Benchmarking LLM Causal Reasoning with Scientifically Validated Relationships

Donggyu Lee[1,2], Sungwon Park[1,2], Yerin Hwang[2,3], Hyoshin Kim[1], Hyunwoo Oh[1], Jungwon Kim[1], Meeyohng Cha[1,2], Sangyoon Park[4], Jihee Kim[1]

[1]Korea Advanced Institute of Science and Technology,
[2]Max Planck Institute for Security and Privacy,
[3]Seoul National University,
[4]Hong Kong University of Science and Technology

Abstract

Causal reasoning is fundamental for Large Language Models (LLMs) to understand genuine cause-and-effect relationships beyond pattern matching. Existing benchmarks suffer from critical limitations such as reliance on synthetic data and narrow domain coverage. We introduce a novel benchmark constructed from causally identified relationships extracted from top-tier economics and finance journals, drawing on rigorous methodologies including instrumental variables, difference-in-differences, and regression discontinuity designs. Our benchmark comprises 29,972 evaluation items covering four task types across domains such as health, environment, technology, law, and culture. Experimental results on eight state-of-the-art LLMs reveal substantial limitations, with the best model achieving only 60.58% accuracy. Moreover, model scale does not consistently translate to superior performance, and even advanced reasoning models struggle with fundamental causal relationship identification. These findings underscore a critical gap between current LLM capabilities and demands of reliable causal reasoning in high-stakes applications.

1 Introduction

Recent advances in large language models (LLMs) have driven their adoption in a wide range of high-stakes applications. Motivated by the promise that LLMs' causal reasoning could enable robust multi-hop reasoning and complex problem-solving, researchers have begun to probe these models beyond surface-level pattern matching (Wei et al., 2022; Yao et al., 2023). Unlike surface-level pattern recognition, causal reasoning enables a model to understand cause-and-effect relationships, moving beyond mere statistical correlations (Pearl, 2009; Spirtes et al., 2000). In response, several works evaluate the causal competencies of LLMs: studies investigate whether they can infer causal graphs (Long et al.), answer counterfactual queries (Jin et al., 2023), and support causal discovery pipelines (Zečević et al., 2023).

Despite these advances, the field still lacks a benchmark that systematically evaluates LLMs' causal reasoning capabilities. Existing benchmarks for evaluating LLMs' causal reasoning suffer from key limitations: they often rely on low-fidelity synthetic data, focus on narrow or biased domains, and reduce causality to simplistic cause-effect identification. As a result, they fail to capture the complexity and diversity of real-world causal reasoning, making it difficult to assess whether LLMs truly reason causally or merely mimic patterns from training data.

To bridge this gap, we propose a benchmark grounded in causally identified relationships published in peer-reviewed economics and finance top journals. This benchmark targets a central research question: *Can contemporary LLMs understand and reason about complex, scientifically validated causal relationships across diverse real-world domains?* To this end, it offers (1) real-world causal relationships verified through rigorous scientific methods rather than synthetic generation, (2) broad domain coverage that extends beyond narrow contexts, and (3) multi-level evaluation tasks designed to distinguish genuine causal reasoning from pattern matching.

To address these limitations, we construct our benchmark through a systematic pipeline that extracts causally identified relationships from 14,977 papers published between 2000 and 2025 across eight top-tier economics and finance journals. Our approach leverages GPT-5-mini to extract candidate causal relationships from paper abstracts five times, employing clustering and consensus mechanisms to ensure reliability. This process yields 11,869 validated causal relationships spanning diverse economic and financial domains. From these relationships, we generate four distinct task types that progressively test different aspects of causal

reasoning—from simple cause-effect identification to complex multi-hop reasoning and directional inference—resulting in a final benchmark of 29,972 questions after filtering out trivially easy items.

Our evaluation of eight state-of-the-art LLMs reveals that all models exhibit significant limitations in causal reasoning. Even the best-performing model, Qwen3-32B, achieves only 60.58% accuracy, and notably, the latest large-scale models including GPT-5 demonstrate surprisingly poor performance. These results suggest that despite advances in model scale and recency, current LLMs still face fundamental challenges in genuine causal reasoning beyond pattern matching, revealing substantial room for improvement in understanding scientifically validated, complex real-world causal relationships.

