EconCausal: A Context-Aware Economic Reasoning Benchmark for Large Language Models

Anonymous ACL submission

Abstract

Socio-economic causal effects depend heavily on their institutional and environmental contexts. The same intervention can produce different, even opposite, effects across regulatory regimes, market conditions, time periods, or populations. This poses a challenge for large language models (LLMs) in decision-support roles: can they infer the direction of a causal effect under a specified context, and revise that judgment when the context changes?

To address this, we introduce ECONCAUSAL, a large-scale benchmark of 10,490 context-annotated causal triplets extracted from 2,595 high-quality empirical studies in top-tier economics and finance journals, constructed through a rigorous four-stage pipeline with multi-run consensus, context refinement, and multi-critic filtering.

Across models, LLMs often fail to condition their predictions on context. While top models reach $\sim$88% accuracy in fixed, explicit contexts, accuracy falls by 32.6 pp on cases that require revising the sign across contexts (73.9% to 41.3%), and drops below 50% once misleading signed evidence is introduced. Models also over-commit to directional ($+/-$) signs, recognizing null effects only 13.8% of the time while remaining poorly calibrated on these categories. The dataset and benchmark are publicly available at https://anonymous.4open.science/r/econcausal-benchmark-6F12.

1 Introduction

Understanding causal relationships in real-world socio-economic settings underpins decisions ranging from government policy and corporate strategy to everyday household choices. As Large Language Models (LLMs) are increasingly integrated into decision-support tools—not only summarizing information but also recommending actions and, in some settings, triggering downstream decisions in semi-autonomous or agentic workflows—a key question is whether they can make reliable judgments in complex socio-economic settings.

Crucially, socio-economic causality is often context-dependent. In many natural-science settings, the key conditions are usually measurable, and established laws specify how they enter the mechanism—so accounting for them in a model is relatively direct. In contrast, socio-economic "context" is qualitatively different: it is often high-dimensional, only partially observed, difficult to quantify (e.g., enforcement intensity or informal norms), rarely subject to controlled manipulation, and sometimes endogenous to outcomes, since institutions and behavior can co-evolve with the system being studied. An illustrative example is the employment response to a minimum-wage increase: the sign of the estimated effect can vary across environments shaped by regulation, enforcement, and market conditions—some of which are hard to quantify and may co-evolve with outcomes.

This context sensitivity poses a practical challenge for LLM-based decision support. While LLMs show promise in general reasoning (Kojima et al., 2022; Srivastava et al., 2023; Wang et al., 2024; Wei et al., 2022), applying them to economic decision-making requires reasoning that is conditioned on context and robust to context shifts. Without accounting for the environment in which a claim is made, causal judgments can be superficial and potentially misleading. We formalize the challenge as follows: given a socio-economic context (C) and a candidate treatment–outcome pair (T,O), can an LLM infer the directional sign of the causal effect—and adjust this judgment appropriately when C changes? We refer to this as a causal triplet: a treatment–outcome pair together with the directional sign of the effect. We focus on the directional sign of an effect, taking effect sizes, identification, and mechanisms as established by the source studies rather than re-deriving them.

Existing causal reasoning benchmarks are insuf-

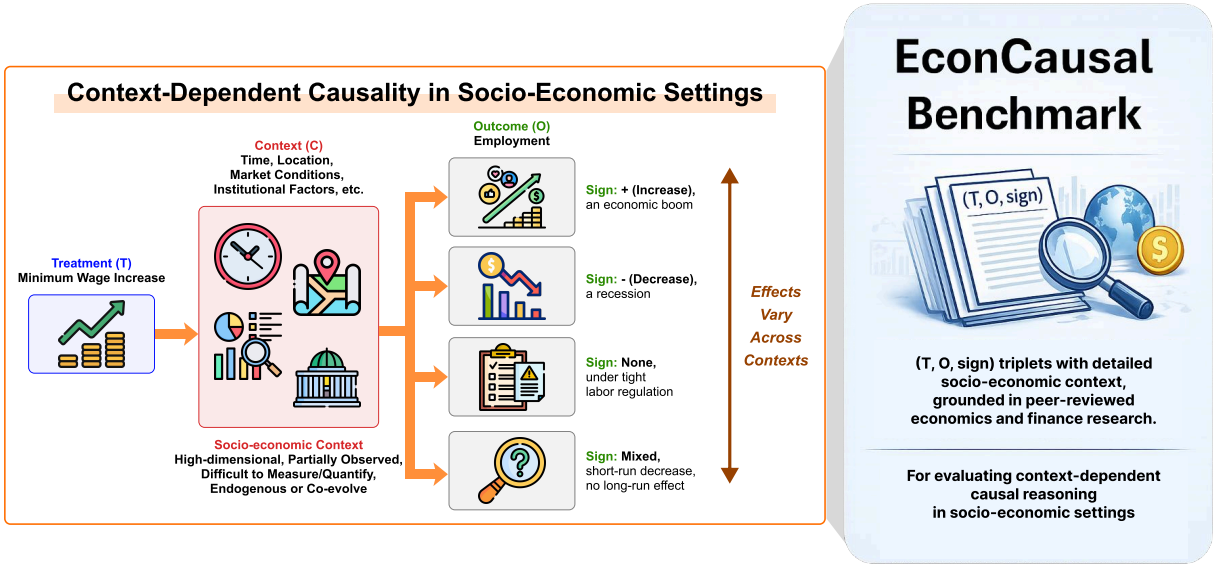


Figure 1: Context-dependent causality in socio-economic settings. Causal effects can vary across high-dimensional and partially observed contexts, posing a challenge for LLMs for socio-economic settings.

ficient for this setting for three reasons. First, they provide little coverage of empirically grounded causal claims in socio-economic contexts—most focus on generic, context-light relations or synthetic causal chains rather than evidence-based findings from the economics and finance literature. Second, they typically omit the institutional and empirical context needed to interpret a claim, treating causal relations as isolated facts. Third, they rarely test robustness under context shifts, even though socio-economic causal effects depend on high-dimensional, partially observed, and potentially endogenous contexts (Figure 1).

To address these limitations, we introduce **EconCausal**, a dataset and benchmark for evaluating context-dependent causal reasoning in economics and finance. EconCausal represents empirical findings as context-annotated treatment–outcome–sign triplets, enabling evaluation of whether LLMs can make causal judgments under explicit socio-economic contexts and revise them when the context changes.

Contributions. This paper makes three main contributions. First, we introduce **EconCausal**, a large-scale **context-dependent economic causal reasoning benchmark** built from high-quality empirical economics and finance research, where context-annotated treatment–outcome–sign triplets are constructed through a rigorous LLM-based pipeline with multi-run agreement, context refinement, multi-critic filtering, and expert validation. Second, we design **context-shift evaluation tasks** that test whether models can identify causal signs under explicit contexts, revise judgments when similar treatment–outcome pairs appear in different socio-economic contexts, and remain robust to misleading signed evidence. Lastly, experiments across diverse LLMs reveal **systematic failures under sign flips and misleading evidence**, showing that models often conflate plausible economic patterns with robust context-grounded causal judgments.

2 Related Work

2.1 Context-Dependent Reasoning in NLP

Recent work has studied context-dependent reasoning, where the correct answer or judgment depends on explicitly specified conditions rather than being fixed for an input. Situated and temporal QA benchmarks test whether models can answer questions whose answers vary with time, geography, or evolving facts (Zhang and Choi, 2021; Chu et al., 2024), while ambiguous QA studies underspecified questions that require additional context to resolve multiple valid interpretations (Min et al., 2020). Beyond QA, situated reasoning benchmarks examine whether models adjust their judgments when scenario conditions or normative constraints change (Ashida and Sugawara, 2022; Emelin et al., 2021).

