (in Python or R).

The dataset is sourced from two main categories, described below.

Research Papers. We curated 120 queries from 79 papers, drawing from three large-scale reanalysis studies in political science:

- **IV**: Lal et al. [58] provide a replication of 67 instrumental variable studies.
- **RDD**: Stommes et al. [85] re-evaluate 44 regression discontinuity designs.
- **DiD**: Chiu et al. [9] conduct a reanalysis of 62 difference-in-differences studies.

We selected cases from these corpora where the original paper presented a clear and defensible identification strategy, and where the documentation was sufficient to reconstruct the analysis. The research-paper subset spans three top political science journals, providing a diverse set of real-world causal problems.

Textbook and Instructional Collections. To include classic and pedagogical examples, we added 53 queries from three popular causal inference textbooks:

- *Causal Inference: The Mixtape* [20]
- *The Effect: An Introduction to Research Design and Causality* [49]
- *Causal Inference: What If* [45]

Several of these examples also appeared in *causaldata* R-package, [50] and *QR Dataset* [65] which served as a starting point for us. The textbook subset complements the research-paper subset by

Table 1: CausalReasoningBenchmark composition by identification strategy.

Identification strategy	#queries	#datasets
Difference-in-Differences	67	37
Regression Discontinuity	44	39
Instrumental Variable	22	22
Conditional Exogeneity	39	39
RCT	1	1

Table 2: Query counts by source group and identification strategy.

Source group	DiD	RDD	IV	Cond. Exog.	RCT
Research papers	62	39	19	0	0
Textbook	5	5	3	39	1

providing well-documented examples with clear pedagogical intent, and by adding coverage of Conditional Exogeneity designs (39 queries) that are not represented in the research-paper subset (Table 2).

5 Evaluation Task and Metrics

5.1 Task Definition

For each query in the benchmark, an agent is provided with the following inputs:

- **Question**: A causal query in natural language.
- **Dataset**: A CSV file containing the data.
- **Metadata**: A text file with column descriptions and study context.

The agent must produce two outputs:

(1) **Identification Specification**: A structured JSON object that adheres to the schema described in Section C, detailing the chosen identification strategy and all its components.
(2) **Estimation Output**: A point estimate of the causal effect (`effect_estimate`) and its standard error (`standard_error`).

6 LLM Baseline Evaluation

To demonstrate the utility of our benchmark, we evaluated a simple LLM-agent baseline. The baseline was run with `gpt-5.3` with reasoning. For each query, the runner supplied the model with the natural-language causal question, an inline metadata description, and the corresponding CSV dataset. For the primary execution path, the CSV file was uploaded as a file attachment and made available to the API's Python code-interpreter environment. The model was instructed to inspect the data, write and execute Python code for the estimate and standard error, and return a single JSON object containing both a schema-conformant identification specification and the estimation code. The returned identification payload was validated against our output schema before being scored. The full prompt template is shown in Appendix E.

Table 3: Aggregate evaluation of the GPT-5.3 baseline on all 173 queries. Identification metrics are exact-match or set-based checks against the gold specification; estimation metrics compare the returned effect and uncertainty to the gold solution. Values in brackets denote the interquartile range.

Metric	Value
Identification Metrics	
Strategy correct	79.2% (137/173)
Causal quantity correct	73.4% (127/173)
Treatments correct	86.1% (149/173)
Outcomes correct	93.6% (162/173)
Minimal controlling set included	77.5% (134/173)
Post-treatment set excluded	89.6% (155/173)
Controls correct	67.6% (117/173)
Strategy-specific fields correct	89.0% (154/173)
Identification spec correct (all checks)	**34.1% (59/173)**
Estimation Metrics	
Median absolute error $\lvert\widehat{\tau} - \tau^{\star}\rvert$	0.070 [0.010, 0.421]
Median percentage error	18.1%
Median CI Jaccard overlap	0.57
CI Overlap	88 %
Estimate Within gold CI	82 %
Null Hypothesis Agreement	80 %
Opposite Direction Flag	1.73 %

6.1 Aggregate Results

Table 3 shows the aggregate performance of the baseline across all 173 queries. The model correctly identifies the high-level strategy in 79.2% of cases and the outcome variables in 93.6% of cases. However, performance drops sharply on more nuanced aspects of identification: causal quantity is correct in only 73.4% of cases, and the overall identification specification is fully correct in only 34.1% of cases. This gap between high-level strategy recognition and full specification correctness is the central finding of our baseline evaluation, and it validates the design of CausalReasoningBenchmark: a single-score evaluation based on the final estimate would have obscured this important distinction.

6.2 Per-Strategy Breakdown

Table 4 provides a per-strategy breakdown of the baseline results. The model performs best on Regression Discontinuity queries, where the strategy is correct in 95.5% of cases, though overall identification correctness is significantly lower. Second, Difference-in-Differences queries prove more challenging, particularly in specifying the correct time and group variables in panel data. Third, Instrumental Variable queries show the largest gap between strategy-level correctness and full specification correctness, suggesting that the model struggles with the nuances of IV designs (e.g., identifying the correct instrument).

Table 4: Per-strategy breakdown of the GPT-5.3 baseline. "Strategy" = fraction with correct strategy label; "Full ID" = fraction with fully correct identification specification; "Med. %Err" = median percentage error on the effect estimate.

Design	#Queries	Strategy (%)	Full ID (%)	Med. %Err
Difference-in-Differences	67	92.5	52.2	27.0
Regression Discontinuity	44	95.5	11.4	20.0
Instrumental Variable	22	77.3	31.8	48.5
Conditional Exogeneity	39	38.5	28.2	4.0
RCT	1	100.0	100.0	0.0
Overall	**173**	**79.2**	**34.1**	**18.1**

6.3 Analysis

The baseline results reveal several important insights. First, the large gap between strategy-level correctness (79.2%) and full identification correctness (34.1%) confirms that the bottleneck in automated causal reasoning lies not in recognizing the broad category of research design, but in specifying its detailed components. A single final-estimate score would not distinguish these failure modes.

Second, the estimation errors (median 18.1% relative error, median Jaccard overlap of 0.57) are non-trivial but secondary to the identification errors. In many cases, the model produces a reasonable estimate even when the identification specification is incorrect, because it may still use a plausible (but not gold-standard) approach. This further underscores the importance of evaluating identification separately.

Third, the per-strategy breakdown reveals that different designs pose different challenges. RDD queries are relatively easy to identify but may still have estimation errors due to bandwidth selection. DiD queries require understanding temporal structure and group assignments. IV queries demand the identification of a valid instrument—a task that requires deep domain knowledge.

7 Hosting and Maintenance

CausalReasoningBenchmark is publicly available on Hugging Face Datasets.[1] All associated code for evaluation, including the evaluator script and the gold-standard estimation scripts, is available in the same repository.