2 Related work

Causal reasoning benchmarks for LLMs. A growing line of work evaluates whether LLMs move beyond associative pattern-matching toward interventional and counterfactual reasoning (Long et al.; Zečević et al., 2023). *CaLM* systematizes targets, adaptation regimes, metrics, and error analysis to define a broad design space for causal evaluation (Chen et al., 2024). *CausalBench* widens the lens across modalities (text, math, code) to probe causal understanding from multiple perspectives (Wang, 2024). *CLadder* operationalizes Pearl's ladder with natural-language tasks derived from oracle causal engines to directly test associational vs. interventional vs. counterfactual abilities (Jin et al., 2023; Pearl, 2009). Complementing these, "fresh-data" efforts such as *CausalProbe-2024* explicitly control training-set contamination and report significant sensitivity, indicating many models remain largely at level-1 (associational) reasoning (Chi et al., 2024; Zečević et al., 2023). Despite this progress, most benchmarks still emphasize synthetic or LLM-generated tasks that lack the validity of empirically verified causal relations from peer-reviewed academic papers. Moreover, their focus on symbolic languages, mathematics, and code—domains far removed from everyday life and social phenomena—limits their applicability to real-world causal understanding in human-centric contexts.

Causality extraction and economics/social-science reasoning. From the IE perspective, the *FinCausal* shared tasks (2020, 2022) formalized cause–effect detection in financial text (Mariko et al., 2020), while *EconLogicQA* targets sequential economic logic (Quan and Liu, 2024) and *EconNLI* casts economic implications as NLI (Guo and Yang, 2024). However, these three benchmarks (FinCausal/EconLogicQA/EconNLI) are primarily constructed from news, Wikipedia, and other secondary narratives, which are lower-quality and less vetted than peer-reviewed academic papers. In parallel, the "causal claims" line (e.g., Garg & Fetzer) aggregates paper-level causal graphs to map characteristics of the literature, but largely stops at corpus characterization rather than yielding fine-grained LLM evaluation tasks (Garg and Fetzer, 2024).

Overall, existing evaluations often rely on synthetic or secondary texts and focus narrowly on simple, direct causal relationships without testing diverse causal directions or analogical reasoning. We introduce a domain-grounded dataset built from peer-reviewed economics/finance articles with explicitly verified causal relations, diverse causal directions, and analogical inference tasks—offering a higher-quality, harder, and more credible test of LLM causal reasoning.

3 Benchmark construction

3.1 Workflow Overview

The benchmark construction in this study proceeds in two main stages. In the first stage, we extract causal relations in the form of $(X, Y, \text{direction})$ from 8 top-tier economics and finance journals. In the second stage, based on the extracted causal relations, we generate diverse natural language questions designed to evaluate LLMs' causal reasoning capabilities.

3.1.1 Data Collection

This study utilizes the top 5 economics journals (American Economic Review, Econometrica, Journal of Political Economy, Quarterly Journal of Economics, Review of Economic Studies) and the top 3 finance journals (Journal of Finance, Journal of Financial Economics, Review of Financial Studies) as data sources. Data collection was conducted using the OpenAlex API[1], which provides access to comprehensive metadata for academic publications. For each paper, we collected the title, abstract, publication year, and journal name. The collection period

[1]OpenAlex is an open-source bibliographic database providing comprehensive metadata for scholarly works (https://openalex.org).

