Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ29689

Bridging the Reliability Gap: Challenges and Prospects for Large Language Models in Economic Causal Inference

Wenzhuo Wang

Faculty of Humanities and Social Sciences of Arts, University of Nottingham Ningbo, Ningbo, China
hmyww2@nottinngham.edu.cn

Abstract. Large Language Models (LLMs) are driving a paradigm shift in economic causal inference, thereby enabling the direct quantification of causal effects from unstructured text. However, this transformation comes with a significant reliability gap. Existing approaches, whether using text as a proxy, extracting causal chains, or treating LLMs as world models, are constrained by three interconnected challenges: persistent confounding, a lack of robust validation standards, and limited interpretability. Through a review of more than 30 studies in text analysis, causal science, and computational economics, the results show that, unless the reliability gap is directly addressed, LLMs are likely to remain promising black boxes and cannot yet serve as reliable tools for policy analysis or scientific discovery. To enhance credibility, research efforts should go beyond exploring model capabilities, and reliability can be improved via a multi-pronged approach involving hybrid models, human-machine collaboration, and Explainable AI (XAI). Consequently, the paper aims to guide this critical transition and future research to develop reliable and accountable LLMs for economics.

Keywords: Large Language Models (LLMs), Text Analysis, Causal Inference, Economics, Explainable Artificial Intelligence (XAI)

1. Introduction

The emergence of digital text and large language models (LLMs) is reshaping economic research. Traditionally, economists have mostly relied on structured data and causal analysis methods, such as difference-in-differences (DiD), instrumental variables (IV), regression discontinuity designs (RDD), and randomized controlled trials (RCTs), rather than on simple correlations [1]. However, economic information embedded in unstructured text, including central bank speeches, corporate filings, and news, has often been omitted from causal analysis due to processing and interpretation challenges [2,3]. This landscape has shifted with the advent of powerful NLP systems, notably Transformer-based models including BERT and GPT-3 [4-6]. These models can identify, infer, and even simulate causality from text, making it an important source of causal information and, in some cases, functioning as a "world model" for economic behavior [7]. Thus, new avenues have emerged, including market reactions to central bank announcements and corporate responses to political risk [8-10]. However, this promising shift is constrained by a significant reliability gap. Unstructured text can amplify confounding and misleading associations, robust validation benchmarks are often lacking, and the black-box nature of LLMs obscures internal mechanisms and accountability. This

DOI: 10.54254/2755-2721/2026.TJ29689

review aims to examine the potential and challenges of large language models in economic causal analysis and to propose strategies for narrowing the reliability gap. To address these challenges, a framework and roadmap are proposed that combine hybrid modeling with explainable artificial intelligence (XAI) to improve reliability and ensure responsible use of LLMs in economics.

2. The evolution of text analysis in economic causal inference

The utilization of textual information in economic research has advanced alongside computational linguistics and AI, and it has followed three interconnected approaches, as shown in Figure 1. These include using text as a substitute for economic indicators, automatically detecting causality from text, and using LLMs to perform causal reasoning and simulation directly.

2.1. Paradigm I: text as a proxy for economic variables

This approach transforms unstructured text into quantitative proxies for use in formal econometric models. It mainly converts complex qualitative narratives into quantifiable data, extending beyond the direct extraction of causal relationships from the text.

The paradigm has gained prominence by enabling the quantification of previously unmeasurable concepts, like economic uncertainty, exemplified by the Economic Policy Uncertainty (EPU) Index [11]. This method has been tailored to specific contexts, for instance an index for Russia's economy, and applied to monetary policy via forward guidance extracted from central bank communications to study policy impacts [8,9,12]. At the firm level, a similar analysis of earnings call transcripts has measured political risk and linked it to changes in corporate behavior [10]. These techniques utilize natural language processing (NLP) tools such as topic modeling and sentiment analysis to generate numerical representations of text [13,14]. Despite some progress in understanding latent concepts, this approach has drawbacks. The main issue lies in the inherent separation between text processing and causal inference, thus making the quality of the resulting proxy variable critical for valid causal conclusions. The method cannot directly control latent confounders in the text and relies on external econometric methods to handle endogeneity and confounding. This paradigm quantifies qualitative insights well but needs a sound econometric framework downstream for valid causal inference.