However, these benchmarks typically use compact contexts centered on a single salient factor, such as time, location, ambiguity, or scenario condition. Multi-dimensional socio-economic contexts—where institutions, markets, policies, regions, populations, and baseline conditions jointly shape how

a relationship should be interpreted—remain comparatively underexplored.

2.2 Economic and Causal Reasoning Benchmarks for LLMs

Recent work has evaluated economic reasoning in LLMs through domain-specific benchmarks and extraction tools. STEER and STEER-ME (Raman et al., 2024, 2026) assess microeconomic rationality and game-theoretic optimality, while EconNLI (Guo and Yang, 2024) and EconLogicQA (Quan and Liu, 2024) evaluate inference and sequential reasoning in economic corpora. Large-scale extraction of causal claims has also been used to analyze empirical economics research (Garg and Fetzer, 2025). Related causal-reasoning benchmarks, such as CLADDER (Jin et al., 2023) and CausalBench (Wang, 2024), probe formal or commonsense causal understanding, with complementary work evaluating counterfactual and interventional reasoning in controlled settings (Huang et al., 2024).

However, these benchmarks mainly target rationality, inference, claim extraction, or controlled causal reasoning, rather than context-dependent sign prediction across changing socio-economic settings. EconCausal is complementary but distinct: it does not probe formal causal structure (e.g., graphical models or counterfactual identification), but instead tests whether models apply empirically established signs to the right contexts. It provides context-annotated treatment–outcome–sign triplets grounded in peer-reviewed economics and finance research, and tests whether LLMs can revise causal judgments across institutional, temporal, market, and policy contexts.

3 EconCausal

3.1 Causality Dataset Construction

The dataset is constructed from the **NBER Working Paper series** (National Bureau of Economic Research, 2025) (1991–2025), which serves as the primary preprint platform for influential, policy-relevant research. To ensure high empirical standards and external validity, the corpus is restricted to working papers subsequently published in the "top-five" economics journals—*American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*—and the "top-three" finance journals—*Journal of Finance*, *Journal of Financial Economics*, and *Review of Financial Studies*. These venues are globally recognized for their stringent review standards and rigorous causal identification strategies. This focus aligns with the **"credibility revolution"** in empirical economics (Angrist and Pischke, 2010), which marked a methodological shift toward explicit identification and the rigorous treatment of endogeneity. We rely on this rigor at the source: identification is performed by the original studies, and we adopt their validated effect directions as ground truth. Working papers were matched to their final published versions using official NBER metadata, resulting in a curated corpus of **5,006 papers** spanning more than three decades of research.

3.2 Dataset Extraction Pipeline

We build our causal-relation dataset using a four-stage LLM-based extraction pipeline. Figure 2 depicts the overall workflow of the data construction process.

Step 1: Triplet extraction with multi-run consensus. To ensure consistent treatment/outcome extraction, we run `GPT-5 mini` three independent times per paper and retain treatment–outcome pairs that appear in at least two of the three runs. Pairs with minor wording differences are merged using cosine similarity over `text-embedding-3-small` vectors; a pair is considered identical when both the treatment similarity and the outcome similarity reach or exceed 0.8 (see Appendix A for details). For each retained pair, the sign label $\in \{+, -,$ `None`, `Mixed`$\}$ is assigned by majority vote, and up to three supporting evidence paragraphs are stored verbatim for traceability.

Step 2: Paper-level metadata, global context, and identification methods. We extract paper-level metadata (empirical vs. theoretical), a concise global context paragraph summarizing core elements such as setting and unit of analysis, and a deduplicated set of identification methods mapped to a fixed ontology (e.g., DiD, IV, RDD, RCT). These global summaries serve as defaults for subsequent triplet-level refinement.

Step 3: Triplet-specific context and identification method. From this step onward, we consider only triplets from papers classified as *empirical* in Step 2, restricting the dataset to causal claims validated by established identification methods. For each triplet, we verify whether the global context and identification-method summaries remain valid

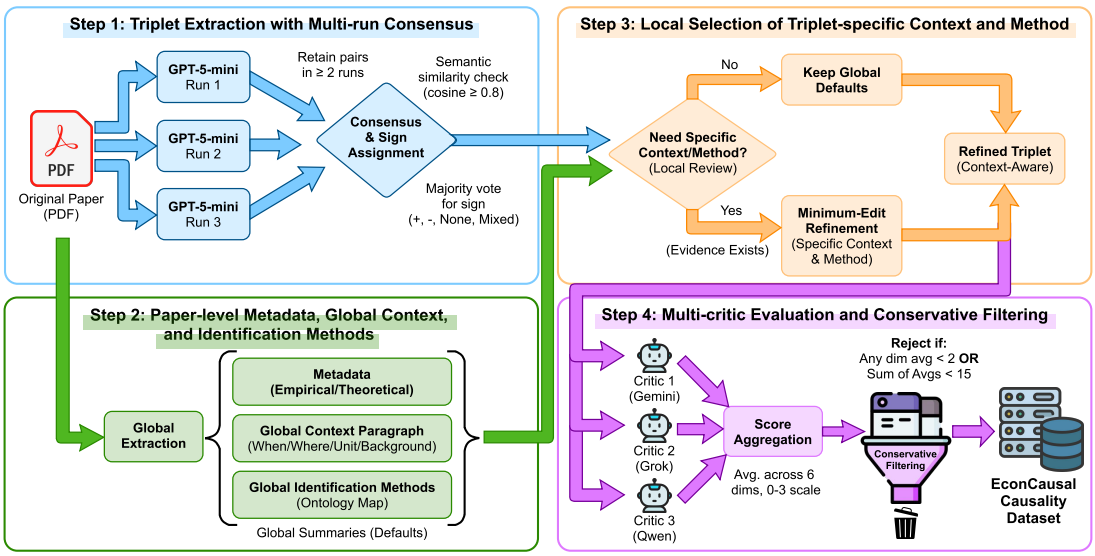


Figure 2: Overview of the EconCausal dataset extraction pipeline.

at the claim level. If the paper explicitly indicates that a given triplet is associated with a different or more specific context or identification method, we apply a minimal edit to produce a triplet-specific context; otherwise, the global defaults are retained unchanged.

Step 4: Multi-critic evaluation and conservative filtering. We apply an LLM-as-a-critic stage using three independent critic models, each scoring every triplet on six quality dimensions using a 0–3 rubric (see Appendix B for model and rubric details). Scores are averaged across critics, and triplets failing conservative quality thresholds are removed, filtering out 27.3% of candidates.

In total, the pipeline produces 10,490 context-annotated triplets drawn from 2,595 papers. The source-journal distribution is reported in Appendix C. The full extraction prompts for the four construction stages are provided in Appendix G.

3.3 Validation by Economists

We validate the critic-based filtering by comparing LLM critic scores with evaluations from three economics professors, who assessed 206 triplets within their areas of expertise using the same six criteria. The expert validation suggests that the filtering rule is conservative: while economists approved 83.5% of the evaluated Step 3 triplets, Step 4 retained 76.2%, and 86.6% of the retained triplets were also approved by economists. Criterion-level scores are also high, and further details are provided in Appendix D.

3.4 EconCausal's LLM Benchmark Tasks

Research Questions. Our benchmark is designed to probe three core capabilities of LLMs in economic causal reasoning. First, we ask whether state-of-the-art LLMs have internalized the high-quality economic causalities established in peer-reviewed economics and finance research (**RQ1**). Second, because causal relationships in socio-economic settings are inherently context-dependent—the same treatment can produce opposite effects under different institutional, temporal, or demographic conditions—we examine whether LLMs understand that causal signs can shift when the context changes (**RQ2**). Third, we investigate whether LLMs can reason robustly rather than merely reproducing patterns from their training corpora; specifically, we test whether they can filter out deliberately injected misinformation and still arrive at the correct causal judgment (**RQ3**).