Licensing and Access. The benchmark metadata, evaluation code, and gold-standard identification specifications are released under the MIT license. The underlying datasets are redistributed under the terms of their original licenses; we provide attribution and licensing information for each dataset.

Maintenance Plan. We plan to update the benchmark periodically: adding queries as new reanalysis studies appear, incorporating additional designs (synthetic control, event studies), and revising metrics in response to feedback. We welcome contributions from the community.

8 Limitations and Future Work

CausalReasoningBenchmark has several limitations that we plan to address in future work.

[1] https://huggingface.co/datasets/syrgkanislab/CausalReasoningBenchmark

Domain Coverage. The research-paper subset is drawn entirely from political science, reflecting the availability of large-scale reanalysis studies in that field. While the textbook subset provides some cross-domain coverage, the benchmark would benefit from the inclusion of datasets from economics, epidemiology, and other fields where causal inference is central.

Strategy Coverage. The current benchmark focuses on five identification strategies. Important designs such as synthetic control methods, event studies, and regression kink designs are not yet covered. We plan to expand the strategy coverage in future releases.

Single Gold Standard. For each query, we provide a single gold-standard identification specification. In practice, there may be multiple defensible identification strategies for a given dataset and question. Future work could explore evaluation frameworks that accommodate multiple valid specifications.

Estimation Sensitivity. The gold-standard estimates are produced by specific estimation scripts. Different but equally valid estimation choices (e.g., different bandwidth selectors for RDD, different standard error clustering for DiD) could produce different estimates.

Scale. With 173 queries, CausalReasoningBenchmark is smaller than some existing benchmarks. However, we prioritize quality and depth of evaluation over quantity: each query requires a full identification specification, a python script to produce the causal estimate. We plan to expand the benchmark over time.

9 Conclusion

CausalReasoningBenchmark provides a new, challenging, and realistic benchmark for evaluating automated causal reasoning systems. By separating the evaluation of identification and estimation, it offers a more nuanced view of model capabilities than existing benchmarks. Our baseline results demonstrate that state-of-the-art LLMs struggle with the detailed specification of causal research designs, even when they can correctly identify the broad design family. This finding highlights the need for more sophisticated reasoning capabilities in automated causal inference systems. We hope that CausalReasoningBenchmark will help with the development of more robust and reliable AI systems for causal inference, and we welcome contributions from the research community.

10 Acknowledgement

Vasilis Syrgkanis, Ayush Sawarni and Jiyuan Tan were supported by NSF Award IIS-2337916.

References

[1] Kenichi Ariga. 2015. Incumbency Disadvantage under Electoral Rules with Intraparty Competition: Evidence from Japan. *The Journal of Politics* (2015). doi:10.1086/681718

[2] Jeroen Berrevoets, Julianna Piskorz, Robert Davis, Harry Amad, Jim Weatherall, and Mihaela van der Schaar. 2025. Technical Report: Facilitating the Adoption of Causal Inference Methods Through LLM-Empowered Co-Pilot. *arXiv preprint arXiv:2508.10581* (2025). doi:10.48550/arXiv.2508.10581

[3] Graeme Blair, Darin Christensen, and Valerie Wirtschafter. 2022. How Does Armed Conflict Shape Investment? Evidence from the Mining Sector. *The Journal of Politics* (2022). doi:10.1086/715255

[4] Taylor C. Boas, F. Daniel Hidalgo, and Neal P. Richardson. 2014. The Spoils of Victory: Campaign Donations and Government Contracts in Brazil. *The Journal of Politics* (2014). doi:10.1017/s002238161300145x

[5] David E. Broockman and Timothy J. Ryan. 2015. Preaching to the Choir: Americans Prefer Communicating to Copartisan Elected Officials. *American Journal of Political Science* (2015). doi:10.1111/ajps.12228

[6] Allison Carnegie and Nikolay Marinov. 2017. Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment. *American Journal of Political Science* (2017). doi:10.1111/ajps.12289

[7] Jamie L. Carson and Joel Sievert. 2017. Congressional Candidates in the Era of Party Ballots. *The Journal of Politics* (2017). doi:10.1086/688077

[8] Devin Caughey, Christopher Warshaw, and Yiqing Xu. 2017. Incremental Democracy: The Policy Effects of Partisan Control of State Government. *The Journal of Politics* (2017). doi:10.1086/692669

[9] Albert Chiu, Xingchen Lan, Ziyi Liu, and Yiqing Xu. 2026. Causal Panel Analysis under Parallel Trends: Lessons from a Large Reanalysis Study. *American Political Science Review* 120, 1 (2026), 245–266. doi:10.1017/S0003055425000243

[10] Alberto Chong, Gianmarco León-Ciliotta, Vivian Roza, Martín Valdivia, and Gabriela Vega. 2018. Urbanization Patterns, Information Diffusion, and Female Voting in Rural Paraguay. *American Journal of Political Science* (2018). doi:10.1111/ajps.12404

[11] Darin Christensen and Francisco Garfias. 2021. The Politics of Property Taxation: Fiscal Infrastructure and Electoral Incentives in Brazil. *The Journal of Politics* (2021). doi:10.1086/711902

[12] Joanie Hayoun Chung, Chaemyung Lim, Sumin Lee, Songseong Kim, and Sungbin Lim. 2025. ORCA: ORchestrating Causal Agent. *arXiv preprint arXiv:2508.21304* (2025). doi:10.48550/arXiv.2508.21304

[13] Andrew J. Clarke. 2020. Party Sub-Brands and American Party Factions. *American Journal of Political Science* (2020). doi:10.1111/ajps.12504

[14] Amanda Clayton and Pär Zetterberg. 2018. Quota Shocks: Electoral Gender Quotas and Government Spending Priorities Worldwide. *The Journal of Politics* (2018). doi:10.1086/697251

[15] Alexander Coppock and Donald P. Green. 2015. Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities. *American Journal of Political Science* (2015). doi:10.1111/ajps.12210

[16] Alexander Coppock and Donald P. Green. 2015. Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities. *American Journal of Political Science* (2015). doi:10.1111/ajps.12210

[17] Alexander Coppock and Donald P. Green. 2016. Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities. *American Journal of Political Science* 60, 4 (2016), 1044–1062. doi:10.1111/ajps.12210

[18] Benjamin Hans Creutzfeldt. 2016. China y EE. UU. en Latinoamérica. *Revista Científica General José María Córdova* (2016). doi:10.21830/19006586.1

[19] Kevin Croke, Guy Grossman, Horacio A. Larreguy, and John Marshall. 2016. Deliberate Disengagement: How Education Can Decrease Political Participation in Electoral Authoritarian Regimes. *American Political Science Review* (2016). doi:10.1017/s0003055416000253

[20] Scott Cunningham. 2021. *Causal Inference: The Mixtape.* Yale University Press, London. https://mixtape.scunning.com/

[21] Carl Dahlström and Mikael Holmgren. 2023. Loyal Leaders, Affluent Agencies: The Budgetary Implications of Political Appointments in the Executive Branch. *The Journal of Politics* (2023). doi:10.1086/717756

[22] Justin de Benedictis-Kessner. 2018. Off-Cycle and Out of Office: Election Timing and the Incumbency Advantage. *The Journal of Politics* (2018). doi:10.1086/694396

[23] Greg Distelhorst and Richard M. Locke. 2018. Does Compliance Pay? Social Standards and Firm-level Trade. doi:10.31235/osf.io/tcrhq

[24] Paul Castañeda Dower, Evgeny Finkel, Scott Gehlbach, and Steven Nafziger. 2018. Collective action and representation in autocracies: Evidence from Russia's great reforms. *American Political Science Review* 112, 1 (2018), 125–147.