2.2. Paradigm II: automated discovery of causal chains from text

The second paradigm shifts from quantifying variables to deriving causal relationships from text, using explicit cause-effect claims in sources such as news and financial reports to build structured representations of economic mechanisms. Initial research mined financial news for causal signals to construct databases and trace the propagation of events, with these causal chains later incorporated into machine learning models, thus moving the emphasis from static indicators to dynamic, causally informed forecasting [15,16]. Recently, LLMs have been applied to thousands of papers, and in one study over 44,000 working papers were analyzed to extract causal claims and flag those supported by rigorous methods [1]. However, this paradigm confronts substantial obstacles. In textual data, a major challenge lies in separating genuine causal relationships from mere correlations or sequential patterns [1]. Moreover, the complexity of language presents a related difficulty. For instance, vague expressions and technical jargon can reduce the model's ability to accurately identify and interpret causal relationships [14,17,18]. This shows that the richness of narrative is meaningful only when it is supported by credible causal identification.

DOI: 10.54254/2755-2721/2026.TJ29689

2.3. Paradigm III: LLMs for simulation and causal reasoning

LLMs provide a third transformative method, identifying presuppositional relations and functioning as flexible "world models" to simulate economic behavior and perform direct causal reasoning. This approach harnesses the extensive knowledge embedded in models like GPT3 and Large Language Model Meta AI (LLaMA) to enable novel analyses of economic conditions [5,6,19]. For example, LLMs can be used as virtual economic agents. When provided with specific goals and data, they can inexpensively simulate behavioral experiments, enabling rapid testing of how behavior shifts across different conditions [7]. At the macroeconomic level, systems such as EconAgent employ LLM-driven assistants to model complex interactions and policy effects, incorporating individual heterogeneity often ignored by conventional models [20,21]. These models also exhibit the ability for direct "zero-shot causal reasoning,", which can be enhanced via fine-tuning on domain-specific corpora like economic research literature [19,22-24]. However, this approach still faces key limits. LLMs reason based on statistical patterns in their training data rather than a true understanding of causality [25]. Research shows that their reasoning can be unreliable; while the conclusions may be correct, the underlying logic can be flawed or merely reflect correlations, forming a clear reliability gap that poses a major obstacle to their widespread adoption in economics [26].

3. The reliability gap: key challenges in text-based causal inference

Though LLMs are powerful, they have a reliability gap. They rely on statistical patterns rather than a true understanding of causality, making them prone to errors in situations requiring strict logical reasoning. This gap involves three interrelated issues, leaving LLM conclusions unreliable for key policy decisions and scientific research.

3.1. Challenge I: the confounding bias and spurious associations

As text is inherently observational, it is strongly susceptible to confounding. In addition, economic narratives are influenced by numerous unobserved factors, such as author bias, political context, or concurrent events, which LLMs are not designed to identify or control. Traditional econometrics has devised various tools, like instrumental variables and control functions, to handle confounding; however, applying them to unstructured text and opaque neural network models remains difficult [1]. LLMs learn from co-occurrence patterns in language, making them prone to detecting spurious associations. For instance, if a policy announcement and a market downturn occur simultaneously, an LLM might incorrectly conclude that the policy caused the downturn, while both might actually result from an unmentioned underlying economic crisis. This problem is intensified by the linguistic complexity of economic texts, where sentiment, tone, and context can act as latent confounders that are challenging to model explicitly [14,27]. Moreover, LLMs are known to "hallucinate" plausible yet factually incorrect causal narratives, as they depend on statistical patterns instead of real domain expertise [24]. In the absence of formal confounder adjustment, a key principle of the credibility revolution in economics, LLM causal claims are fundamentally unreliable [28].

3.2. Challenge II: the validation vacuum

The second challenge is the lack of "gold standard" benchmarks to validate causal inferences from text. In traditional econometrics, causal claims are evaluated against real-world data or controlled experimental results. Nevertheless, the construction of reliable ground truth is particularly difficult for text-based causal inference. This results in a validation vacuum, preventing rigorous assessment

DOI: 10.54254/2755-2721/2026.TJ29689

of LLM-generated causal claims. Although certain datasets have been created to assess economic reasoning, they are often narrow in scope and emphasize logical classification rather than rigorous causal identification from complex narratives [24].