Ground-truth causal signs. Each causal triplet is labeled based on the authors' preferred empirical specification, using four categories:

- $+$: The treatment significantly increases the outcome.
- $-$: The treatment significantly decreases the outcome.
- **None**: No statistically significant effect is found.
- **Mixed**: Effects are heterogeneous or multiple equally central results with opposite signs are reported.

Task #	Question	Evaluation Goal
Task 1 *Causal Sign Identification*	Given a context c and a treatment–outcome pair (T, O), what is the causal sign of $T \rightarrow O$ under c?	Does the LLM internalize economic causalities established in high-quality peer-reviewed research? (**RQ1**)
Example Given context c: "minimum-wage increase ($4.25 to $5.05) on fast-food establishments in New Jersey in April 1992, ..., amid a concurrent regional recession", T: minimum-wage increase, O: employment, what is the causal sign of T to O?		
Task 2 *Context-Dependent Sign Prediction*	Given triplet (T,O, sign) under context c_1, what is the causal sign of the same (or comparable) $T \rightarrow O$ under a different context c_2?	Does the LLM understand that causality is context-dependent and adjust its reasoning when the context changes? (**RQ2**)
Example Given the triplet (T: higher subsidy, O: insurance take-up rate, $+$) under context c_1: "a two-year (2010–2011) randomized field experiment in Jiangxi province, China, ...", what is the sign of the pair (T, O) under a different context c_2: "Massachusetts' subsidized insurance exchange *CommCare* from 2009–2013, ..."?		
Task 3 *Misinformation-Robust Sign Prediction*	Given the triplet (T,O, sign) under context c_1, together with a statement that deliberately reports an *incorrect* sign for c_1 (noise), what is the causal sign of $T \rightarrow O$ under a different context c_2?	Can the LLM discount misinformation and still perform robust, context-grounded causal reasoning? (**RQ3**)
Example Same as in Task #2 example, except the given triplet is (T: higher subsidy, O: insurance take-up rate, $-$), where the sign is incorrect.		

Table 1: Overview of the three EconCausal benchmark tasks.

Task descriptions. Guided by these research questions, we operationalize three benchmark tasks designed to probe progressively deeper dimensions of economic causal reasoning. Each task addresses an increasingly complex level of context-dependent interpretation. Table 1 provides an overview of each task, including representative questions and evaluation goals; the full task prompts are provided in Appendix H.

1. **Task 1 (Causal Sign Identification).** Given a context c and a treatment–outcome pair (T, O), predict the causal sign of $T \rightarrow O$ under c. We sample up to 30 triplets per publication year from each domain, yielding 947 Economics and 860 Finance questions (1,807 total).

2. **Task 2 (Context-Dependent Sign Prediction).** Given examples in which a treatment–outcome pair (T, O) exhibits an observed causal sign under a context c_1, predict the causal sign of the same (or a comparable) (T, O) under a different context c_2. We identify semantically matched treatment–outcome pairs by computing cosine similarity between treatment embeddings and between outcome embeddings, and retaining pairs whose average similarity exceeds 0.8. Questions are constructed using example and target contexts drawn from different papers, yielding 284 instances.

3. **Task 3 (Misinformation-Robust Sign Prediction).** Given examples of a treatment–outcome pair (T, O) with an observed causal sign under context c_1, together with an additional statement that deliberately reports an incorrect sign for c_1, predict the causal sign of the same (or a comparable) (T, O) under a different context c_2. We extend Task 2 by replacing the original sign with each of the remaining three labels among $\{+, -, \text{None}, \text{Mixed}\}$, resulting in three noisy variants per instance and a total of 852 questions.

4 Experiments

4.1 Baseline Models

We evaluate a diverse set of LLMs, spanning both proprietary and open-source systems.

Proprietary Models. We include representative proprietary models from major providers: *Gemini family* (Comanici et al., 2025) (Gemini 3 Flash, Gemini 2.5 Pro, Gemini 2.5 Flash), *GPT family* (Achiam et al., 2023) (GPT-5.2, GPT-5 mini, GPT-5 nano, GPT-4o, GPT-4o mini),

and *Grok family* (xAI, 2025) (Grok-4.1 Fast, Grok-3 Mini, Grok-3).

Open-Source Models. We also evaluate widely used open-source models: *Llama family* (Grattafiori et al., 2024) (Llama 3.3 70B, Llama 3.1 8B, Llama 3.2 3B, Llama 3.2 1B), and *Qwen family* (Yang et al., 2025) (Qwen3 32B, Qwen3 14B, Qwen3 8B).

Evaluation protocol. Each benchmark result is computed from a single model response per test instance under the corresponding task prompt. We

Model	Task 1 (Econ)	Task 1 (Finance)	Task 2		Task 3
			Overall	Sign-Mismatch	
Closed-Source LLMs					
Gemini 3 Flash	**0.884 (0.548)**	**0.868 (0.613)**	**0.824 (0.587)**	**0.634 (0.511)**	**0.624 (0.452)**
Gemini 2.5 Pro	0.829 (0.482)	0.808 (0.503)	0.747 (0.521)	0.535 (0.443)	0.552 (0.408)
Gemini 2.5 Flash	0.810 (0.432)	0.800 (0.521)	0.725 (0.498)	0.426 (0.342)	0.386 (0.295)
GPT-5.2	0.771 (0.440)	0.782 (0.518)	0.782 (0.571)	0.485 (0.414)	0.572 (0.426)
GPT-5 mini	0.753 (0.442)	0.750 (0.456)	0.750 (0.517)	0.416 (0.338)	0.534 (0.382)
GPT-5 nano	0.736 (0.422)	0.720 (0.435)	0.729 (0.477)	0.366 (0.284)	0.447 (0.326)
GPT-4o	0.665 (0.411)	0.704 (0.443)	0.658 (0.469)	0.346 (0.303)	0.357 (0.290)
GPT-4o mini	0.679 (0.364)	0.634 (0.339)	0.690 (0.467)	0.168 (0.150)	0.338 (0.250)
Grok-4.1 Fast	0.834 (0.476)	0.813 (0.521)	0.775 (0.525)	0.465 (0.363)	0.541 (0.405)
Grok-3	0.802 (0.424)	0.767 (0.461)	0.739 (0.501)	0.346 (0.267)	0.531 (0.394)
Grok-3 mini	0.828 (0.485)	0.798 (0.490)	0.711 (0.490)	0.356 (0.298)	0.547 (0.411)
Average	0.781 (0.448)	0.768 (0.482)	0.739 (0.511)	0.413 (0.338)	0.493 (0.367)
Open-Source LLMs					
Llama 3.3 70B	0.761 (0.418)	**0.718 (0.436)**	**0.747 (0.490)**	0.327 (0.245)	**0.525 (0.352)**
Llama 3.1 8B	0.565 (0.290)	0.509 (0.295)	0.676 (0.353)	0.297 (0.141)	0.473 (0.242)
Llama 3.2 3B	0.582 (0.230)	0.585 (0.252)	0.609 (0.244)	0.257 (0.114)	0.445 (0.212)
Llama 3.2 1B	0.620 (0.191)	0.599 (0.187)	0.486 (0.132)	**0.386 (0.112)**	0.511 (0.171)
Qwen3 32B	0.692 (0.380)	0.654 (0.402)	0.725 (0.463)	0.307 (0.230)	0.412 (0.302)
Qwen3 14B	0.626 (0.348)	0.600 (0.372)	0.630 (0.423)	0.317 (0.247)	0.392 (0.298)
Qwen3 8B	**0.773 (0.419)**	0.714 (0.413)	0.704 (0.474)	0.337 (0.265)	0.343 (0.262)
Average	0.660 (0.325)	0.626 (0.337)	0.654 (0.368)	0.318 (0.194)	0.443 (0.262)

Table 2: Main results on EconCausal benchmark. We report accuracy (Macro F1) for each task. "Sign-Mismatch" denotes the subset where the example sign contradicts the ground-truth sign of the target question. **Bold** indicates the best in each column.

use each provider's default decoding settings.