[25] Laurel Eckhouse. 2021. Metrics Management and Bureaucratic Accountability: Evidence from Policing. *American Journal of Political Science* (2021). doi:10.1111/ajps.12661

[26] Andrew C. Eggers and Jens Hainmueller. 2009. MPs for Sale? Returns to Office in Postwar British Politics. *American Political Science Review* (2009). doi:10.1017/s0003055409990190

[27] Andrew C. Eggers and Arthur Spirling. 2017. Incumbency Effects and the Strength of Party Preferences: Evidence from Multiparty Elections in the United Kingdom. *The Journal of Politics* (2017). doi:10.1086/690617

[28] Robert S. Erikson, Olle Folke, and James M. Snyder. 2015. A Gubernatorial Helping Hand? How Governors Affect Presidential Elections. *The Journal of Politics* (2015). doi:10.1086/680186

[29] Jane Esberg and Alexandra A. Siegel. 2022. How Exile Shapes Online Opposition: Evidence from Venezuela. *American Political Science Review* (2022). doi:10.1017/s0003055422001290

[30] Jeremy Ferwerda and Nicholas L. Miller. 2014. Political Devolution and Resistance to Foreign Rule: A Natural Experiment. *American Political Science Review* (2014). doi:10.1017/s0003055414000240

[31] Olle Folke and James M. Snyder. 2012. Gubernatorial Midterm Slumps. *American Journal of Political Science* (2012). doi:10.1111/j.1540-5907.2012.00599.x

[32] Alexander Fouirnaies and Andrew B. Hall. 2014. The Financial Incumbency Advantage: Causes and Consequences. *The Journal of Politics* (2014). doi:10.1017/s0022381614000139

[33] Adriane Fresh. 2018. The Effect of the Voting Rights Act on Enfranchisement: Evidence from North Carolina. *The Journal of Politics* (2018). doi:10.1086/697592

[34] Francisco Garfias. 2019. Elite Coalitions, Limited Government, and Fiscal Capacity Development: Evidence from Bourbon Mexico. *The Journal of Politics* (2019). doi:10.1086/700105

[35] Alan S. Gerber, Gregory A. Huber, and Ebonya Washington. 2010. Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment. *American Political Science Review* (2010). doi:10.1017/s0003055410000407

[36] Jacob M. Grumbach. 2022. Laboratories of Democratic Backsliding. *American Political Science Review* (2022). doi:10.1017/s0003055422000934

[37] Jacob M. Grumbach and Charlotte Hill. 2022. Rock the Registration: Same Day Registration Increases Turnout of Young Voters. *The Journal of Politics* (2022). doi:10.1086/714776

[38] Jacob M. Grumbach and Alexander Sahn. 2019. Race and Representation in Campaign Finance. *American Political Science Review* (2019). doi:10.1017/s0003055419000637

[39] Laurenz Guenther. 2024. Correcting Misperceptions Can Increase Anti-Immigration Attitudes. doi:10.2139/ssrn.5001788

[40] Jens Hainmueller and Dominik Hangartner. 2014. Does Direct Democracy Hurt Immigrant Minorities? Evidence from Naturalization Decisions in Switzerland. *SSRN Electronic Journal* (2014). doi:10.2139/ssrn.2503141

[41] Andrew B. Hall. 2015. What Happens When Extremists Win Primaries? *American Political Science Review* (2015). doi:10.1017/s0003055414000641

[42] Andrew B. Hall and Daniel M. Thompson. 2018. Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in US Elections. *American Political Science Review* (2018). doi:10.1017/s0003055418000023

[43] Michael Hankinson and Asya Magazinnik. 2023. The Supply-Equity Trade-Off: The Effect of Spatial Representation on the Local Housing Supply. *The Journal of Politics* (2023). doi:10.1086/723818

[44] Andrew Healy and Neil Malhotra. 2013. Childhood Socialization and Political Attitudes: Evidence from a Natural Experiment. *The Journal of Politics* (2013). doi:10.1017/s0022381613000996

[45] Miguel A. Hernán and James M. Robins. 2020. *Causal Inference: What If.* Chapman & Hall/CRC, Boca Raton. https://miguelhernan.org/whatifbook

[46] F. Daniel Hidalgo and Simeon Nichter. 2015. Voter Buying: Shaping the Electorate through Clientelism. *American Journal of Political Science* (2015). doi:10.1111/ajps.12214

[47] Shigeo Hirano, Jaclyn Kaslovsky, Michael P. Olson, and James M. Snyder. 2022. The Growth of Campaign Advertising in the United States, 1880–1930. *The Journal of Politics* (2022). doi:10.1086/719008

[48] John B. Holbein and D. Sunshine Hillygus. 2015. Making Young Voters: The Impact of Preregistration on Youth Turnout. *American Journal of Political Science* (2015). doi:10.1111/ajps.12177

[49] Nick Huntington-Klein. 2022. *The Effect: An Introduction to Research Design and Causality.* CRC Press, Taylor & Francis Group, Boca Raton.

[50] Nick Huntington-Klein and Malcolm Barrett. [n. d.]. causaldata: Example Data Sets for Causal Inference Textbooks, 2021. *URL https://github. com/nickch-k/causaldata. R package version 0.1* 4 ([n. d.]).

[51] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. Can Large Language Models Infer Causation from Correlation? *arXiv preprint arXiv:2306.05836* (2023). doi:10.48550/arXiv.2306.05836

[52] Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. 2023. CLadder: Assessing Causal Reasoning in Language Models. In *Advances in Neural Information Processing Systems*, Vol. 36. 31038–31065.

[53] Mitchell Kilborn and Arjun Vishwanath. 2021. Public Money Talks Too: How Public Campaign Financing Degrades Representation. *American Journal of Political Science* (2021). doi:10.1111/ajps.12625

[54] Jeong Hyun Kim. 2019. Direct democracy and women's political engagement. *American Journal of Political Science* 63, 3 (2019), 594–610.

