← Go to **ACL ARR 2024 October** homepage (/group?id=aclweb.org/ACL/ARR/2024/October)

Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey

(/pdf?id=76qJ1ZaD4m)

Xiaoyu Liu (/profile?id=~Xiaoyu_Liu3), Paiheng Xu (/profile?id=~Paiheng_Xu1), Junda Wu (/profile?id=~Junda_Wu1), Jiaxin Yuan (/profile?id=~Jiaxin_Yuan1), Yifan Yang (/profile?id=~Yifan_Yang5), Yuhang Zhou (/profile?id=~Yuhang_Zhou1), Fuxiao Liu (/profile?id=~Fuxiao_Liu1), Tianrui Guan (/profile?id=~Tianrui_Guan1), Haoliang Wang (/profile?id=~Haoliang_Wang1), Tong Yu (/profile?id=~Tong_Yu3), Julian McAuley (/profile?id=~Julian_McAuley1), Wei Ai (/profile?id=~Wei_Ai1), Furong Huang (/profile?id=~Furong_Huang1)

13 Oct 2024 (modified: 20 Dec 2024) ACL ARR 2024 October Submission October, Senior Area Chairs, Area Chairs, Reviewers, Authors, Commitment Readers Revisions (/revisions?id=76qJ1ZaD4m)

Abstract:
Causal inference has demonstrated significant potential to enhance Natural Language Processing (NLP) models in areas such as predictive accuracy, fairness, robustness, and explainability by capturing causal relationships among variables. The rise of generative Large Language Models (LLMs) has greatly impacted various language processing tasks. This survey focuses on research that evaluates or improves LLMs from a causal view in the following areas: reasoning capacity, fairness and safety issues, explainability, and handling multimodality. Meanwhile, LLMs can assist in causal inference tasks, such as causal relationship discovery and causal effect estimation, by leveraging their generation ability and knowledge learned during pre-training. This review explores the interplay between causal inference frameworks and LLMs from both perspectives, emphasizing their collective potential to further the development of more advanced and robust artificial intelligence systems.

Paper Type: Long
Research Area: Machine Learning for NLP
Research Area Keywords: causality
Contribution Types: Surveys
Languages Studied: English
Reassignment Request Action Editor: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: N/A
B Use Or Create Scientific Artifacts: No
B1 Cite Creators Of Artifacts: N/A
B2 Discuss The License For Artifacts: N/A
B3 Artifact Use Consistent With Intended Use: N/A
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A
B5 Documentation Of Artifacts: N/A
B6 Statistics For Data: N/A

C Computational Experiments: No
C1 Model Size And Budget: N/A
C2 Experimental Setup And Hyperparameters: N/A
C3 Descriptive Statistics: N/A
C4 Parameters For Packages: N/A
D Human Subjects Including Annotators: No
D1 Instructions Given To Participants: N/A
D2 Recruitment And Payment: N/A
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: N/A
E Ai Assistants In Research Or Writing: No
E1 Information About Use Of Ai Assistants: N/A
Reviewing Volunteers: Paiheng Xu (/profile?id=~Paiheng_Xu1), Yuhang Zhou (/profile?id=~Yuhang_Zhou1)
Reviewing No Volunteers Reason: N/A - An author was provided in the previous question.
Reviewing Volunteers For Emergency Reviewing: The volunteers listed above are only willing to serve as regular reviewers.
Preprint: no
Preprint Status: There is a non-anonymous preprint (URL specified in the next question).
Existing Preprints: https://arxiv.org/abs/2403.09606 (https://arxiv.org/abs/2403.09606)
Preferred Venue: NAACL 2025
Consent To Share Data: yes
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
Author Submission Checklist: I confirm that the paper is anonymous and that all links to data/code repositories in the paper are anonymous., I confirm that the paper has proper length (Short papers: 4 content pages maximum, Long papers: 8 content pages maximum, Ethical considerations and Limitations do not count toward this limit), I confirm that the paper is properly formatted (Templates for *ACL conferences can be found here: https://github.com/acl-org/acl-style-files (https://github.com/acl-org/acl-style-files).)
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I agree
Submission Number: 450

Discussion (/forum?id=76qJ1ZaD4m#discussion)

Filter by reply type... | Filter by author... | Search keywords...

Sort: Newest First

Everyone | Submission450... | Submission450 Area... | Submission450 Authors

10 / 11 replies shown

Submission450... | Program Chairs | Submission450... | Submission450... | Submission450...

Submission450... | Submission450... | Submission450...