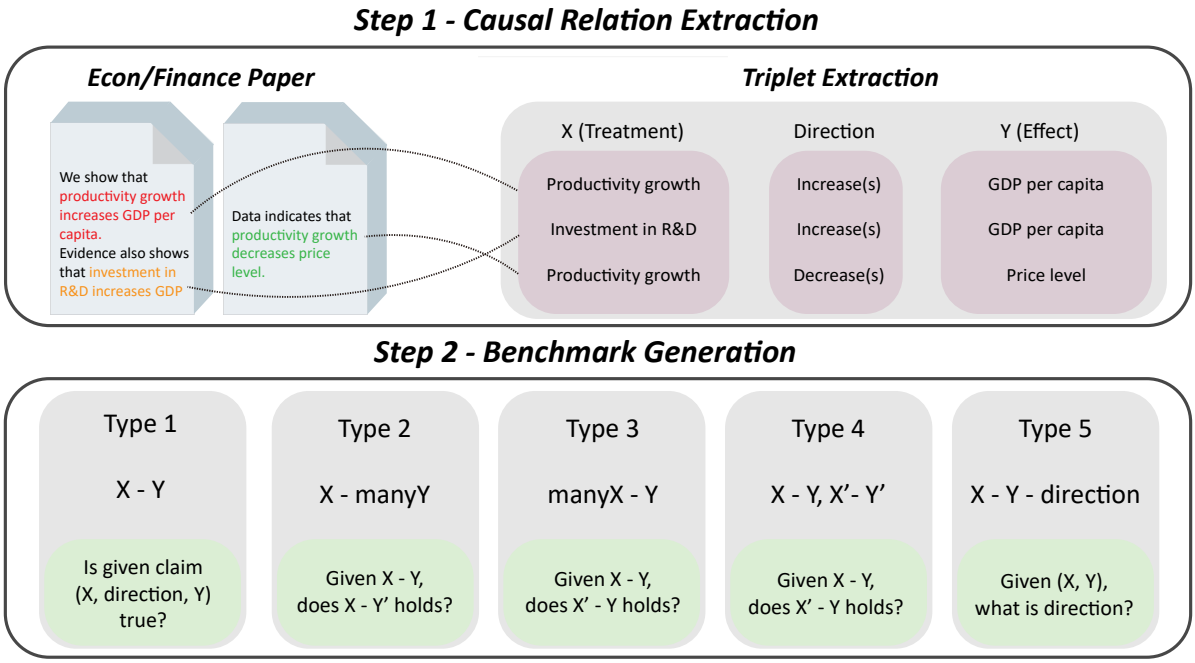

Figure 1: Overall Process of Causal Relation Extraction and Benchmark Generation

spans from 2000 to 2025, encompassing 14,977 papers in total.

3.1.2 Causal Relation Extraction Process

For each collected paper, we employed the GPT-5-mini model to extract causal relation information from the title and abstract. Let $\mathcal{D} = \{(X, d, Y) \mid X, Y \in \text{variables}, d \in \{\text{increase}, \text{decrease}, \text{none}\}\}$ denote the set of all extracted causal relation triplets, where X represents a cause variable, Y represents an effect variable, and d represents the direction of the causal effect. To ensure validity and robustness of extraction, we implemented the following procedure:

1. **Multiple Extraction**: We repeated the extraction process 5 times for each paper, yielding sets $\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_5$.

2. **Entity Identification**: Each extracted variable X and Y was converted into an embedding vector[2], and pairs with cosine similarity ≥ 0.9 were considered identical entities.

3. **Consensus-based Filtering**: The final dataset was defined as $\mathcal{D} = \{(X, d, Y) \in \bigcup_{i=1}^{5} \mathcal{D}_i \mid |\{i : (X, d, Y) \in \mathcal{D}_i\}| \geq 4\}$, containing only triplets that appeared in at least 4 out of 5 extractions.

[2]We used OpenAI's text-embedding-3-large model for generating embedding vectors.

Through this process, $|\mathcal{D}| = 11{,}869$ validated causal relations were extracted.

3.1.3 Quality Validation

To evaluate the reliability and validity of the extracted causal relations, we conducted a human assessment. Two independent annotators reviewed 104 causal relations randomly sampled from the full set of extractions, with the sampling procedure designed to ensure temporal coverage by including four relations per year from 2000 to 2025.

Each relation was evaluated along three complementary dimensions. First, annotators assessed whether the two entities, X and Y, had been correctly identified depending on accuracy and completeness (0-2 point). Second, they determined whether the causal direction between X and Y had been correctly assigned (0–1 point). Third, they judged whether the relation represented a genuine causal link rather than a spurious correlation (0–1 point). These criteria yielded a maximum of 4 points per relation per annotator.

Inter-rater reliability was quantified using the Intraclass Correlation Coefficient (ICC). The resulting coefficient was 0.57, which is generally interpreted as moderate agreement.

3.1.4 Characteristics of Causal Relations

Variable Classification Based on JEL Codes JEL (Journal of Economic Literature) classification

Table 1: Statistics of Extracted Causal Relations

Metric	Value
Unique X entities	9,657
Unique Y entities	10,202
Average causal relations per paper	1.87
Consistent extractions ($\geq$4/5 attempts)	79%

codes are a standardized classification system used to categorize economic research by topic, ranging from microeconomics and macroeconomics to specialized fields like health economics and environmental economics.[3]To analyze the topical coverage of our benchmark, we matched the extracted X and Y variables with JEL codes using embedding-based similarity. Specifically, we computed embedding vectors for each variable and for JEL secondary classification codes, then assigned variables to the JEL category with the highest cosine similarity.