[55] Marko Klašnja and Rocío Titiunik. 2017. The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability. *American Political Science Review* (2017). doi:10.1017/s0003055416000575

[56] Mary Kroeger and Maria Silfa. 2023. Motivated Corporate Political Action: Evidence from an SEC Experiment. *The Journal of Politics* (2023). doi:10.1086/723998

[57] Nicholas Kuipers and Alexander Sahn. 2022. The Representational Consequences of Municipal Civil Service Reform. *American Political Science Review* (2022). doi:10.1017/s0003055422000521

[58] Apoorva Lal, Mackenzie Lockhart, Yiqing Xu, and Ziwen Zu. 2024. How Much Should We Trust Instrumental Variable Estimates in Political Science? Practical Advice Based on 67 Replicated Studies. *Political Analysis* 32, 4 (2024), 521–540. doi:10.1017/pan.2024.2

[59] Alan Lambert, Fade Eadeh, and Emily Hanson. 2018. Anger and its Consequences for Judgment and Behavior: Recent Developments in Social and Political Psychology. doi:10.31234/osf.io/svcux_v1

[60] Audrey Latura and Ana Catalano Weeks. 2022. Corporate Board Quotas and Gender Equality Policies in the Workplace. *American Journal of Political Science* (2022). doi:10.1111/ajps.12709

[61] Donggyu Lee, Sungwon Park, Yerin Hwang, Hyoshin Kim, Hyunwoo Oh, Jungwon Kim, Meeyoung Cha, Sangyoon Park, and Jihee Kim. 2025. Benchmarking LLM Causal Reasoning with Scientifically Validated Relationships. *arXiv preprint arXiv:2510.07231* (2025). doi:10.48550/arXiv.2510.07231

[62] Yphtach Lelkes, Gaurav Sood, and Shanto Iyengar. 2015. The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect. *American Journal of Political Science* (2015). doi:10.1111/ajps.12237

[63] Amy E. Lerman and Katherine T. McCabe. 2017. Personal Experience and Public Opinion: A Theory and Test of Conditional Policy Feedback. *The Journal of Politics* (2017). doi:10.1086/689286

[64] Steven Liao. 2023. The Effect of Firm Lobbying on High-Skilled Visa Adjudication. *The Journal of Politics* (2023). doi:10.1086/723984

[65] Zeqi Liu, Ke Li, Yu Cheng, Lichao Xue, Xuhui Fan, Yue Chen, Aobo Yang, Kun Ma, Zhiyuan Zhao, Peng Jiang, Yuxiang Zhou, Hao Wang, Jianxing Yu, Qian Zhang, Yang Liu, and Yangfeng Ji. 2024. Are LLMs Capable of Data-based Statistical and Causal Reasoning? Benchmarking Advanced Quantitative Reasoning with Data. In *Findings of the Association for Computational Linguistics: ACL 2024*. 9215–9235. doi:10.18653/v1/2024.findings-acl.548

[66] Beatriz Magaloni, Edgar Franco-Vivanco, and Vanessa Melo. 2020. Killing in the Slums: Social Order, Criminal Governance, and Police Violence in Rio de Janeiro. *American Political Science Review* (2020). doi:10.1017/s0003055419000856

[67] Wayde Z. C. Marsh. 2022. Trauma and Turnout: The Political Consequences of Traumatic Events. *American Political Science Review* (2022). doi:10.1017/s0003055422001010

[68] Gwyneth H. McClendon. 2013. Social Esteem and Participation in Contentious Politics: A Field Experiment at an LGBT Pride Rally. *American Journal of Political Science* (2013). doi:10.1111/ajps.12076

[69] Michael McDevitt and Steven Chaffee. 2002. From Top-Down to Trickle-Up Influence: Revisiting Assumptions About the Family in Political Socialization. *Political Communication* (2002). doi:10.1080/01957470290055501

[70] Marc Meredith. 2013. Exploiting Friends-and-Neighbors to Estimate Coattail Effects. *American Political Science Review* (2013). doi:10.1017/s0003055413000439

[71] Gareth Nellis and Niloufer Siddiqui. 2017. Secular Party Rule and Religious Violence in Pakistan. *American Political Science Review* (2017). doi:10.1017/s0003055417000491

[72] Lucas M. Novaes. 2017. Disloyal Brokers and Weak Parties. *American Journal of Political Science* (2017). doi:10.1111/ajps.12331

[73] Ana L. De La O. 2012. Do Conditional Cash Transfers Affect Electoral Behavior? Evidence from a Randomized Experiment in Mexico. *American Journal of Political Science* (2012). doi:10.1111/j.1540-5907.2012.00617.x

[74] Agustina S. Paglayan. 2022. Education or Indoctrination? The Violent Origins of Public School Systems in an Era of State-Building. *American Political Science Review* (2022). doi:10.1017/s0003055422000247

[75] Maxwell Palmer and Benjamin Schneer. 2016. Capitol Gains: The Returns to Elected Office from Corporate Board Directorships. *The Journal of Politics* (2016). doi:10.1086/683206

[76] Julia A. Payson. 2020. The Partisan Logic of City Mobilization: Evidence from State Lobbying Disclosures. *American Political Science Review* (2020). doi:10.1017/s0003055420000118

[77] Jan H. Pierskalla and Audrey Sacks. 2018. Unpaved Road Ahead: The Consequences of Election Cycles for Capital Expenditures. *The Journal of Politics* (2018). doi:10.1086/694547

[78] Nico Ravanilla, Renard Sexton, and Dotan Haim. 2022. Deadly Populism: How Local Political Outsiders Drive Duterte's War on Drugs in the Philippines. *The Journal of Politics* (2022). doi:10.1086/715257

[79] Miguel R. Rueda. 2016. Small Aggregates, Big Manipulation: Vote Buying Enforcement and Collective Monitoring. *American Journal of Political Science* (2016). doi:10.1111/ajps.12260

[80] Jerome Schafer, Enrico Cantoni, Giorgio Bellettini, and Carlotta Berti Ceroni. 2022. Making unequal democracy work? The effects of income on voter turnout in Northern Italy. *American Journal of Political Science* 66, 3 (2022), 745–761.

[81] Eric Schickler, Kathryn Pearson, and Brian D. Feinstein. 2010. Congressional Parties and Civil Rights Politics from 1933 to 1972. *The Journal of Politics* (2010). doi:10.1017/s0022381610000095

[82] Sophie Schuit and Jon C Rogowski. 2017. Race, representation, and the voting rights act. *American Journal of Political Science* 61, 3 (2017), 513–526.