4.2 Main Results

Overview. Table 2 reports model performance across Tasks 1–3 of the EconCausal benchmark. Overall, current LLMs perform well when the causal sign is queried under a single explicit context, but their performance becomes less reliable when the task requires cross-context transfer or robustness to conflicting exemplar cues. Among proprietary models, Gemini 3 Flash achieves the highest accuracy, with $88.4\%/86.8\%$ on Task 1 (Economics/Finance), 82.4% on Task 2, and 62.4% on Task 3. While top models approach $\sim$90% in Task 1 and remain strong in aggregate Task 2 performance, Task 3 is substantially harder: closed- and open-source averages drop to 49.3% and 44.3% accuracy, respectively, with Macro F1 scores of 36.7% and 26.2%. This suggests that models are highly sensitive to conflicting signed examples when the displayed sign is deliberately corrupted or disagrees with the target ground truth.

Bias toward binary signs ($+/-$) over `None`/`Mixed`. Despite high aggregate accuracy in Tasks 1–2, Macro F1 scores remain much lower, revealing systematic class-level imbalance. Models tend to over-predict directional signs $(+)$ or $(-)$ while struggling to identify `None` and `Mixed` cases. For instance, averaging across the four task rows in Table 9, accuracy on `None` is only 13.83%, and `Mixed` reaches just 22.82%, both far below 74.82% for $(+)$ and 61.14% for $(-)$. This pattern suggests that many apparently correct predictions may reflect a tendency to commit to dominant directional signs rather than reliably distinguishing null or heterogeneous effects, a tendency likely amplified by the skewed ground-truth sign distribution (Figure 4 in the Appendix). Because aggregate Task 2 accuracy can obscure failures in context-dependent transfer, we analyze Task 2 sign-mismatch cases in Section 4.3.

4.3 Context-Dependent Transfer Failures

Sign-mismatch transfer. We focus on Task 2 *sign-mismatch* cases, where the source and target share semantically similar treatment–outcome relations but have different empirical signs across contexts. These cases are central to EconCausal because the source sign is valid in its original context but non-portable to the target context, directly testing whether models can decide when an analogous empirical finding should not transfer. Sign-mismatched cases constitute 35.6% of Task 2. While aggregate Task 2 performance is rel-

	Correct (%)	Error (%)	
		Example sign	Other signs
Task 1 (Question only)	52.70	27.28	20.02
Task 2 (With Example)	37.62	49.23	13.15
Δ	-15.08	+21.95	-6.87

Table 3: Prediction distribution on Task 2 sign-mismatch cases, with errors split by direction.

atively high, accuracy drops sharply on this subset: closed-source models decline from 73.9% overall to 41.3%, and open-source models decline from 65.4% to 31.8%. Thus, aggregate Task 2 scores substantially overestimate models' ability to reason about context-dependent transfer.

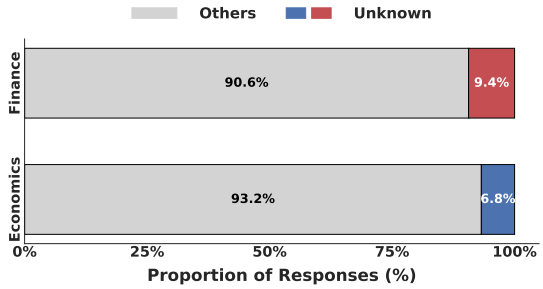
Source-sign over-transfer. We next examine whether this drop reflects random degradation on harder examples, or whether predictions systematically shift toward the source sign. We compare the standard Task 2 setting against a target-context-only setting that provides the same target context, treatment, and outcome, but omits the signed source example.

Table 3 shows that adding the signed source example shifts predictions toward the source-example sign. Correct target predictions decrease by 15.08 pp, while source-example-sign predictions increase by 21.95 pp; other incorrect predictions decrease. Among incorrect predictions, the share matching the source-example sign also rises from 57.67% to 78.99%. This distributional shift suggests that Task 2 failures are not merely random errors on difficult examples, but reflect over-transfer toward a source sign that is valid in its original context but non-portable to the target context. Additional sensitivity analyses for matching thresholds and prompt variants are provided in Appendix E.

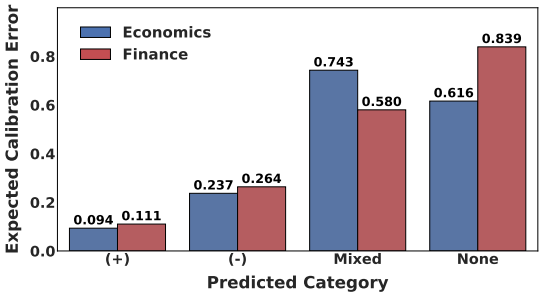
4.4 Calibration Analysis

In the domain of economic decision-making, where errors carry significant risks, a model's utility relies not only on predictive accuracy but also on the reliability of its confidence. To assess this reliability, we analyze GPT-4o's calibration, specifically determining if it resists the tendency toward 'over-commitment'—the projection of false certainty driven by surface-level correlations.

Experiment 1: Abstention under missing context. Using the Task 1 setup, we test whether the model can recognize when a causal judgment is unsupported due to insufficient information. We remove all contextual information, provide only the treatment–outcome pair, and add an explicit



(a) Proportion of 'Unknown' Responses



(b) ECE by Category

Figure 3: Comparative analysis of model uncertainty in Economics and Finance domains.

`Unknown` option. A well-calibrated model should frequently select `Unknown` in this setting. However, as shown in Figure 3a, the proportion of `Unknown` responses remains strikingly low across both domains. Even without any identifying context, GPT-4o commits to a deterministic causal sign, indicating a systematic failure to represent epistemic uncertainty as a first-class decision outcome. This over-commitment is especially concerning in policy-relevant settings involving null or heterogeneous effects, where premature sign assignment can mislead downstream decisions.

Experiment 2: Context-aware confidence calibration. We next evaluate whether model confidence aligns with empirical accuracy under the same Task 1 setting. We extract GPT-4o's log-probabilities over $\mathcal{Y} = \{+, -, \texttt{None}, \texttt{Mixed}\}$ and measure calibration using Expected Calibration Error (ECE) (Guo et al., 2017) :

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \quad (1)$$

where B_m denotes the set of instances in bin m. We further compute option-wise variants ECE_y by restricting to instances with label $y \in \mathcal{Y}$.

As shown in Figure 3b, GPT-4o is relatively well-calibrated for clear directional effects ($+$ and $-$), but severely miscalibrated for ambiguous categories: `None` in Finance (ECE = 0.839) and `Mixed` in Economics (ECE = 0.743). This pattern is consistent with the broader sign-level accuracy dis-

parity in Table 9: averaging across the four task rows, models achieve 74.82% on $(+)$ and 61.14% on $(-)$, but only 13.83% on `None` and 22.82% on `Mixed`.

The calibration analysis reveals that this accuracy gap is not merely a classification failure: GPT-4o is also poorly calibrated in precisely the categories where it performs worst. In particular, the model assigns high confidence to many incorrect `None` and `Mixed` predictions, suggesting that its bias toward binary signs is accompanied by overconfident miscalibration. These results show that the model's internal confidence scores do not reliably identify its weakest regimes, especially when causal effects are null, heterogeneous, or context-dependent.