As shown in Table 2, our benchmark covers a diverse range of topics beyond traditional economics, including health, environmental economics, technological change, and legal systems, demonstrating the interdisciplinary nature of causal inference questions.

Topic	Frequency
General Financial Markets	2357
Intertemporal Choice and Growth	868
Corporate Finance and Governance	738
Market Structure and Pricing	510
Prices, Business Fluctuations, and Cycles	508
Game Theory and Bargaining Theory	455
Technological Change; Research and Development	346
Health	178
Environmental Economics	120
Legal Procedure, the Legal System, and Illegal Behavior	104

Table 2: Top 5 Most Frequent JEL Tertiary-Level Topics and Top 5 Diverse Topics Beyond Core Economics (selected from top 50 topics)

3.2 Benchmark Generation

3.2.1 Question Types

Type 1: Causal Relation Identification (X-Y) This type assesses whether a causal relationship exists for a given triplet (X, d, Y), evaluating the most fundamental ability of identifying valid causal claims. All generated questions are ground-truth cases from academic journals, requiring models to confirm the validity of the causal relationship.

Type 2: Effect Variation (X-manyY) For relations where $\{(X, d, Y), (X, d', Y')\} \subset \mathcal{D}$, this type assesses whether a given triplet (X, d, Y) is valid, then (X, d', Y') is true, evaluating the understanding of causal spillover effects. Positive cases use the actual direction from the dataset, while negative cases replace d' with alternative directions (e.g., increase $\rightarrow$ decrease/none). This measures the ability to distinguish correct effect directions from incorrect ones.

Type 3: Cause Variation (manyX-Y) For relations where $\{(X, d, Y), (X', d', Y)\} \subset \mathcal{D}$, this type assesses whether a given triplet (X, d, Y) is valid, then (X', d', Y) is true, evaluating understanding of multiple causality. Positive cases use the actual direction from the dataset, while negative cases replace d with alternative directions (e.g., increase $\rightarrow$ decrease/none). This measures the ability to consider confounding and alternative explanations.

Type 4: Causal Direction Identification (X-Y-direction) For relations where $(X, d, Y) \in \mathcal{D}$, given a variable pair (X, Y), this type asks models to predict the direction $d \in \{\text{increase}, \text{decrease}, \text{none}\}$ of the causal effect. Unlike Types 1-4 which are binary true/false questions, this type requires selecting the correct direction among three options, evaluating the ability to predict qualitative outcomes of interventions. Understanding whether a policy promotes or suppresses a target variable is essential for practical decision-making.

3.2.2 Filtering Easy Questions

To exclude trivially easy questions, we filtered the dataset using three small language models (Llama 3.2 3B, Qwen 3 4B, Ministral 3B). Questions cor-

[3]The JEL classification system consists of three hierarchical levels: primary categories (single letters, e.g., "I" for Health, Education, and Welfare), secondary categories (two-digit codes, e.g., "I1" for Health), and tertiary categories (three-digit codes, e.g., "I12" for Health Behavior). The complete classification is available at https://www.aeaweb.org/econlit/jelCodes.php.

Table 3: Statistics of the Benchmark Dataset. The first row shows the number of samples for each question type, and the remaining rows show the distribution of ground-truth answers.

	All	X-Y	X-manyY	manyX-Y	X-Y-direction
# of samples	29972	10012	4003	6557	9400
True	17795	10012	2797	4986	—
False	2777	—	1206	1571	—
increase	6151	—	—	—	6151
decrease	2579	—	—	—	2579
none	670	—	—	—	670

rectly answered by all three models were removed, as they likely rely on simple pattern matching rather than genuine causal reasoning. This ensures our benchmark maintains sufficient difficulty for evaluating true causal inference capabilities. After filtering, 29,972 questions remained.

4 Benchmark Results

4.1 Experimental Setup

We evaluated causal reasoning using eight state-of-the-art LLMs. For reasoning models, GPT-5, GPT-5-mini, DeepSeek-R1-0528, and QwQ-32B were used. For non-reasoning models, Llama-3.3-70B, Llama-3.1-8B, Qwen3-32B, and Mistral-medium-2505 were used.

Models span large-scale, medium-scale (32B$\sim$70B), and small-scale (8B) to analyze scale-dependent capabilities. Temperature was set to 0 where supported, with a 2000-token limit. All experiments used identical prompts and evaluation criteria.