Add: **Author-Editor Confidential Comment** **Withdrawal**

Meta Review of Submission450 by Area Chair PpZT

Meta Review by Area Chair PpZT 06 Dec 2024, 09:39 (modified: 20 Dec 2024, 12:23)

Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers

Revisions (/revisions?id=LoESwTEp5A)

Metareview:
This is a survey of the intersection between causal inference and LLMs. It examines two main themes: (1) using causal inference to improve LLMs in areas such as reasoning, fairness, safety, and explainability, and (2) leveraging LLMs to enhance causal inference tasks like causal relationship discovery and treatment effect estimation. The survey analyzes key trends, challenges, and opportunities.

Summary Of Reasons To Publish:
The intersection of causal inference and LLMs addresses pressing challenges in NLP, including reasoning, fairness, safety, and interpretability.

The paper organizes a wide range of recent developments, which could be useful for researchers interested in the intersection of causality and LLMs.

Summary Of Suggested Revisions:
- Clarify the methodology for paper selection and categorization to address concerns about transparency and improve the rigor of the survey process.
- Expand on underrepresented areas, such as Event Causality Identification and multimodal applications, by incorporating additional works and discussing their relevance more deeply.
- Provide stronger insights and discussions that go beyond summarizing existing literature, highlighting novel trends, challenges, and opportunities specific to LLMs' generative capabilities.
- Address the overlap with Feder et al.'s 2022 survey, emphasizing how this work builds incrementally on their framework while contributing updated findings on LLM advancements and new causal inference applications.

Overall Assessment: 3 = There are major points that may be revised
Best Paper Ae: No
Ethical Concerns:
There are no concerns with this submission

Author Identity Guess: 3 = From the contents of the submission itself, I know/can guess at least one author's name.
Poor Reviews: Jfts
Reported Issues: No

Add: **Author-Editor Confidential Comment**

Official Comment by Authors

Official Comment

by Authors (Tianrui Guan (/profile?id=~Tianrui_Guan1), Fuxiao Liu (/profile?id=~Fuxiao_Liu1), Haoliang Wang (/profile?id=~Haoliang_Wang1), Paiheng Xu (/profile?id=~Paiheng_Xu1), +9 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission450/Authors))

25 Nov 2024, 08:14 (modified: 02 Jan 2025, 08:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors, Commitment Readers

Revisions (/revisions?id=ZZPtrIRhJV)

Comment:
We thank all reviewers for their valuable feedback and insightful questions. We are especially encouraged by their recognition of the importance of the intersection between causal inference and LLMs, as well as their appreciation of the paper’s organization and exploration of this field. Individual questions are addressed in separate responses.

Add: **Author-Editor Confidential Comment**

Official Review of Submission450 by Reviewer Jfts

Official Review by Reviewer Jfts 21 Nov 2024, 22:37 (modified: 20 Dec 2024, 12:23)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Jfts, Commitment Readers

Revisions (/revisions?id=L0P2QvE6VK)

Paper Summary:
This is a review paper which reviews the related work for the causal inference with LLMs. It is separated into two main perspectives: one is the causal for LLMs and another is LLMs for Causal. They have a more details separate especially on the reasoning capacity related works. However, this work as review paper is not really suitable. There is no specific methods applied when collecting the related work. Lack of insights of how the current work and how other people can get some insighnts from the current related work.

Summary Of Strengths:
A good start and grouping for all the related paper.

Summary Of Weaknesses:
- As a review paper, lack of review methdology
- The explanation of the paper just repeats other people's work with limited insights of the review

Comments Suggestions And Typos:
- Need to have methodology described for the review/survey paper on how the papers are selected
- more insights on the current study

Confidence: 3 = Pretty sure, but there's a chance I missed something. Although I have a good feel for this area in general, I did not carefully check the paper's details, e.g., the math or experimental design.
Soundness: 1 = Major Issues: This study is not yet sufficiently thorough to warrant publication or is not relevant to ACL.
Overall Assessment: 1 = Major Revisions Needed: This paper has significant flaws, and needs substantial work before it would be of interest to the community.
Best Paper: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Official Comment by Authors

Official Comment

by Authors (Tianrui Guan (/profile?id=~Tianrui_Guan1), Fuxiao Liu (/profile?id=~Fuxiao_Liu1), Haoliang Wang (/profile?id=~Haoliang_Wang1), Paiheng Xu (/profile?id=~Paiheng_Xu1), +9 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission450/Authors))

25 Nov 2024, 07:55 (modified: 02 Jan 2025, 08:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer Jfts, Commitment Readers

Revisions (/revisions?id=iRHCyoYx46)

Comment:
Thank you for the valuable comments. Below, we address your questions in detail.

> **W1**: As a review paper, lack of review methodology

The intersection of causal inference and LLMs represents a rapidly evolving research frontier where significant developments appear across NLP/ML venues and preprints. Given this field's dynamic nature, our selection methodology prioritizes comprehensive coverage of key developments over exhaustive enumeration. For preprints, the authors of this paper manually reviewed them to assess their quality and relevance to the topic.