5 Discussion and Future Directions

5.1 Reliability and Uncertainty Awareness

Our findings in §4.4 reveal a significant gap between model performance and the ability to recognize uncertainty. Despite the inclusion of an explicit "Unknown" option, GPT-4o exhibited a profound failure to abstain, attempting to predict causal signs even when all contextual information was removed. This tendency toward overcommitment suggests that, in this GPT-4o case study, the model does not reliably abstain when contextual information is insufficient, instead often producing plausible-sounding directional answers. Such behavior is particularly problematic in economic settings, where insufficient context should ideally trigger a demand for more information rather than a speculative judgment.

Furthermore, model calibration degrades substantially as causal complexity increases. While models appear relatively well-calibrated for simple directional effects, they are severely miscalibrated for complex categories such as `None` and `Mixed`. In these regimes, high confidence scores often mask systematically incorrect predictions, failing to signal when the underlying structural mechanism is ambiguous. These reliability concerns suggest that the model still exhibits important gaps in context-dependent economic reasoning, limiting its reliability for high-stakes economic decision-making.

5.2 Plausibility versus Portability

The source-sign over-transfer pattern in Section 4.3 suggests that the failure exposed by EconCausal is not simple context-insensitivity. Models can often use target context in isolation: in our Task 2 target set, adding target context improves accuracy from 64.42% to 72.10% over a treatment–outcome-only baseline. The difficulty arises when a semantically similar empirical precedent is provided. In Task 2, the source sign is not a random distractor, but a valid empirical sign in the source context. Models may therefore treat it as a plausible causal interpretation while failing to determine whether it should transfer.

This distinction is central to socio-economic causal reasoning. Many treatment–outcome relations do not have a single context-free sign: their effects may vary across institutions, time periods, populations, and baseline conditions. Thus, the key challenge is not merely generating a plausible causal sign, but deciding whether that sign applies in the target context. The Task 2 failures suggest that current LLMs often conflate plausibility with portability, over-transferring empirical findings across contexts where they should be revised or rejected.

6 Conclusion

This paper presents ECONCAUSAL, a large-scale benchmark for evaluating context-dependent reasoning over empirical causal findings in socio-economic domains. Constructed through a systematic four-stage pipeline, it curates context-annotated triplets that link a treatment and outcome to a directional sign within a specific socio-economic context, providing a rigorous foundation for evaluating how LLMs integrate institutional and environmental constraints into their judgments.

Across three tasks requiring increasing contextual generalization, our experiments reveal a persistent gap. While LLMs perform strongly in explicit settings, they exhibit a persistent anchoring effect under context shifts and misleading premises, applying empirically valid signs to contexts where they no longer hold rather than re-evaluating them. These findings suggest that current systems may not yet be robust enough for reliable deployment in high-stakes economic or policy-sensitive analyses. ECONCAUSAL thus provides a meaningful step toward systems that judge not only whether a causal sign is plausible, but whether it transfers to the context at hand.

Limitations

Dependence on LLM-assisted extraction and filtering. The dataset construction pipeline relies on LLMs for triplet extraction, context refinement, and multi-critic filtering. Multi-run consensus, conservative filtering thresholds, and economist validation reduce noise, but they cannot fully eliminate systematic extraction errors, model-specific biases, or omissions of subtle identification details. LLM-based annotators may still simplify nuanced empirical findings or assign contexts that are faithful at a high level but incomplete for fine-grained causal interpretation.

Corpus selection bias. EconCausal is constructed from NBER working papers that were subsequently published in top economics and finance journals. This design improves empirical quality, traceability, and consistency, but may overrepresent influential, English-language, publication-oriented research. As a result, our benchmark should be interpreted as a high-quality but selective subset of empirical economics and finance, rather than the full distribution of socio-economic causal claims.

Granularity of causal labels. EconCausal represents empirical findings using a compact directional label space: `+`, `-`, `None`, and `Mixed`. This abstraction enables scalable benchmark construction and consistent evaluation, but omits information such as effect size, economic magnitude, statistical strength, p-values, and confidence intervals. Thus, correctly predicting the sign should not be interpreted as fully understanding the empirical result.

Compression of heterogeneous effects. The current `Mixed` label compresses several distinct empirical patterns into a single category. Mixed effects may arise from heterogeneity across groups, sign changes across specifications or institutional settings, nonlinear responses, or temporal reversals. Future versions could split `Mixed` into interpretable subtypes and record the dimension along which the effect varies.

Scope of context and external validity. Finally, the context field in EconCausal is tied to the specific empirical setting of the source paper. This improves grounding and traceability, but limits our ability to evaluate external validity directly. Our context-shift tasks test whether models revise causal signs across paired settings, but they do not explicitly annotate portability-relevant features such as institutions, market structure, baseline conditions, or enforcement environment.

Behavioral scope of evaluation. Our evaluation is primarily behavioral. Beyond output predictions and their associated confidence (§4.4), we do not probe model internals such as intermediate representations or attention, nor attribute failures to specific computational mechanisms. Connecting these patterns to internal model processes is a direction for future work.

Ethical Considerations

EconCausal is constructed from scholarly economics and finance papers, including NBER working papers and their matched published versions. NBER does not hold the copyright for its Working Paper Series; rather, copyright remains with the authors of each paper. We therefore use the source papers solely for research purposes and provide attribution through bibliographic metadata and source identifiers. We do not redistribute source PDFs, full-text papers, or verbatim excerpts from the original articles. Any verbatim evidence used during dataset construction is retained only for internal traceability and quality control, and is excluded from the public release. The released EconCausal artifact will be distributed under the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

EconCausal is intended as a research benchmark for evaluating context-dependent causal reasoning in language models. It should not be used as a substitute for reading the original empirical studies, nor as a basis for policy, financial, legal, medical, or investment decisions. Since causal signs in socio-economic settings are context-dependent and may omit effect sizes, uncertainty, and heterogeneous mechanisms, systems evaluated on this benchmark should not be deployed in high-stakes decision-making settings without expert review and source verification.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joshua D. Angrist and Jörn-Steffen Pischke. 2010. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30.

Mana Ashida and Saku Sugawara. 2022. Possible stories: Evaluating situated commonsense reasoning under multiple possible scenarios. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3606–3630.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1204–1228.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

Prashant Garg and Thiemo Fetzer. 2025. Causal claims in economics. *arXiv preprint arXiv:2501.06873*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

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.

Yinya Huang, Ruixin Hong, Hongming Zhang, Wei Shao, Zhicheng Yang, Dong Yu, Changshui Zhang, Xiaodan Liang, and Linqi Song. 2024. Clomo: Counterfactual logical modification with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11012–11034.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and 1 others. 2023. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:31038–31065.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. Ambigqa: Answering ambiguous open-domain questions. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 5783–5797.

National Bureau of Economic Research. 2025. NBER working paper series. https://www.nber.org/papers. Accessed: 2026-02-09.

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.

Narun Raman, Taylor Lundy, Thiago Amin, Kevin Leyton-Brown, and Jesse Perla. 2026. Steer-me: Assessing the microeconomic reasoning of large language models. *Advances in Neural Information Processing Systems*, 38.

Narun Krishnamurthi Raman, Taylor Lundy, Samuel Joseph Amouyal, Yoav Levine, Kevin Leyton-Brown, and Moshe Tennenholtz. 2024. Steer: Assessing the economic rationality of large language models. In *International Conference on Machine Learning*, pages 42026–42047. PMLR.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

xAI. 2025. Grok 4.1 model card. https://data.x.ai/2025-11-17-grok-4-1-model-card.pdf. Accessed: 2025-11-17.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Michael Zhang and Eunsol Choi. 2021. Situatedqa: Incorporating extra-linguistic contexts into qa. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7371–7387.