The lack of uniform, broad evaluation criteria means that models are often evaluated ad hoc, thus preventing meaningful performance comparisons and reducing confidence in their outputs when faced with new situations [22,29]. This problem is particularly evident in zero-shot scenarios, where LLMs draw conclusions without specific fine-tuning for specific areas [16]. For example, an LLM may concoct a persuasive explanation of the causes of inflation, but without a rigorous verification process, it is unclear whether the explanation has a practical basis or is merely a plausible story [23]. This shortcoming hinders the establishment of reliable systems, and solutions such as model cards are rarely used in economics because of the lack of specially designed verification indicators in the field [25,29].

3.3. Challenge III: the black box and the need for explainability

The third key challenge lies in the inherent opacity of large language models. Due to their highly complex architecture, often comprising billions of parameters, it is virtually impossible to discern how these models reach their conclusions. In a field like economics, where decision-making hinges on clear and verifiable reasoning, this black-box nature poses a fundamental barrier to acceptance [30]. Even if a prediction from an LLM is correct, its output cannot be trusted without transparency in the underlying logic [21]. In addition, this opacity also prevents the diagnosis of errors and biases. Past studies on chain-of-thought reasoning have shown that LLMs can produce justifications that are unfaithful to the model's actual inference process [26]. This is particularly unacceptable in economics, where causal claims must be supported by a rigorous and transparent chain of logic. As a result, the demand for XAI goes beyond technical preference, serving as a key requirement for building trust and maintaining accountability [14,31]. Without robust methods to open the black box and scrutinize the model, LLMs remain an unconvincing tool for serious economic analysis.

4. Bridging the reliability gap: strategies and technical pathways

A strategic change in research is needed to close the "reliability gap" caused by inherent opacity, a validation vacuum, and persistent confounding bias. By linking LLM strengths with econometric rigor, this shift goes beyond testing LLMs to firmly proving their trustworthiness. According to Figure 1, three main strategies use domain XAI to open the black box, human-machine help to close the validation gap, and hybrid models to fix confounding

DOI: 10.54254/2755-2721/2026.TJ29689

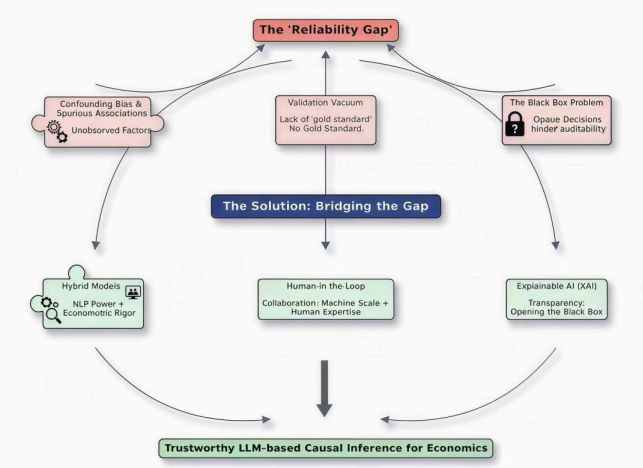

Figure 1. Bridging the reliability gap and proposed solutions

4.1. The mitigation of confounding bias through mixed modeling

When using text as observational data, confounding bias presents an inherent challenge, particularly for LLMs, which are prone to detecting spurious associations owing to their reliance on statistical co-occurrences. Given the “black box” nature of these models, traditional econometric tools cannot be directly applied to their internal mechanisms to address this problem. By combining the strengths of LLMs with the methodological rigor of traditional econometrics, the hybrid approach provides a practical solution. This approach primarily employs a division of labor, with LLMs first extracting latent features, cues, and contextual information from unstructured text, and then feeding these variables into established causal inference tools, such as DiD, IV, and VAR, to obtain more reliable estimates of causal effects [22,27]. For instance, an LLM can be used to analyze documents issued by a central bank, generating a detailed policy sentiment indicator that is then incorporated into a VAR framework to examine its impact on inflation, while controlling for other macroeconomic variables [11,28]. If these models are trained on domain-specific texts written by economic experts, the extracted signals will be more precise and better at capturing economic-related information [19]. This division of labor allows researchers to leverage advanced NLP tools without compromising the strict criteria required for credible causal inference. However, hybrid methods pose challenges. LLMs pick up or amplify biases in the input text while generating features, potentially leading to unobserved confounders. Additionally, validating the credibility of these black-box-derived proxies is challenging, as unreliable proxies can bias results. Besides, potential endogeneity between LLM feature extraction and later causal identification must be addressed, as it can undermine the entire strategy.