4.2 Main Results

The benchmark results clearly reveal that current state-of-the-art LLMs demonstrate lower-than-expected performance in causal reasoning within economics and finance domains. Qwen3-32B achieved the highest overall accuracy (ALL) at 60.6%, yet this still leaves substantial room for improvement. Notably, model size or recency does not necessarily guarantee performance. GPT-5 recorded one of the lowest accuracies at 30.4%, comparable to GPT-5-mini (38.6%).

4.3 Difficulty in Causality Judgment

All models struggled to accurately identify causal relationships even in the **Type 1 (X-Y)**, which contains high-quality, high-context causal statements directly extracted from peer-reviewed social science papers. The average accuracy across all models for X-Y was 41.0%, with F1 scores averaging 26.9%. Surprisingly, even the most advanced reasoning model, GPT-5, achieved merely 29.3% accuracy in this category, underperforming stronger models like Llama-3.3-70B (54.4%), QwQ-32B (58.5%), and Qwen3-32B (71.3%).

Beyond this, performance degradation was observed across most task types. For **Type 2 (X-manyY)**, the average accuracy dropped to 43.9% (F1: 42.1%), while **Type 3 (manyX-Y)** showed the poorest results with an average accuracy of 32.0% (F1: 30.0%). GPT-5's consistently low performance across all categories (ranging from 24.7% to 34.4% accuracy) further emphasizes the challenging nature of our dataset. These results demonstrate that our benchmark presents a novel and difficult challenge that effectively distinguishes model capabilities in causal reasoning.

4.4 Performance Degradation in Tasks

The **directional reasoning task (Type 4 (X-Y-direction))** exhibits large dispersion across models: the average accuracy is 51.3% (F1: 42.2%), with scores ranging from 31.5% to 71.1% in accuracy and 27.6% to 52.6% in F1. QwQ-32B attains the highest accuracy (71.1%), while Qwen3-32B achieves the best F1 (52.6%); three models exceed 0.50 F1 on this task (GPT-5-mini: 52.2%, QwQ-32B: 50.9%, Qwen3-32B: 52.6%).

Overall, complex causal reasoning that requires directional inference (**Type 4 (X-Y-direction)**) remains difficult and highly model-dependent. QwQ-32B and Qwen3-32B are comparatively stronger on directional inference, underscoring that architectural choices and tuning lead to distinct strengths across task types.

Type	X	d	Y	X'	d'	Y'	Question
Type 1 (X-Y)	sourcing patterns (advanced firms sourcing inputs from other advanced firms)	Increases	magnification effect of technology adoption				Is given claim (X, d, Y) true?
Type 2 (X-manyY)	introduction of highly subsidized, universally accessible child care in Quebec	Increases	hostile parenting (parenting hostility)		Decreases	parental health	Given (X, d, Y), is (X, d', Y') true?
Type 3 (manyX-Y)	direct promotion of ATM investment (policy)	Increases	welfare	menus of contracts (when second-period offers cannot be contingent on initial contract choice)	Increases		Given (X, d, Y), is (X', d', Y) true?
Type 4 (X-Y-direction)	human capital acquired while working in other industries before joining fund management		fund managers' information advantage				What is the direction d of causal effect from X to Y?

Table 4: Examples of Each Type of Question

Task	ALL		X-Y		X1-manyY	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
GPT-5	0.3037	0.3235	0.2929	0.2266	0.3285	0.2853
GPT-5-mini	0.3857	0.4183	0.2594	0.2060	0.3100	0.2568
DeepSeek-R1-0528	0.4364	0.4197	0.4794	0.3241	0.4372	0.4333
Llama-3.3-70B	0.4918	0.4487	0.5440	0.3524	0.4744	0.4743
Llama-3.1-8B	0.3854	0.3867	0.2493	0.1996	**0.5586**	**0.5443**
Mistral-medium-2505	0.2562	0.2571	0.1367	0.0547	0.3765	0.3537
QwQ-32B	0.5480	0.4859	0.5847	0.3690	0.4954	0.4953
Qwen3-32B	**0.6058**	**0.5161**	**0.7132**	**0.4163**	0.5299	0.5234
Average	0.4266	0.4070	0.4102	0.2685	0.4388	0.4208

Table 5: Model Performance Across Question Types (Part 1). Bold numbers indicate best performance for each metric. Weighted by sample count

4.5 By JEL Category

We analyzed performance across JEL classification codes at the primary level using **Type 1 (X-Y)** tasks