[83] Ishaan Sheth, Zheng Yuan, Keyi Fu, et al. 2025. CausalGraph2LLM: Evaluating LLMs for Causal Queries. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 2076–2098. doi:10.18653/v1/2025.findings-naacl.110

[84] Leah C. Stokes. 2015. Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy. *American Journal of Political Science* (2015). doi:10.1111/ajps.12220

[85] Drew Stommes, Peter M. Aronow, and Fredrik Sävje. 2023. On the Reliability of Published Findings Using the Regression Discontinuity Design in Political Science. *Research & Politics* 10, 2 (2023), 20531680231166457. doi:10.1177/20531680231166457

[86] David Szakonyi. 2016. Businesspeople in Elected Office: Identifying Private Benefits from Firm-Level Returns. *SSRN Electronic Journal* (2016). doi:10.2139/ssrn.2844901

[87] Anjali Thomas. 2018. Targeting Ordinary Voters or Political Elites? Why Pork Is Distributed Along Partisan Lines in India. *American Journal of Political Science* (2018). doi:10.1111/ajps.12374

[88] Jessica Trounstine. 2020. The Geography of Inequality: How Land Use Regulation Produces Segregation. *American Political Science Review* (2020). doi:10.1017/s0003055419000844

[89] Wei Xu, Yuncheng Zhang, Rui Guo, Xiaowei Wang, Qian Liu, Xiaoxiao Li, et al. 2025. MRAgent: an LLM-based automated agent for causal knowledge discovery in disease via Mendelian randomization. *Briefings in Bioinformatics* 26, 2 (2025), bbaf140. doi:10.1093/bib/bbaf140

[90] Qi Zhang, Dong Zhang, Mingxing Liu, and Victor Shih. 2021. Elite Cleavage and the Rise of Capitalism under Authoritarianism: A Tale of Two Provinces in China. *The Journal of Politics* (2021). doi:10.1086/711131

[91] Yuan Zhou, Zifan Wang, Chenyang Gao, Xiaocheng Li, Junfeng Lou, Bo Li, and Jian Tang. 2024. CausalBench: A Comprehensive Benchmark for Causal Learning Capability of LLMs. *arXiv preprint arXiv:2404.06349* (2024). doi:10.48550/arXiv.2404.06349

A Related Benchmark Comparison

Table 5 summarizes the key differences between CausalReasoningBenchmark and the most closely related benchmarks.

Table 5: Comparison of CausalReasoningBenchmark with related benchmarks. "ID eval" indicates whether identification is evaluated separately from estimation. "Real data" indicates whether the benchmark uses real-world (non-synthetic) datasets. "Quant. eval" means evaluating causal effect estimation from data (e.g., ATE / ATT / LATE / CATE). "Design-specific" indicates whether the benchmark requires specification of design-specific elements (e.g., instruments, running variables).

Benchmark	# Queries	Real data	ID eval	Quant. eval	Design-specific	Designs covered
QRData [65]	899	✓		Partial		Mixed
CausalBench [91]	495	Partial				Mixed
CLadder [52]	6.6k		✓			Graph-based
Corr2Cause [51]	413k		✓			Graph-based
CausalGraph2LLM [83]	700k+		✓			Graph-based
CausalReasoning Benchmark	**173**	✓	✓	✓	✓	IV, RDD, DiD, CE, RCT

B Additional Dataset Composition

Dataset Composition. Tables 6, 1, and 2 provide a detailed breakdown of the benchmark's composition. Table 6 shows the split between research papers and textbooks. Table 1 shows the distribution across identification strategies: DiD is the most common (67 queries), followed by RDD (44), Conditional Exogeneity (39), IV (22), and RCT (1). Table 2 reveals that the research-paper subset is dominated by DiD, RDD, and IV, while the textbook subset provides the bulk of the Conditional Exogeneity examples. The research-paper subset spans three top political science journals—*The Journal of Politics*, *American Journal of Political Science*, and *American Political Science Review*—which account for the vast majority of the research-paper queries.

Table 6: CausalReasoningBenchmark queries and datasets by source group.

Source group	#queries	#datasets
Research papers	120	79
Textbook	53	53
Total	173	132

C Identification Strategies

A key feature of CausalReasoningBenchmark is that it requires systems to produce a *structured identification specification* for each query. This section provides a brief formal description of each identification strategy covered by the benchmark, along with the specific fields that the system must specify.

C.1 Instrumental Variables (IV)

An instrumental variable design exploits an exogenous source of variation (the *instrument*, Z) that affects the treatment (D) but has no direct effect on the outcome (Y) except through D. The key assumptions are: (i) *relevance*: Z is correlated with D; (ii) the *exclusion restriction*: Z affects Y only through D; and (iii) *independence*: Z is independent of the potential outcomes (unobserved confounders).

In many applications, these assumptions may only be plausible after conditioning on a set of pre-treatment covariates $\mathbf{X}$. The IV assumptions are thus relaxed to hold conditionally: (i) *conditional relevance*: $\mathrm{Cov}(D, Z \mid \mathbf{X}) \neq 0$; (ii) *conditional exclusion*: Z is independent of potential outcomes $Y(d)$ conditional on D and $\mathbf{X}$; and (iii) *conditional independence*: Z is independent of potential outcomes conditional on $\mathbf{X}$ ($Z \perp\!\!\!\perp Y(d) \mid \mathbf{X}$). When these conditions hold, the LATE can be estimated using methods like two-stage least squares with controls. Under the unconditional assumptions, the LATE for compliers is identified as:

$$\mathrm{LATE} = \frac{\mathrm{Cov}(Y, Z)}{\mathrm{Cov}(D, Z)}. \tag{1}$$

Required fields: `strategy` = `Instrumental Variable`; `instrument` (column name(s) of the instrument); `is_encouragement_design` (whether the instrument is a randomized binary encouragement); `treatments`; `outcomes`; `controls`; `causal_quantity` (typically `LATE`).

C.2 Regression Discontinuity (RDD)

A regression discontinuity design exploits a known threshold (the *cutoff*) on a continuous *running variable* (X) that determines treatment assignment. Units just above and just below the cutoff are assumed to be comparable, so the causal effect is identified as the discontinuity in the conditional expectation of the outcome at the cutoff:

$$\tau_{\mathrm{RDD}} = \lim_{x \downarrow c} E[Y \mid X = x] - \lim_{x \uparrow c} E[Y \mid X = x], \tag{2}$$

where c is the cutoff value. In a *sharp* design, treatment is a deterministic function of the running variable; in a *fuzzy* design, the probability of treatment changes discontinuously at the cutoff, and the design is analogous to an IV with the threshold indicator as the instrument.

Required fields: `strategy` = `Regression Discontinuity`; `running_variable` (column name); `cutoff` (numeric threshold); `treatments`; `outcomes`; `controls`; `causal_quantity`.

C.3 Difference-in-Differences (DiD)

A difference-in-differences design compares the change in outcomes over time between a treated group and a control group. The key assumption is *parallel trends*: in the absence of treatment, the average outcomes for the treated and control groups would have followed parallel paths over time.