A Triplet Extraction Details

Embedding model. We use OpenAI's `text-embedding-3-small` model to encode each extracted treatment and outcome phrase into a dense vector representation.

Similarity-based merging. Across the three independent extraction runs, minor surface-level wording differences (e.g., "physical activity" vs. "physical exercise") are common. To merge such near-duplicate pairs, we compute the cosine similarity between the embedding vectors of each treatment phrase and each outcome phrase separately. A candidate pair from one run is considered identical to a pair from another run if and only if *both* the treatment similarity and the outcome similarity meet or exceed a threshold of 0.8:

$$\mathrm{sim}_{\cos}(\mathbf{t}_i, \mathbf{t}_j) \geq 0.8 \quad \wedge \quad \mathrm{sim}_{\cos}(\mathbf{o}_i, \mathbf{o}_j) \geq 0.8, \tag{2}$$

where $\mathbf{t}$ and $\mathbf{o}$ denote the embedding vectors of the treatment and outcome phrases, respectively.

Consensus filtering. After merging, we retain only those treatment–outcome pairs that appear in at least two out of the three runs ($\geq 2/3$ consensus). This multi-run consensus strategy filters out hallucinated or spurious extractions that arise in only a single run.

Sign assignment. For each retained pair, the directional sign label $s \in \{+, -, \texttt{None}, \texttt{Mixed}\}$ is determined by majority vote across the runs in which the pair appears.

B Multi-Critic Evaluation Details

B.1 Critic Models

Each extracted triplet is independently evaluated by three LLM critic models from different providers to reduce single-model bias:

- **Gemini** (Google): `gemini-3-flash-preview`
- **Grok** (xAI): `grok-4-1-fast-reasoning`
- **Qwen** (OpenRouter): `qwen3-vl-30b-a3b-thinking`

Each critic receives the extracted triplet (treatment, outcome, sign), the verbatim evidence paragraph(s), selected context from the extraction pipeline, and the original paper. Scores from the three critics are averaged per criterion.

B.2 Scoring Rubric (0–3 Scale)

All six criteria are scored on a uniform four-point scale:

- **3** – *Clearly supported / correct.* The aspect is unambiguously correct and well-grounded in the paper.
- **2** – *Mostly supported / minor ambiguity.* The aspect is largely correct but contains minor imprecision or ambiguity.
- **1** – *Weak support / substantial ambiguity.* The aspect has notable issues, e.g., misspecification, vague grounding, or debatable interpretation.
- **0** – *Not supported / contradicted / clearly wrong.* The aspect is unsupported by or contradicts the paper.

B.3 Quality Criteria

1. **Variable Extraction.** Whether the treatment and outcome are extracted as concise, concrete noun phrases explicitly mentioned or defined in the paper, with pronouns, acronyms, or shorthand correctly expanded.
2. **Direction.** Whether the triplet correctly captures the intended causal direction (Treatment $\rightarrow$ Outcome) as asserted by the authors, without reversal due to ambiguous wording.
3. **Sign.** Whether the assigned sign ($+$/$-$/None/Mixed) matches the authors' preferred or baseline estimate. Signs of $+$ or $-$ require statistical significance; `None` applies when the preferred estimate is insignificant; `Mixed` is reserved for genuinely heterogeneous headline results, not mere sensitivity to alternative specifications.
4. **Causality.** Whether the relationship is presented as a causal effect claim supported by an identification strategy (e.g., instrumental variables, difference-in-differences), rather than a correlation, descriptive statistic, or theoretical conjecture.
5. **Main Claim.** Whether the triplet represents a core causal claim emphasized by the authors as a central contribution (e.g., in the abstract, introduction, or conclusion), rather than a peripheral finding.

6. **Context Appropriateness.** Whether the accompanying context includes the key elements required to interpret the causal claim without omitting paper-critical setting information, while avoiding encoding or implying the correctness of the triplet's sign, direction, or causal validity.

B.4 Conservative Filtering

After scoring, we apply a conjunctive filtering rule: a triplet is retained only if (i) its critic-averaged score is at least 2.0 on *every* individual criterion, and (ii) the sum of its six criterion-averaged scores is at least 15 (i.e., a mean of 2.5 or above across all criteria). A triplet failing either condition is discarded. This conservative strategy removes 27.3% of candidate triplets, prioritizing precision over recall in the final dataset.

C Source Paper Distribution

Table 4 reports the distribution of the final source papers by journal and domain. The final corpus contains 2,595 papers, drawn from top economics and finance journals matched from the NBER Working Paper series.

Journal	**# Papers**
Economics journals	
American Economic Review	751
Quarterly Journal of Economics	440
Journal of Political Economy	258
Review of Economic Studies	177
Econometrica	123
Finance journals	
Journal of Financial Economics	356
Journal of Finance	255
Review of Financial Studies	235
Total	**2,595**

Table 4: Distribution of source papers by journal and domain.

D Economist Validation Details

Three economics professors individually assessed **206** triplets within their areas of expertise, using the same six 0–3 criteria described in Appendix B. These expert evaluations allow us to validate both the LLM-critic scores and the downstream Step 4 filtering outcomes.

Expert validation protocol. The validation was conducted by three economics professors: two author-evaluators and one external expert recruited through the authors' professional network. The evaluators were not separately compensated. The external expert was informed that the ratings would be used to validate the dataset construction pipeline and consented to this use.

Validation summary. Table 5 summarizes score-level agreement, economist approval rates, and Step 4 filtering outcomes on the expert-rated triplets. The Step 4 acceptance rate is 76.2%, lower than the economist approval rate of 83.5%, suggesting that the filtering rule is relatively conservative.

Metric	Value
Score-level agreement	
Per-criterion MAE	0.229
Sum-of-scores MAE	1.018
Filtering and approval outcomes	
Economist approval rate before filtering	83.5%
Step 4 acceptance rate	76.2%
Economist approval among Step 4 accepted triplets	86.6%

Table 5: Economist validation summary for LLM-critic scores and Step 4 filtering.

Criterion-level expert scores. Table 6 reports economist scores for each of the six validation criteria. Across all criteria, mean scores are high, and most triplets receive scores of at least 2 out of 3.

Criterion	Mean score	Share ≥ 2
Variable extraction	2.87 / 3	97.57%
Direction	2.99 / 3	99.51%
Sign score	2.93 / 3	97.57%
Causality	2.83 / 3	94.66%
Main claim	2.72 / 3	90.78%
Context	2.94 / 3	98.06%

Table 6: Criterion-level economist validation scores. Mean scores are reported on a 0–3 scale, and the final column reports the share of triplets receiving a score of at least 2.

E Sensitivity Analyses

E.1 Sensitivity to Matching Thresholds

We examine whether the cross-context results are sensitive to the treatment–outcome similarity threshold used to construct matched pairs. Table 7 reports results under stricter matching thresholds.

Task 2 accuracy remains comparable under stricter thresholds, suggesting that the cross-context transfer pattern is not driven by loosely matched treatment–outcome pairs. In contrast,

	Task 2		Task 3	
Threshold	N	Δ Acc.	N	Δ Acc.
0.80	284	baseline	852	baseline
0.85	123	+3.4	369	-5.0
0.90	65	+2.9	195	-10.1

Table 7: Sensitivity to stricter treatment–outcome similarity thresholds, averaged across all models. Accuracy changes are reported in percentage points relative to the 0.80 threshold.

	Task 2		Task 3	
Prompt	Acc.	F1	Acc.	F1
BASE	70.22	45.63	46.92	31.88
SIMPLE	72.24	45.56	49.65	31.75
EXPERT	70.07	44.66	48.34	31.87

Table 8: Prompt sensitivity results for Tasks 2 and 3, averaged across all models. Accuracy and Macro F1 are reported as percentages.