Task	manyX-Y1		X1-Y1-direction	
	Accuracy	F1	Accuracy	F1
GPT-5	0.2469	0.2102	0.3441	0.3492
GPT-5-mini	0.2411	0.2023	0.6532	0.5216
DeepSeek-R1-0528	0.2765	0.2645	0.5019	0.4411
Llama-3.3-70B	0.3187	0.3089	0.5641	0.4613
Llama-3.1-8B	**0.5864**	**0.5580**	0.3165	0.2944
Mistral-medium-2505	0.2480	0.2127	0.3150	0.2759
QwQ-32B	0.2910	0.2977	**0.7105**	0.5094
Qwen3-32B	0.3509	0.3477	0.7016	**0.5256**
Average	0.3199	0.3002	0.5134	0.4223

Table 6: Model Performance Across Question Types (Part 2). Bold numbers indicate best performance for each metric. Weighted by sample count

(causal relations directly extracted from papers). The reported accuracies represent averages across all eight models.

Domain-specific analysis revealed differentiated performance across economics fields. The highest accuracy was observed in Other Special Topics (Z, 45.0%), Political Economy and Comparative Economic Systems (P, 44.8%), and History of Economic Thought, Methodology, and Heterodox Approaches (B, 44.2%). In contrast, relatively lower performance was found in General Economics and Teaching (A, 32.5%), Health, Education, and Welfare (I, 35.4%), and Miscellaneous Categories (Y, 34.4%).

The performance variation across JEL categories reveals interesting patterns in LLMs' domain-specific capabilities. Superior performance in Other Special Topics (Z), Political Economy (P), and History of Economic Thought (B) suggests LLMs excel in fields characterized by qualitative reasoning and theoretical discourse, which align well with their text-based training. Conversely, lower accuracy in General Economics (A), Health, Education, and Welfare (I), and Miscellaneous Categories (Y) indicates challenges in handling empirically-driven or specialized technical content. The 12.5 percentage point gap between highest and lowest performing categories demonstrates that LLMs exhibit non-uniform expertise across economic subfields, with performance correlating more with linguistic characteristics than theoretical complexity.

5 Ablation Study: Open-Book Setting

5.1 Motivation

Our main results revealed a surprising pattern: state-of-the-art large-scale models such as GPT-5 and GPT-5-mini achieved unexpectedly low performance on our benchmark. One hypothesis is that these models may possess relevant economic and financial intuitions but struggled to confine their reasoning strictly to the provided textual evidence when instructed to *"Use ONLY the information provided... Do NOT rely on external knowledge."* To isolate the effect of this constraint, we conducted an ablation experiment comparing **closed-book** (original) versus **open-book** conditions.

5.2 Experimental Design

- **Closed-book**: The default prompt explicitly instructs models to ignore external knowledge to isolate pure causal reasoning ability, independent of whether domain-specific knowledge is present or absent.
- **Open-book**: The same items with the constraint instruction removed, allowing models to leverage their pre-trained domain knowledge.
- **Sampling**: We drew a stratified random 10% sample from each of Tasks 1–5, preserving the

Table 7: Closed-book vs. Open-book performance of GPT-5-mini on a stratified 10% sample. Δ denotes the score difference between Close and Open (Open - Close)

Task	Accuracy (%)			F1 score (%)		
	Close book	Open book	Δ_{Acc}	Close book	Open book	Δ_{F1}
All	38.57	54.27	+15.7	41.83	48.49	+6.66
Type 1 (X–Y)	25.94	58.74	+32.8	20.60	37.00	+16.40
Type 2 (X–manyY)	31.00	40.65	+9.65	25.68	38.84	+13.16
Type 3 (manyX–Y)	24.11	26.68	+2.57	20.23	24.01	+3.78
Type 4 (X–Y-direction)	65.32	74.57	+9.25	52.16	52.22	+0.06

number of items and True/False class balance per task.

- **Metrics**: Accuracy and Macro-F1 (averaged over both classes).

5.3 Results

Across the board, allowing *open-book* reasoning improved performance (Table 7): average accuracy rose by $+15.70$ points and Macro-F1 by $+6.66$ points, indicating that the "no external knowledge" constraint had been a binding limitation. The effect, however, is highly non-uniform across task types.

Impact of Prior Knowledge Type 1 (items extracted directly from economics/finance abstracts) shows a dramatic jump ($\Delta_{\text{Acc}} = +32.80$), ($\Delta_{\text{F1}} = +16.40$). This substantial improvement suggests that the model had already internalized these literature-derived causal patterns during pretraining. Removing the prohibition on external knowledge essentially allows the model to access its pre-existing representations of canonical mechanisms, stylized facts, and domain-specific terminology. In other words, when the target relation resembles those commonly documented in academic literature, the model's prior training on such corpora provides an effective substitute for—and complement to—evidence strictly within the item.