This assumption can be relaxed to a *conditional parallel trends* assumption, which posits that parallel trends hold after conditioning on a set of pre-treatment covariates $\mathbf{X}$. This allows the baseline trends to differ, as long as they are parallel within strata defined by $\mathbf{X}$. Formally, the assumption is $E[Y(0)_{\mathrm{post}} - Y(0)_{\mathrm{pre}} \mid D = 1, \mathbf{X}] = E[Y(0)_{\mathrm{post}} - Y(0)_{\mathrm{pre}} \mid D = 0, \mathbf{X}]$. Under the unconditional assumption, the Average Treatment Effect on the Treated (ATT) is identified

as:

$$\tau_{\text{DiD}} = \big(E[Y_{\text{post}} \mid D=1] - E[Y_{\text{pre}} \mid D=1]\big) - \big(E[Y_{\text{post}} \mid D=0] - E[Y_{\text{pre}} \mid D=0]\big). \quad (3)$$

Required fields: `strategy` = `Difference-in-Differences`; `time_variable` (column name of the time index); `group_variable` (column name of the unit/group identifier); `treatments`; `outcomes`; `controls`; `causal_quantity` (typically `ATT`).

C.4 Conditional Exogeneity (Selection on Observables)

Under conditional exogeneity, treatment assignment is assumed to be independent of potential outcomes after conditioning on a set of observed covariates $\mathbf{X}$:

$$(Y(0), Y(1)) \perp\!\!\!\perp D \mid \mathbf{X}. \quad (4)$$

This assumption, also known as *unconfoundedness* or *selection on observables*, allows identification of the ATE (or ATT) via regression adjustment, inverse probability weighting, or matching.
Required fields: `strategy` = `Conditional Exogeneity`; `treatments`; `outcomes`; `controls` (the conditioning set $\mathbf{X}$, which must include a minimal sufficient adjustment set); `causal_quantity`.

C.5 Randomized Controlled Trials (RCT)

In a randomized controlled trial, treatment is assigned randomly, so identification is straightforward: the ATE is simply the difference in mean outcomes between the treated and control groups. While RCTs are the gold standard for causal inference, they are included in CausalReasoningBenchmark primarily for completeness (1 query).
Required fields: `strategy` = `RCT`; `treatments`; `outcomes`; `controls`; `causal_quantity`.

D Full Evaluation Metrics

D.1 Identification Metrics

We evaluate the identification specification by comparing it field-by-field with the gold standard. The evaluator checks the following conditions:

- **Strategy**: Exact match of the identification strategy label (e.g., `Instrumental Variable`).
- **Causal quantity**: Exact match of the estimand label (e.g., `ATE`, `LATE`).
- **Treatments and outcomes**: Exact set match of the variable names specified by the agent against the gold standard.
- **Controls**: We check two conditions: (1) the agent's specified controls must be a superset of the gold-standard *minimal sufficient adjustment set*—the smallest set of pre-treatment covariates needed for identification (e.g., to satisfy conditional exogeneity or conditional parallel trends); and (2) the agent's controls must not include any variables from the gold-standard *bad controls* list, which includes post-treatment variables, mediators, and colliders whose inclusion would bias the estimate.
- **Strategy-specific fields**: Depending on the strategy, we require correct specification of all compulsory fields: for IV, the `instrument` list and `is_encouragement_design` flag; for RDD, the `running_variable` and `cutoff`; for DiD, the `time_variable` and `group_variable`.
- **Overall identification correctness**: A binary indicator that is `true` only if *all* of the above checks pass. This is the strictest metric and captures whether the model has fully specified a valid research design.

D.2 Estimation Metrics

Given a predicted effect $\widehat{\tau}$ and standard error $\widehat{\text{SE}}$, and gold-standard values $\tau^\star$ and $\text{SE}^\star$, we form 95% Wald confidence intervals $\text{CI}_{\text{pred}} = [\widehat{\tau} - 1.96\,\widehat{\text{SE}},\ \widehat{\tau} + 1.96\,\widehat{\text{SE}}]$ and $\text{CI}_{\text{gold}} = [\tau^\star - 1.96\,\text{SE}^\star,\ \tau^\star + 1.96\,\text{SE}^\star]$, and compute:

- **Point-estimate error**: Absolute error $|\widehat{\tau} - \tau^\star|$, signed error $\widehat{\tau} - \tau^\star$, and (when $\tau^\star \neq 0$) relative absolute error $\frac{|\widehat{\tau} - \tau^\star|}{|\tau^\star|} \times 100\%$.
- **Estimate within gold CI**: Whether $\widehat{\tau} \in \text{CI}_{\text{gold}}$.
- **Null-hypothesis agreement**: Whether both intervals lead to the same reject/fail-to-reject decision for $H_0 : \tau = 0$.
- **Opposite-direction flag**: Whether both intervals reject H_0 but imply opposite effect signs—a particularly dangerous type of error.
- **Interval overlap (Jaccard)**: We measure the overlap between the two confidence intervals using the Jaccard index:

$$J(\text{CI}_{\text{pred}}, \text{CI}_{\text{gold}}) = \frac{|\text{CI}_{\text{pred}} \cap \text{CI}_{\text{gold}}|}{|\text{CI}_{\text{pred}} \cup \text{CI}_{\text{gold}}|}, \quad (5)$$

  which equals 0 when the intervals are disjoint and 1 when they coincide.
- **CI Overlap**: A binary indicator of whether the predicted and gold-standard confidence intervals overlap, i.e., $\mathbf{1}[\text{CI}_{\text{pred}} \cap \text{CI}_{\text{gold}} \neq \emptyset]$.
- **Standard-error gap**: $|\widehat{\text{SE}} - \text{SE}^\star|$ and the relative gap $\frac{|\widehat{\text{SE}} - \text{SE}^\star|}{\text{SE}^\star}$.

Auto-rescaling. A common source of spurious estimation error is a unit mismatch (e.g., an effect reported in percentage points vs. proportions). For example, if the gold-standard effect is 0.05 (a 5 percentage point increase) and the model predicts 5.0, a naive error calculation would be enormous. To mitigate this, the evaluator can optionally rescale the predicted effect and standard error by a multiplicative factor from a fixed candidate set (e.g., $\{0.01, 0.1, 10, 100\}$). The evaluator selects the factor that minimizes the absolute error. In the example above, multiplying the prediction of 5.0 by 0.01 yields 0.05, which perfectly matches the gold standard. The evaluation would proceed with this rescaled value. This ensures that trivial unit-conversion errors do not dominate the estimation metrics.