4.2. The remediation of validation gaps through human-AI collaboration

In the absence of “gold standard” benchmarks, fully automated methods cannot reliably validate causal claims produced by LLMs. Therefore, a human-in-the-loop framework stands as the most viable strategy [1]. This approach uses the expertise of human economists to guide, evaluate, and refine LLM outputs, combining algorithmic speed with essential human judgment. For example, a practical workflow involves using LLMs to rapidly generate causal hypotheses or identify potential relationships from large volumes of text. Subsequently, the machine-generated results are subject to thorough assessment [22]. Based on economic theory and existing research data, these findings can

DOI: 10.54254/2755-2721/2026.TJ29689

be revised, confirmed, or enhanced. This method has been proven to enhance the effectiveness of models in different fields [32]. Furthermore, interactive tools can further refine the causal networks generated by the model [15]. This collaborative approach turns validation into an active, iterative exchange between artificial intelligence and domain expertise, ensuring that the final conclusions are both computationally robust and endorsed by experts. Though this paradigm demonstrates great potential by integrating computational power with human expertise, it still has inherent limitations. In particular, the main issues concern scalability and potential reviewer bias. Moreover, relying on detailed evaluations by domain experts is both time-consuming and costly, making it impractical for research involving massive text corpora. Besides, experts' biases and perspectives may enter the validation process, affecting the objectivity of the conclusions. Therefore, the design of an efficient workflow that limits expert bias is a central challenge for future research

4.3. The decoding of the black box through explainable artificial intelligence

Given the inherent opacity of LLMs and the need for transparent and verifiable causal reasoning in economics, applying XAI methods to LLMs is an urgent task [1]. Thus, it seeks to make the models' reasoning process understandable, ensuring that causal judgments are verifiable and reliable.

To ensure step-by-step reasoning before a conclusion, an effective approach is chain-of-thought prompting. If models can generate "faithful" chains that accurately reflect their decision-making paths, their conclusions become more transparent and credible [26]. For example, an LLM could be required to outline the theoretical economic mechanism through which a policy intervention affects a specific outcome. Another powerful XAI method is leveraging the attention mechanisms inherent to the Transformer architecture of most LLMs [4]. By visualizing which words or phrases in a text the model paid the most attention to when making a causal inference, researchers can gain insight into the textual evidence driving the conclusion.

However, generating explanations alone is insufficient to establish trust, and future studies must develop operational metrics to validate the reliability of these explanations. Existing studies have shown that LLMs can produce "unreliable" justifications that fail to reflect their true reasoning [26]. Thus, a critical challenge is to establish methods for measuring the fidelity of these explanations, ensuring they are not only plausible but truthful representations of the model's decision-making. The continued development of domain-specific XAI tools is essential to transform LLMs from opaque black boxes into transparent, trustworthy partners in economic research [17,31].

5. Conclusion

This paper reviews text-based methods for economic causal inference, from early attempts using text as economic proxies, to automated causal extraction, and finally to the use of LLMs for direct inference and simulation. The study finds that although LLMs are powerful, their reliability remains limited. This is mainly reflected in confusion bias, difficulties in verification, and the opacity of their decision-making processes. These issues arise because LLMs rely on statistical patterns rather than rigorous econometric principles, making it difficult to meet requirements for interpretability. To narrow this reliability gap, LLMs can be combined with established econometric methods to reduce bias and spurious correlations, while promoting human-machine collaboration to rigorously validate hypotheses and using explainable AI techniques to enhance decision-making transparency, thereby improving credibility and accountability. In addition, these approaches require consistent evaluation, reduced text noise, and mitigation of algorithmic biases. To render LLMs reliable for economic analysis and policy work, they should they should be tailored for economic applications and

DOI: 10.54254/2755-2721/2026.TJ29689

integrated with econometric methods, which requires close collaboration between NLP experts and economists.