Task 3 accuracy decreases as the threshold becomes stricter, consistent with the injected incorrect source sign becoming a more plausible distractor when source and target pairs are more semantically similar.

E.2 Prompt Sensitivity

We also test whether the cross-context results are sensitive to prompt wording. In addition to the BASE prompt used in the main experiments, we evaluate two variants: SIMPLE and EXPERT. SIMPLE removes role-specific guidance and sign definitions, while EXPERT adds an expert-economist role, explicit sign definitions, and an instruction to use reference examples only as background because causal signs may vary across contexts. The full prompt templates for BASE, SIMPLE, and EXPERT are provided in Appendix H.

The results are broadly stable across prompt variants. SIMPLE improves accuracy by 2.03 percentage points on Task 2 and 2.73 percentage points on Task 3, while Macro-F1 remains nearly unchanged. EXPERT yields similar accuracy to BASE on Task 2 and slightly higher accuracy on Task 3, again without improving Macro-F1. This suggests that the main findings are not driven by a particular prompt wording.

F Accuracy by Various Categories

Both analyses in this section focus on **Task 1**, where models identify the causal sign given a single context, to isolate domain- and time-specific effects from cross-context reasoning.

Task	+	−	*None*	*Mixed*	Overall
Task 1 (Econ)	81.58	65.43	9.89	10.23	73.47
Task 1 (Finance)	81.49	63.48	10.18	21.52	71.42
Task 2	81.38	71.04	16.90	24.60	70.60
Task 3	54.84	44.62	18.36	34.92	47.38

Table 9: Accuracy by Expected Sign and Task

Performance by JEL Category. Performance varies markedly across JEL categories. The top-five fields—History of Economic Thought (B; acc = 86.1%), Economic History (N; 79.7%), Law and Economics (K; 78.7%), Urban, Rural, and Regional Economics (R; 78.1%), and International Economics (F; 77.9%)—share a common characteristic: they typically involve concrete and well-documented causal mechanisms grounded in specific historical episodes, institutional settings, or policy shocks. In contrast, the bottom-five fields—General Economics (A; 59.3%), Mathematical and Quantitative Methods (C; 65.7%), Economic Systems (P; 67.6%), Other Special Topics (Z; 70.2%), and Microeconomics (D; 70.7%)—tend to rely on more abstract or structural reasoning, such as formal modeling, system-level comparisons, or methodological frameworks. Detailed results for all JEL categories are provided in Table 10 of the Appendix.

Performance by Publication Year. Performance in the early 1990s is somewhat volatile (ranging from 70% to 81%), reflecting sparse coverage in early years. From the late 1990s through 2023, when each year is represented by approximately 60 triplets, combined accuracy stabilizes in the 66–81% range, with no systematic decline after 2015. This pattern suggests that model performance is driven more by task structure than by memorization of older studies. Year-level results are shown in Table 11 of the Appendix.

JEL	Field	Accuracy	N
B	History of Economic Thought	86.1%	34
N	Economic History	79.7%	585
K	Law and Economics	78.7%	642
R	Urban/Rural/Regional	78.1%	594
F	International Economics	77.9%	840
L	Industrial Organization	74.8%	801
Q	Agricultural & Natural Resource	74.5%	467
I	Health, Education, Welfare	73.7%	838
J	Labor and Demographic	73.7%	892
O	Economic Development & Tech Change	73.4%	833
M	Business Administration	73.0%	522
H	Public Economics	72.9%	870
E	Macroeconomics	72.2%	871
G	Financial Economics	71.1%	843
D	Microeconomics	70.7%	839
Z	Other Special Topics	70.2%	271
P	Economic Systems	67.6%	365
C	Mathematical & Quantitative	65.7%	709
A	General Economics	59.3%	45

Table 10: Accuracy by JEL code

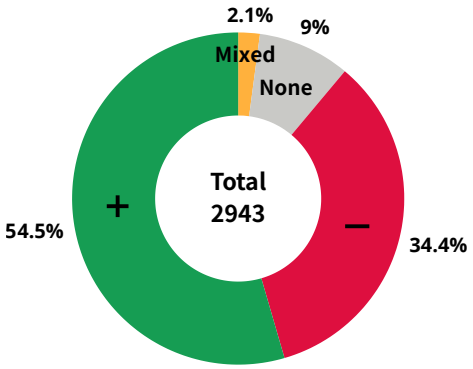


Figure 4: Distribution of ground-truth causal signs in EconCausal benchmark (Task 1/2/3). Each triplet is labeled as positive (+), negative (−), none, or Mixed based on the authors' preferred empirical specification.

Year	Econ	Finance	Combined
1991	78.2%	59.9%	72.9%
1992	76.3%	52.8%	74.0%
1993	67.3%	94.4%	69.9%
1994	79.7%	67.6%	73.8%
1995	90.9%	70.4%	81.2%
1996	75.5%	89.6%	78.3%
1997	91.8%	69.5%	78.7%
1998	70.9%	83.3%	74.9%
1999	70.5%	70.4%	70.3%
2000	72.9%	74.8%	73.8%
2001	70.5%	74.6%	72.5%
2002	67.6%	64.1%	65.8%
2003	66.3%	70.7%	68.5%
2004	78.5%	68.6%	73.5%
2005	73.8%	78.0%	75.8%
2006	69.6%	78.7%	74.1%
2007	83.3%	79.2%	81.3%
2008	76.4%	60.9%	68.6%
2009	71.5%	73.5%	72.5%
2010	73.8%	65.9%	70.1%
2011	76.5%	65.7%	70.9%
2012	73.9%	75.4%	74.7%
2013	73.7%	73.9%	73.7%
2014	74.6%	65.2%	69.8%
2015	79.2%	75.6%	77.4%
2016	69.1%	70.1%	69.6%
2017	69.8%	63.5%	66.7%
2018	67.1%	74.5%	70.9%
2019	74.0%	62.4%	67.6%
2020	74.2%	68.7%	71.4%
2021	73.5%	71.2%	72.5%
2022	73.1%	73.2%	73.1%
2023	63.5%	72.3%	67.7%
2024	—	90.1%	90.1%
2025	72.8%	83.3%	73.9%

Table 11: Accuracy by Publication Year

G Causality Extraction Pipeline Prompts

Role: You are an expert Research Assistant in Economics and Finance, specialized in causal inference and systematic literature reviews.

Goal: From the provided economics paper text, extract all causal effect claims explicitly asserted by the authors, and return them as a JSON array of objects.

Definitions & Scope:

- **Causal Claim**: A statement in which the authors argue that a change in X (Treatment) causes a change in Y (Outcome).
- **Strict Exclusion**: Exclude simple correlations, descriptive statistics, predictive relationships, or purely theoretical conjectures.
- **Identification Focus**: Prioritize claims supported by empirical identification strategies (e.g., DiD, IV, RDD, Fixed Effects, or RCT).
- **No Context at This Stage**: Extract only the treatment and outcome variables as abstract concepts. Do not include who/when/where, population, or institutional context.
- **Expand Shorthand**: If treatments or outcomes use pronouns, acronyms, or shorthand labels, recover the fully specified variable as originally defined.

Extraction Rules:

(1) **treatment**: A concise noun phrase for the independent variable (<10 words).

(2) **outcome**: A concise noun phrase for the dependent variable (<10 words).

(3) **sign**: Direction of the causal effect—`'+'` (significant increase), `'-'` (significant decrease), `'None'` (not significant), or `'Mixed'` (heterogeneous or conflicting headline results). Assign based only on the authors' preferred/main result.

(4) **supporting_evidence**: Up to three verbatim paragraphs reporting the main results or causal conclusions.