E Prompt Template

LLM prompt template

```
SYSTEM:
You are a meticulous causal inference research assistant. You can
      read
structured metadata, inspect CSV datasets, craft valid identification
strategies, and run Python code to obtain causal effect estimates.

USER:
You are provided with a causal question, an inline metadata
      description,
and a CSV dataset (available as a file in your Python environment).

CAUSAL QUESTION:
{...}

METADATA ABOUT THE DATASET:
{...}

TASKS:
1) Read metadata to understand variables/context.
2) Load and analyze the CSV via Python to estimate the causal effect;
   report point estimate and standard error.
3) Fill out a JSON object with the identification specification,
   adhering to the provided schema.
4) Return the full Python script executed to produce the numbers.

FORMAT: Output a single JSON object with keys:
{ "identification_output": ..., "estimation_code": { "language": "
      python",
  "code": "...", "explanation": "..." } }
```

F Additional Error Analysis

To understand why the model fails, we categorize incorrect identification across the 114 failed cases. Table 7 summarizes the error types. Figure 1 further decomposes these errors by identification strategy, revealing that each design family has a distinct error profile.

Table 7: Taxonomy of identification errors across all 114 incorrect cases. A single case may contribute to multiple categories. Counts are ordered by frequency.

Error Family	Error Type	Count
Control variables	Missing required controls	39
	Included bad / post-treatment controls	18
Estimand	CATE → LATE	29
	CLATE → LATE	7
	Other estimand errors	10
Strategy	CE → RCT	24
	DiD → CE or IV → CE	10
	Other strategy errors	2
Variable specification	Wrong treatment variable	24
	Wrong outcome variable	11
Design-specific fields	Wrong IV-specific fields	10
	Wrong RDD / DiD-specific fields	9

Control variable errors. Control specification is the most error-prone component overall (39 cases with missing necessary controls, 18 with bad or post-treatment controls). A representative

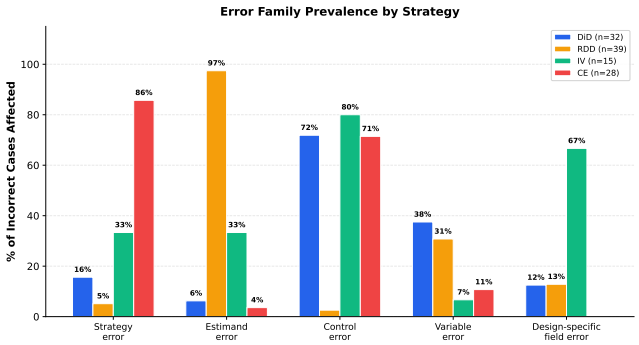

Figure 1: Percentage of incorrect cases affected by each error family, grouped by gold strategy. Categories are *non-exclusive*: a single case may exhibit multiple error types simultaneously (49% of incorrect cases do). Each bar shows the fraction of that strategy's incorrect cases containing the given error family. RDD is dominated by estimand errors (97%), CE by strategy errors (86%), and DiD, IV, and CE all show high rates of control errors (72%, 80%, 71%).

bad-control error occurs in a DiD study of parliamentary gender quotas and public-health spending [14], where the model includes the quota-induced change in women's legislative representation as a covariate. Because this variable is a post-treatment mediator that lies on the causal path from quota adoption to health spending, conditioning on it blocks part of the treatment effect and biases the estimate. The correct specification controls only for pre-treatment country characteristics, yet the model fails to distinguish a downstream consequence of treatment from a confounder.

IV instrument misidentification. Among the 22 IV queries, 10 cases exhibit errors in design-specific fields. An example comes from a study of peasant unrest and zemstvo representation in Imperial Russia [24], which uses historical religious polarization as an instrument for the frequency of peasant disturbances. The model selects a first-stage fitted value as the instrument rather than the actual instrument (historical religious polarization), suggesting it confuses the mechanics of two-stage estimation with the conceptual identification argument.

Strategy confusion: CE misidentified as RCT.. The single most striking pattern is the model's tendency to classify observational studies requiring covariate adjustment as Randomized Controlled Trials (24 cases, all from the IHDP textbook subset studying the effect of specialist home visits on children's cognitive outcomes). The metadata explicitly lists 25 baseline covariates—including birth weight, gestational age, maternal education, and neonatal health indicators—and a binary treatment indicator, but does *not* state that treatment was randomly assigned. The model appears misled by the word "assignment" in the query text, predicting RCT with an empty control set when the correct strategy is Conditional Exogeneity with all 25 covariates in the adjustment set.

Estimand confusion: CATE → LATE in RDD.. In 29 of 44 RDD queries (66%), the model correctly identifies the strategy but mislabels the estimand as LATE instead of CATE. For instance, in a study

of incumbency advantage in Japanese lower-house elections [1], the estimand is the treatment effect at the vote-margin cutoff—a conditional average treatment effect (CATE)—but the model labels it as LATE, conflating the RDD estimand with the complier effect from an IV design.

G Sample Query

To illustrate the task format, we present an example from Coppock and Green [17].

Example Causal Query

"What is the causal effect of voting in the November 2007 municipal election on the probability of voting in the January 2008 primary, estimated for the subgroup of individuals whose November 2007 turnout is changed by the study's experimental encouragement?"

This query, along with the dataset and detailed metadata (excerpts below), is provided to the agent.

First Few Rows of Provided CSV (excerpt):

```
yob1,id,precinct,...,treatmen,hh,lastelec,voted,avhh0,color,hhsize,...

571129,00000026,506800006,...,control,489243,,,,,1,...
160626,00000030,705804003,...,control,934562,,,,,2,...
421017,00000066,638800041,...,control,880582,,,,,1,...
```

Partial Metadata:

Variable	Description
`treatmen`	Mailing-group label (e.g., "control", "mail-arm A")
`voted`	Binary indicator for voting (0/1)
`hhsize`	Household size (registered individuals)
...	...

Dataset Context:

This dataset combines official voter-registration and voter-history records with household-level mailing assignments from a large mail program around municipal elections in 2007…

Gold-Standard Identification. The correct identification for this query is an **Instrumental Variable** design. The treatment is `voted` (voting in the November 2007 election), the outcome is voting in the January 2008 primary, and the instrument is `treatmen` (the randomized mailing assignment). The causal quantity is `LATE`, because the IV design identifies the effect for compliers—those whose November 2007 turnout was changed by the mailing encouragement. This example illustrates the level of reasoning required: the agent must recognize that the mailing assignment is a randomized encouragement (instrument) for the endogenous treatment (voting), and that the estimand is a LATE rather than an ATE.

H Paper list and citations

Table 8: Research papers included in the benchmark. Each row corresponds to one paper-sourced dataset; some papers contribute two queries.