Tasks requiring heavy reasoning are less affected by prior knowledge. Types 2 and 4 showed moderate improvements when given prior knowledge (Acc $+9.65$, $+9.25$; F1 $+13.16$, $+0.06$). This means that prior knowledge helps clarify relationships or narrow down candidates, but it does not completely solve tasks that require comparing and connecting multiple entities or pairs.

The smallest improvement was seen in Type 3 (many-to-one relationships; Acc $+2.57$, F1 $+3.78$). Here, because the task requires aggregating information and reasoning through indirect steps, general domain knowledge alone does not help much.

These patterns suggest the following: when a task requires transforming or combining information (aggregation, analogy, or constraints across multiple pairs), simply "unlocking" prior knowledge provides limited performance gains. Instead, the ability to reason faithfully within the problem itself becomes more important.

6 Conclusion

This study introduces a novel benchmark for evaluating Large Language Models' causal reasoning capabilities, constructed from scientifically validated causal relationships extracted from top-tier economics and finance journals. By leveraging peer-reviewed research employing rigorous causal inference methodologies—including instrumental variables, difference-in-differences, and regression discontinuity designs—our benchmark of 29,972 evaluation items spans diverse societal domains including health, environment, technology, legal systems, and culture.

Experimental results reveal that current state-of-the-art LLMs demonstrate substantial limitations, with the best-performing model achieving only 60.6% overall accuracy. Notably, model scale does not consistently translate to superior performance, and even fundamental causal relationship identification tasks yield an average accuracy of merely 41.0% across models. This poor performance suggests that these models struggle with genuine causal understanding beyond pattern-matching from training data. These findings underscore a critical gap between current LLM capabilities and the requirements for reliable causal reasoning in high-stakes applications such as healthcare, finance, and policy-making, emphasizing the imperative need to address this capability gap for responsible and effective AI deployment.

Limitations

While our benchmark provides valuable insights into LLMs' causal reasoning capabilities, several limitations should be acknowledged:

- **Domain Specificity**: Our benchmark focuses exclusively on causal relationships extracted from economics and finance journals. While

these domains cover diverse societal topics (health, environment, technology, etc.), the dataset may not fully represent causal reasoning patterns in natural sciences, engineering, or other STEM fields. Future work should expand to incorporate causal relationships from broader scientific domains to enable more comprehensive evaluation of LLMs' general causal reasoning capabilities.

- **Inter-annotator Agreement**: Although 94% of sampled relations received high quality scores (≥7/8 points), the Intraclass Correlation Coefficient (ICC) 0.57 indicates limited statistical agreement between annotators. This mediocre ICC primarily stems from score concentration at maximum values rather than genuine disagreement, but nonetheless suggests that validation with a larger pool of annotators and more extensive sampling would strengthen confidence in extraction quality. Future iterations should employ more annotators and larger validation samples to establish more robust quality metrics.

- **Data Contamination**: We cannot completely rule out the possibility that some causal relationships in our benchmark may have been encountered during LLM pretraining. While our analysis shows relatively even accuracy across publication years, this does not provide absolute guarantees against data contamination.

Acknowledgments

References

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.

Hongzhan Chi, Junlin Zhou, Kangyi Zhu, Yuchuan He, Yangzhe Cai, Peng Li, Yi Yang, Yuxin Meng, Yang Wang, and Zhaopeng Tu. 2024. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Introduces the CausalProbe-2024 benchmark.

Prashant Garg and Thiemo Fetzer. 2024. Causal claims in economics. Technical Report 11462, CESifo Working Paper Series.

Yue Guo and Yi Yang. 2024. Econnli: Evaluating large language models on economics reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 982–994, Bangkok, Thailand. Association for Computational Linguistics.

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. 2023. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, pages 31038–31065. Curran Associates, Inc.

Stephanie Long, Tibor Schuster, and Alexandre Piché. Can large language models build causal graphs? In *NeurIPS 2022 Workshop on Causality for Real-world Impact*.

Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32, Barcelona, Spain (Online). COLING.

Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference*, 2 edition. Cambridge University Press.