Constraints: Verbatim only for evidence. No hallucination. Output valid JSON array only.

Figure 5: Prompt for Step 1: Triplet Extraction with Multi-run Consensus

Role: You are an expert Research Assistant in Economics.

Goal: Extract paper-level metadata, research context, and identification strategies into a structured JSON format.

Task Instructions:

(1) **Metadata**: Classify the paper type as "`empirical`" (uses data/estimation for causal claims) or "`theoretical`" (develops models/theory without empirical estimation).

(2) **Global Context**: Produce one cohesive natural-language paragraph (≤100 words) summarizing the paper-wide context underlying the main causal claims. Integrate when available:

- **When**—time period, years, cohorts, event timing;
- **Where**—geography, country/region, market, industry, institutional setting;
- **Who/Unit**—unit of observation (individuals, households, firms, regions, etc.);
- **Background**—key policy environment, institutional details, economic setting.

(3) **Identification Methods**: Identify all empirical identification methods supporting causal claims, mapped to allowed values: `[DiD, IV, RCT, RDD, event studies, synthetic control, PSM, other panel regressions, other time-series regressions, other cross-sectional regressions, others]`.

Constraints: Output only valid JSON. If paper is theoretical, identification methods must be `[]`.

Figure 6: Prompt for Step 2: Paper-level Metadata, Global Context, and Identification Methods

Role: You are an expert Research Assistant in Economics.

Goal: Perform local review for a single extracted causal triplet. You are given: an extracted causal triplet (treatment, outcome, sign), verbatim evidence paragraphs, a paper-wide global context paragraph, a paper-wide list of identification methods, and the original paper.

A) Context Review.
Default: return `[]` (keep global context).
Rewrite global context into a triplet-specific paragraph **only if** the paper clearly shows the global context is not valid for this triplet because:

- the triplet's context is more specific or different, **or**
- some elements in global context are not applicable.

If rewriting, apply the **minimum-edit rule**: keep parts that still apply, remove/adjust only inapplicable parts, add/narrow details only when explicitly supported. Output $\leq$100 words.

B) Identification-Method Review.
Default: return `[]` (keep global identification methods).
Output a non-empty list only if the evidence explicitly shows the triplet uses an identification method that is different from, stricter than, or missing from the global list.
Map each method to one of: `[DiD, IV, RCT, RDD, event studies, synthetic control, PSM, other panel regressions, other time-series regressions, other cross-sectional regressions, others]`.

Strict Rules: Use only provided data. Do not guess or extrapolate. If no clear difference, return empty lists. Output valid JSON only.

Figure 7: Prompt for Step 3: Local Selection of Triplet-specific Context and Method

Role: You are a causal-inference critic reviewing an extracted causal relation from an Economics/Finance paper.

Goal: Given an extracted causal triplet and supporting materials, evaluate whether the triplet is correctly extracted and score it on six criteria (0–3 each).

Inputs: Extracted triplet (treatment, outcome, sign), verbatim evidence paragraphs, selected context, and the original paper.

Scoring Rubric:
3 = clearly supported/correct; **2** = mostly supported/minor ambiguity; **1** = weak support/substantial ambiguity; **0** = not supported/contradicted/clearly wrong.

Dimensions:

(1) **variable_extraction**: Are treatment and outcome extracted as concise, concrete noun phrases explicitly defined in the paper? Are shorthand references correctly expanded?

(2) **direction**: Does the triplet correctly capture the intended causal direction (Treatment $\rightarrow$ Outcome) without reversal?

(3) **sign**: Does the sign (`+/-/None/Mixed`) match the authors' preferred/main estimate? Is `+/-` used only if statistically significant, `None` if not significant, and `Mixed` only for truly heterogeneous headline results?

(4) **causality**: Is the relationship a causal effect claim supported by an identification strategy, excluding correlations, descriptive statistics, or theoretical conjectures?

(5) **main_claim**: Is the triplet a core causal claim emphasized by the authors (e.g., headline finding in abstract/introduction/conclusion), rather than a peripheral finding?

(6) **context_appropriateness**: Does the context include key elements for interpreting the causal claim without encoding or implying the triplet's correctness?

Output: One valid JSON object with the triplet, scores (0–3 per dimension), and one-sentence justifications per dimension.

Figure 8: Prompt for Step 4: Multi-critic Evaluation and Conservative Filtering

H EconCausal LLM Task Prompts

 # Role
You are an expert economist.

 # Task
Given a context and a treatment–outcome pair, predict the most likely sign of the causal effect.

 # Sign Definitions
- '+': The treatment increases the outcome (positive and statistically significant effect).
- '-': The treatment decreases the outcome (negative and statistically significant effect).
- 'None': No statistically significant effect.
- 'Mixed': The effect varies across subgroups or specifications.

 # Context
{context}

 # Treatment–Outcome Pair
Treatment: {treatment}
Outcome: {outcome}

 # Output Format
Respond with a JSON object containing:
- predicted_sign: "+", "-", "None", or "Mixed"
- reasoning: A concise explanation of your reasoning

Figure 9: Evaluation prompt for Task 1: Causal Sign Identification.

 # Role
You are an expert economist.

 # Task
You are given examples in which the same or comparable treatment–outcome pair is observed across multiple contexts, potentially with different causal signs. Predict the most likely sign for the treatment–outcome pair in the target context.

 # Sign Definitions
- '+': The treatment increases the outcome (positive and statistically significant effect).
- '-': The treatment decreases the outcome (negative and statistically significant effect).
- 'None': No statistically significant effect.
- 'Mixed': The effect varies across subgroups or specifications.

 # Reference Examples
{examples}

 # Target Context
{context_new}

 # Target Treatment–Outcome Pair
Treatment: {treatment}
Outcome: {outcome}

 # Output Format
Respond with a JSON object containing:
- predicted_sign: "+", "-", "None", or "Mixed"
- reasoning: A concise explanation of your reasoning

Figure 10: Evaluation prompt for Task 2 and Task 3: Context-Dependent Sign Prediction and Misinformation-Robust Sign Prediction

 # Task
Predict the causal sign of the treatment–outcome pair in the target context.

 # Reference Examples
{examples}

 # Target
Treatment: {treatment}
Outcome: {outcome}
Context: {context_new}

 # Output Format
Respond with a JSON object containing:
- predicted_sign: "+", "-", "None", or "Mixed"
- reasoning: A concise explanation of your reasoning

Figure 11: SIMPLE prompt variant used in the prompt sensitivity analysis.

 # Role
You are an expert economist.

 # Task
You are given examples related to the target treatment–outcome pair, potentially with different causal signs. Predict the most likely sign for the treatment–outcome pair in the target context. Because causal effects can vary across contexts, use the reference examples as background only and infer the most likely sign under the target context.

 # Sign Definitions
- '+': The treatment increases the outcome (positive and statistically significant effect).
- '-': The treatment decreases the outcome (negative and statistically significant effect).
- 'None': No statistically significant effect.
- 'Mixed': The effect varies across subgroups or specifications.

 # Reference Examples
{examples}

 # Target
Treatment: {treatment}
Outcome: {outcome}
Context: {context_new}

 # Output Format
Respond with a JSON object containing:
- predicted_sign: "+", "-", "None", or "Mixed"
- reasoning: A concise explanation of your reasoning

Figure 12: EXPERT prompt variant used in the prompt sensitivity analysis.

I Implementation details.

We implemented all experiments using custom Python scripts. API-based inference used `openai==2.14.0` and `google-genai==1.66.0`, while open-source inference used `vllm==0.19.0`, `torch==2.10.0`, `transformers==4.57.6`, and `huggingface_hub==1.7.1`. Data processing and evaluation relied on `pandas==3.0.0`, `numpy==2.4.2`, and `scipy==1.17.0`.