Paper	Design	Title	#queries
[60]	DiD	Corporate Board Quotas and Gender Equality Policies in the Workplace	2
[78]	DiD	Deadly Populism: How Local Political Outsiders Drive Duterte's War on Drugs in the Philippines	2
[23]	DiD	Does Compliance Pay? Social Standards and Firm-level Trade	2
[40]	DiD	Does Direct Democracy Hurt Immigrant Minorities? Evidence from Naturalization Decisions in Switzerland	2
[74]	DiD	Education or Indoctrination? The Violent Origins of Public School Systems in an Era of State-Building	2
[90]	DiD	Elite Cleavage and the Rise of Capitalism under Authoritarianism: A Tale of Two Provinces in China	2
[34]	DiD	Elite Coalitions, Limited Government, and Fiscal Capacity Development: Evidence from Bourbon Mexico	2
[3]	DiD	How Does Armed Conflict Shape Investment? Evidence from the Mining Sector	2
[29]	DiD	How Exile Shapes Online Opposition: Evidence from Venezuela	2
[8]	DiD	Incremental Democracy: The Policy Effects of Partisan Control of State Government	2
[66]	DiD	Killing in the Slums: Social Order, Criminal Governance, and Police Violence in Rio de Janeiro	2
[36]	DiD	Laboratories of Democratic Backsliding	2
[21]	DiD	Loyal Leaders, Affluent Agencies: The Budgetary Implications of Political Appointments in the Executive Branch	2
[80]	DiD	Making unequal democracy work? The effects of income on voter turnout in Northern Italy	2
[25]	DiD	Metrics Management and Bureaucratic Accountability: Evidence from Policing	2
[56]	DiD	Motivated Corporate Political Action: Evidence from an SEC Experiment	2
[13]	DiD	Party Sub-Brands and American Party Factions	1
[53]	DiD	Public Money Talks Too: How Public Campaign Financing Degrades Representation	2
[14]	DiD	Quota Shocks: Electoral Gender Quotas and Government Spending Priorities Worldwide	2
[38]	DiD	Race and Representation in Campaign Finance	2
[82]	DiD	Race, representation, and the voting rights act	2
[37]	DiD	Rock the Registration: Same Day Registration Increases Turnout of Young Voters	2
[64]	DiD	The Effect of Firm Lobbying on High-Skilled Visa Adjudication	2
[33]	DiD	The Effect of the Voting Rights Act on Enfranchisement: Evidence from North Carolina	2
[88]	DiD	The Geography of Inequality: How Land Use Regulation Produces Segregation	2
[47]	DiD	The Growth of Campaign Advertising in the United States, 1880–1930	2
[76]	DiD	The Partisan Logic of City Mobilization: Evidence from State Lobbying Disclosures	2
[11]	DiD	The Politics of Property Taxation: Fiscal Infrastructure and Electoral Incentives in Brazil	2
[57]	DiD	The Representational Consequences of Municipal Civil Service Reform	2
[43]	DiD	The Supply-Equity Trade-Off: The Effect of Spatial Representation on the Local Housing Supply	2
[67]	DiD	Trauma and Turnout: The Political Consequences of Traumatic Events	1
[77]	DiD	Unpaved Road Ahead: The Consequences of Election Cycles for Capital Expenditures	2
[28]	RDD	A Gubernatorial Helping Hand? How Governors Affect Presidential Elections	2
[86]	RDD	Businesspeople in Elected Office: Identifying Private Benefits from Firm-Level Returns	1
[75]	RDD	Capitol Gains: The Returns to Elected Office from Corporate Board Directorships	2
[7]	RDD	Congressional Candidates in the Era of Party Ballots	1

Paper	Design	Title	#queries
[81]	RDD	Congressional Parties and Civil Rights Politics from 1933 to 1972	4
[39]	RDD	Correcting Misperceptions Can Increase Anti-Immigration Attitudes	2
[54]	RDD	Direct democracy and women's political engagement	1
[72]	RDD	Disloyal Brokers and Weak Parties	1
[69]	RDD	From Top-Down to Trickle-Up Influence: Revisiting Assumptions About the Family in Political Socialization	4
[31]	RDD	Gubernatorial Midterm Slumps	1
[8]	RDD	Incremental Democracy: The Policy Effects of Partisan Control of State Government	1
[1]	RDD	Incumbency Disadvantage under Electoral Rules with Intraparty Competition: Evidence from Japan	2
[27]	RDD	Incumbency Effects and the Strength of Party Preferences: Evidence from Multiparty Elections in the United Kingdom	1
[15]	RDD	Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities	1
[48]	RDD	Making Young Voters: The Impact of Preregistration on Youth Turnout	1
[26]	RDD	MPs for Sale? Returns to Office in Postwar British Politics	1
[22]	RDD	Off-Cycle and Out of Office: Election Timing and the Incumbency Advantage	2
[30]	RDD	Political Devolution and Resistance to Foreign Rule: A Natural Experiment	1
[5]	RDD	Preaching to the Choir: Americans Prefer Communicating to Copartisan Elected Officials	1
[87]	RDD	Targeting Ordinary Voters or Political Elites? Why Pork Is Distributed Along Partisan Lines in India	1
[32]	RDD	The Financial Incumbency Advantage: Causes and Consequences	2
[55]	RDD	The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability	1
[4]	RDD	The Spoils of Victory: Campaign Donations and Government Contracts in Brazil	1
[46]	RDD	Voter Buying: Shaping the Electorate through Clientelism	1
[41]	RDD	What Happens When Extremists Win Primaries?	2
[42]	RDD	Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in US Elections	1
[59]	IV	Anger and its Consequences for Judgment and Behavior: Recent Developments in Social and Political Psychology	1
[44]	IV	Childhood Socialization and Political Attitudes: Evidence from a Natural Experiment	1
[18]	IV	China y EE. UU. en Latinoamérica	1
[24]	IV	Collective action and representation in autocracies: Evidence from Russia's great reforms	1
[24]	IV	Collective action and representation in autocracies: Evidence from Russia's great reforms	1
[19]	IV	Deliberate Disengagement: How Education Can Decrease Political Participation in Electoral Authoritarian Regimes	1
[73]	IV	Do Conditional Cash Transfers Affect Electoral Behavior? Evidence from a Randomized Experiment in Mexico	1
[84]	IV	Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy	1
[70]	IV	Exploiting Friends-and-Neighbors to Estimate Coattail Effects	1
[6]	IV	Foreign Aid, Human Rights, and Democracy Promotion: Evidence from a Natural Experiment	1
[16]	IV	Is Voting Habit Forming? New Evidence from Experiments and Regression Discontinuities	1
[35]	IV	Party Affiliation, Partisanship, and Political Beliefs: A Field Experiment	1
[63]	IV	Personal Experience and Public Opinion: A Theory and Test of Conditional Policy Feedback	1
[71]	IV	Secular Party Rule and Religious Violence in Pakistan	1
[79]	IV	Small Aggregates, Big Manipulation: Vote Buying Enforcement and Collective Monitoring	1
[68]	IV	Social Esteem and Participation in Contentious Politics: A Field Experiment at an LGBT Pride Rally	1

Paper	Design	Title	#queries
[62]	IV	The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect	2
[57]	IV	The Representational Consequences of Municipal Civil Service Reform	1
[10]	IV	Urbanization Patterns, Information Diffusion, and Female Voting in Rural Paraguay	1