We focused on identifying key areas within this interdisciplinary domain and curating representative works highlighting patterns, trends, and open challenges in each area. This approach is consistent with other surveys in related fields (e.g., Feder et al., 2022), which similarly did not employ exhaustive paper selection strategies such as searching all papers with specific keywords across several venues.

We appreciate your suggestion and will clarify our paper selection methodology in the revised version to address this concern more explicitly.

> **W2**: The explanation of the paper just repeats other people's work with limited insights of the review

Our survey makes several distinctive analytical contributions beyond paper enumeration: 1. We systematically analyze the reciprocal relationship between causality and LLMs, revealing how each field can advance the other; 2. For each subfield we identify, we summarize the common trends and findings. We believe that these insights provide a deeper understanding of the field and highlight open challenges and future research directions.

Add: **Author-Editor Confidential Comment**

Official Review of Submission450 by Reviewer QkCs

Official Review by Reviewer QkCs 21 Nov 2024, 04:37 (modified: 20 Dec 2024, 12:23)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer QkCs, Commitment Readers

Revisions (/revisions?id=JHNeshqduW)

Paper Summary:
This survey explores the intersection of large language models (LLMs) and causal inference. It focuses on two main aspects: 1) Causal Inference to Improve LLMs, and 2) LLMs to Enhance Causal Inference.

Summary Of Strengths:
Combining causal inference and language modeling is a novel and impactful direction. The survey integrates a wide array of studies, offering a thorough overview of this. It captures the latest trends and challenges in integrating causality with LLMs.

Summary Of Weaknesses:

1. The structure of this paper primarily follows the framework proposed by Amir Feder et al.[1], supplementing it with more recent advancements and research on large language models. However, many relevant works are overlooked. For example, it would be beneficial to include a discussion of related works to highlight the connections and differences between this paper and other existing surveys/benchmarking papers [1, 2, 3, 4, 5], particularly the work by Amir Feder et al [1].
2. The review heavily relies on secondary studies and lacks its own experimental benchmarks or validations, which could strengthen the claims.
3. In the area of causal relationship discovery, a highly relevant field like ECI (Event Causality Identification) is barely mentioned. The authors should at least reference some representative works, such as [6, 7, 8].
4. The introduction of LVLMs is indeed an innovative aspect of the paper, but the coverage is too limited. Many related works are not analyzed, such as [9,10,11]. The reviewer suggests that the authors expand this discussion to include

these works, which would significantly enhance the paper's depth and impact.

[1] Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. Amir Feder et al. TACL 2022.

[2] Understanding causality with large language models: Feasibility and opportunities. Zhang et al. 2023

[3] Causal reasoning and large language models: Opening a new frontier for causality. Kıcıman et al. 2023

[4] Causal parrots: Large language models may talk causality but are not causal. Zecevi˘c et al. 2023.

[5] Causal evaluation of language models. Chen et al. 2024.

[6] Modeling Document-level Causal Structures for Event Causal Relation Identification. Gao et al. NAACL 2019.

[7] ERGO: Event Relational Graph Transformer for Document-level Event Causality Identification. Chen et al. COLING 2022.

[8] Knowledge-Enriched Event Causality Identification via Latent Structure Induction Networks. Cao et al. ACL 2021.

[9] Counterfactual VQA: A Cause-Effect Look at Language Bias. Niu et al. CVPR 2021.

[10] Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective. Chen et al. EMNLP 2024

[11] Language prior is not the only shortcut: A benchmark for shortcut learning in VQA. EMNLP 2022.

Comments Suggestions And Typos:
The overall structure is good. However, the writing lacks clarity in some parts. For instance, under the topic of "reasoning capacity," the distinction between "model understanding" and "commonsense reasoning" is unclear. This issue recurs throughout the paper, where many concepts are not explicitly defined, making it difficult for readers without prior knowledge to grasp their meanings.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3 = Acceptable: This study provides sufficient support for its major claims/arguments. Some minor points may need extra support or details.
Overall Assessment: 2.5
Best Paper: No
Ethical Concerns:
There are no concerns with this submission

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Official Comment by Authors

Official Comment

by Authors (Tianrui Guan (/profile?id=~Tianrui_Guan1), Fuxiao Liu (/profile?id=~Fuxiao_Liu1), Haoliang Wang (/profile?id=~Haoliang_Wang1), Paiheng Xu (/profile?id=~Paiheng_Xu1), +9 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission450/Authors))

25 Nov 2024, 08:01 (modified: 02 Jan 2025, 08:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer QkCs, Commitment Readers

Revisions (/revisions?id=Y3X1U0B60e)

Comment:

Thank you for the valuable comments. Below, we address your questions in detail.