References

[1] Garg, P., & Fetzer, T. (2025). Causal claims in economics. https: //www.causal.claims/
[2] Gentzkow, M., Kelly, B. T., & Taddy, M. (2017). Text as data (No. 23276). http: //www.nber.org/papers/w23276
[3] Ash, E., & Hansen, S. (2023). Text algorithms in economics. Annual Review of Economics, 15, 659-688.
[4] Vaswani, A., et al. (2017). Attention is all you need. In 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.
[5] Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT 2019, 4171-4186.
[6] Brown, T. B., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv: 2005.14165
[7] Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? arXiv preprint arXiv: 2301.07543v1.
[8] Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. Journal of International Economics, 99, S114-S133.
[9] Ahrens, M., & McMahon, M. (2021). Extracting economic signals from central bank speeches. In Proceedings of the Third Workshop on Economics and Natural Language Processing, 93-114.
[10] Hassan, T. A., Hollander, S., van Lent, L., & Tahoun, A. (2019). Firm-level political risk: Measurement and effects. NBER Working Paper Series.
[11] Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. Journal of Economic Perspectives, 30(2), 71-96.
[12] Charemza, W., Makarova, S., & Rybiński, K. (2023). Economic uncertainty and natural language processing: The case of Russia. Economic Analysis and Policy.
[13] Khalil, F., & Pipa, G. (2021). Is deep-learning and natural language processing transcending financial forecasting? Investigation through lens of news analytic process. Computational Economics, 60, 147-171.
[14] Alam, M. S., Mrida, M. S. H., & Rahman, M. A. (2025). Sentiment analysis in social media: How data science impacts public opinion knowledge integrates natural language processing (NLP) with artificial intelligence (AI). American Journal of Scholarly Research and Innovation, 4(1), 63-100.
[15] Izumi, K., & Sakaji, H. (2019). Economic causal-chain search using text mining technology. In Proceedings of the First Workshop on Financial Technology and Natural Language Processing (FinNLP@IJCAI 2019), 61-65.
[16] Izumi, K., Sano, H., & Sakaji, H. (2023). Economic causal-chain search and economic indicator prediction using textual data.
[17] Ettaleb, M., Moriceau, V., Kamel, M., & Aussenac-Gilles, N. (2025). The contribution of LLMs to relation extraction in the economic field. In Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, 175-183.
[18] Takala, P., Malo, P., Sinha, A., & Ahlgren, O. (2023). Gold-standard for topic-specific sentiment analysis of economic texts. Journal of Information Science, 49(6), 2152-2167.
[19] Keleş, O., & Bayraklı, Ö. T. (2024). LLaMA-2-econ: Enhancing title generation, abstract classification, and academic Q&A in economic research. In Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, Valletta, Malta: ELRA Language Resource Association, 212-218.
[20] Li, N., Gao, C., Li, M., Li, Y., & Liao, Q. (2024). EconAgent: Large language model-empowered agents for simulating macroeconomic activities. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 15523-15536.
[21] Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI economist: Taxation policy design via two-level deep multiagent reinforcement learning. Science Advances, 8(18), eabk2607.
[22] Wan, G., Lu, Y., Wu, Y., Hu, M., & Li, S. (2025). Large language models for causal discovery: Current landscape and future directions. arXiv preprint arXiv: 2402.11068v2 [cs.CL].
[23] Gueta, A., Feder, A., Gekhman, Z., Goldstein, A., & Reichart, R. (2025). Can LLMs learn macroeconomic narratives from social media? Findings of the Association for Computational Linguistics: NAACL 2025, 57-78.
[24] Guo, Y., & Yang, Y. (2024). Evaluating large language models on economics reasoning. In Findings of the Association for Computational Linguistics: ACL 2024, 5, 982-994.
[25] Li, X., Cai, Z., Wang, S., Yu, K., & Chen, F. (2025). A survey on enhancing causal reasoning ability of large language models. arXiv preprint arXiv: 2503.09326, abs/2503.09326 v1.

DOI: 10.54254/2755-2721/2026.TJ29689

[26] Paul, D., West, R., Bosselut, A., & Faltings, B. (2024). Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2024, Mexico City, Mexico, 15012-15032.
[27] Feder, A., et al. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. Transactions of the Association for Computational Linguistics, 10, 1138-1158.
[28] Jantscher, M., & Kern, R. (2022). Causal investigation of public opinion during the COVID-19 pandemic via social media. In Proceedings of the 13th Conference on Language Resources and Evaluation, 211-226.
[29] Mitchell, M., et al. (2019). Model cards for model reporting. In FAT'19: Conference on fairness, accountability, and transparency* (p. 10). ACM.
[30] Dell, M. (2024). Deep learning for economists. NBER Working Paper Series, No. 32768. http://www.nber.org/papers/w32768
[31] Mumuni, F., & Mumuni, A. (2025). Explainable artificial intelligence (XAI): From inherent explainability to large language models.
[32] Howell, K., et al. (2023). The economic trade-offs of large language models: A case study. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 5, 248-267.