Yinzhu Quan and Zefang Liu. 2024. Econlogicqa: A question-answering benchmark for evaluating large language models in economic sequential reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2273–2282, Miami, Florida, USA. Association for Computational Linguistics.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search*, 2 edition. The MIT Press, Cambridge, MA.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 11809–11822.

Matej Zečević, Moritz Willig, and Devendra Singh Dhami. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*. Published in TMLR (Aug 2023).

A Details on Causal Relation Extraction

Settings From top 5 Economics journals and top 3 Finance journals from 2000–2025 (until September 2025), we extract candidate causal claims as triplets (X, Y, d), where $d \in \{increase, decrease, none\}$.

Model. `GPT-5-mini` (reasoning effort = medium)
Decoding. `max_tokens = 2000`
Format. Strict JSON array
Repetition. Five independent extractions per paper

Causal Relation Validation Two human annotators evaluated 104 causal relations. Four papers were randomly sampled per year. Each annotator received the title, abstract, and extracted triplets (X, Y, d), and assigned scores based on three concise evaluation criteria shown in Table 8.

Table 8: Validation Criteria for Causal Relation Extraction

Label	Description	Score
Correct Entities	X and Y correctly identified as treatment and outcome.	+1 each
Correct Direction	Direction (d: increase / decrease / none) correctly inferred.	+1
Causal Validity	Relation reflects true causality, not mere correlation.	+1

Distribution by Year The temporal distribution shows a relatively even spread across periods, with a gradual increase toward more recent years.

Table 9: Temporal Distribution of Extracted Causal Relations Across Five-Year Periods (2000–2025)

Period	Percentage
2000–2004	17.0%
2005–2009	15.7%
2010–2014	19.1%
2015–2019	26.2%
2020–2025	21.9%

JEL Classification Table 10 lists the top 20 tertiary JEL categories among all extracted variables ($XandY$).

Table 10: Top 20 JEL Tertiary Categories by Frequency

Rank	JEL Category	Count
1	General Financial Markets	2357
2	Intertemporal Choice and Growth	868
3	Corporate Finance and Governance	738
4	Market Structure and Pricing	510
5	Prices, Business Fluctuations, and Cycles	508
6	Particular Labor Markets	500
7	General Equilibrium and Disequilibrium	499
8	Econometric Modeling	463
9	Game Theory and Bargaining Theory	455
10	Taxation, Subsidies, and Revenue	452
11	Market Structure, Firm Strategy, and Market Performance	435
12	Information, Knowledge, and Uncertainty	433
13	Analysis of Collective Decision-Making	424
14	Monetary Policy, Central Banking, and the Supply of Money and Credit	396
15	Money and Interest Rates	356
16	Household Analysis	356
17	Technological Change; Research and Development	346
18	International Finance	336
19	Personnel Economics	334
20	Mobility, Unemployment, and Vacancies	331

B Details on Evaluation

Settings We evaluated models using the following API configurations: OpenAI models (GPT-5 and GPT-5-mini) were accessed through the official OpenAI API, while Mistral Medium and Ministral were deployed via Azure. All other models were evaluated using the Huggingface API (normally

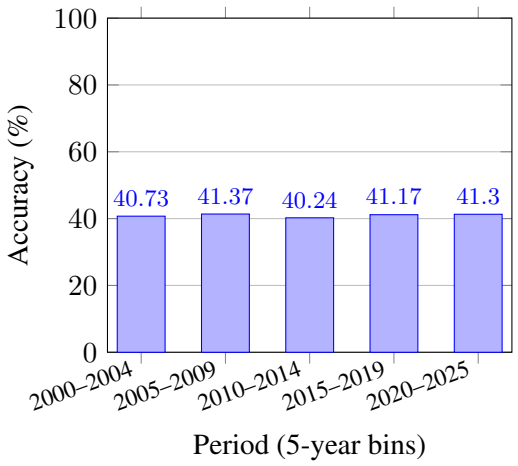

Figure 2: Five-year accuracy on Type 1 data (2000–2025) with 0–100 scale on the y-axis.

through Nebius API). For all models that support temperature configuration, the temperature parameter was set to 0 to ensure deterministic and reproducible outputs.

Accuracy by Year The following table demonstrates that there is no clear temporal variation in model performance. The accuracy scores are based exclusively on Type 1 data, which was originally extracted from causal relations in Economics and Finance journals, and averaged across all 8 LLMs. The lack of year-based trends suggests that our main experimental setting successfully mitigates the influence of domain-specific prior knowledge when measuring causal reasoning capabilities.