> **W1**: The structure of this paper primarily follows the framework proposed by Amir Feder et al.[1], supplementing it with more recent advancements and research on large language models. However, many relevant works are overlooked. For example, it would be beneficial to include a discussion of related works to highlight the connections and differences between this paper and other existing surveys/benchmarking papers [1, 2, 3, 4, 5], particularly the work by Amir Feder et al [1].

Thank you for the suggestion. We would like to clarify that we have discussed works [1, 3, 4] in the submitted version. We agree it is beneficial to highlight the connections and differences.

- We follow [1]'s categorization of how causal framework can improve LLM applications. However, it is worth noting that while [1] focuses on embedding-based methods, our survey reflects the transformative impact of LLMs with strong generative capabilities. We supplement the categorization with additional discussions on LLM's reasoning capacity and multi-modality, capturing research trends driven by the enhanced capacities of LLMs.
- [3,4] benchmarked LLMs on various causal reasoning tasks (Line 244-255). Their findings align with those of other studies reviewed in the same subsection (Section 3.1).
- Thanks for pointing out [2,5]. While [2] is a short commentary on LLMs' ability to answer causal questions, [5] provides a comprehensive benchmark of the LLMs' causal reasoning capacity. We will make sure to include their findings in Section 3.1.

> **W2**: The review heavily relies on secondary studies and lacks its own experimental benchmarks or validations, which could strengthen the claims.

The primary objective of this survey is to provide a comprehensive overview of key trends, challenges, and opportunities at the intersection of causal inference and LLMs, serving as a guide for future research in this rapidly evolving field.

As a survey, this paper is not intended to present new benchmark results. Covering such a wide range of areas while providing original benchmarks would be extremely challenging, particularly within the constraints of a conference paper. Instead, we include results from multiple benchmark studies across the areas we discuss, such as causal reasoning ([3, 4, 5]) and new benchmark datasets (covered in Section 3.6).

We hope this clarifies the scope of our work, and we appreciate your understanding of the limitations inherent to a survey paper.

> **W3**: In the area of causal relationship discovery, a highly relevant field like ECI (Event Causality Identification) is barely mentioned. The authors should at least reference some representative works, such as [6, 7, 8].

We have included papers that study ECI, such as Kıcıman et al. (2023) and Gao et al. (2023), along with a survey on causal relationship extraction from text (Yang et al., 2022). As noted in Lines 157–159 of our paper, the focus of this survey is on works leveraging the strong generative capacity of LLMs. While [6, 7, 8] rely on embedding-based methods, which differ from the primary scope of this survey, we agree that including them would provide additional background and context to this area. We will incorporate these works into the revised version.

> **W4**: The introduction of LVLMs is indeed an innovative aspect of the paper, but the coverage is too limited. Many related works are not analyzed, such as [9,10,11]. The reviewer suggests that the authors expand this discussion to include these works, which would significantly enhance the paper's depth and impact.

Thanks for the suggestions. We will incorporate [9, 10, 11] in the revised version.

Add: **Author-Editor Confidential Comment**

Official Review of Submission450 by Reviewer gbyU

Official Review by Reviewer gbyU 18 Nov 2024, 20:02 (modified: 20 Dec 2024, 12:23)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer gbyU, Commitment Readers

Revisions (/revisions?id=pPso5hv3Z7)

Paper Summary:
The paper surveys the intersection of LLMs and causal inference methods. It examines how causal inference can enhance LLMs in areas such as reasoning, fairness, safety, explainability, and multimodal processing. Conversely, it discusses how LLMs can contribute to causal inference tasks, including causal relationship discovery and causal effect estimation.

Summary Of Strengths:
The links between LLMs and causality is complicate but significant to address challenges faced by either side. The survey is timely to explore these links.

Summary Of Weaknesses:
1. The LLM-based causality part appears to be covered by the existing survey the authors are aware [1].
2. Some work like [2] can be categorised as LLMs for causality, but it is put as causality for LLMs in this survey. I am wondering what the criteria for categorising are.

Comments Suggestions And Typos:
1. Line 223, equation 6 and 7, should be equation 2 and 3.
2. The selection criteria of papers and categorisation criteria should be described.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3.5
Overall Assessment: 4 = This paper represents solid work, and is of significant interest for the (broad or narrow) sub-communities that might build on it.
Best Paper: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 1 = No usable datasets submitted.
Software: 1 = No usable software released.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: After the review process started
Knowledge Of Paper Source: Preprint on arxiv
Impact Of Knowledge Of Paper: Not at all
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Official Comment by Authors

Official Comment
by Authors (Tianrui Guan (/profile?id=~Tianrui_Guan1), Fuxiao Liu (/profile?id=~Fuxiao_Liu1), Haoliang Wang (/profile?id=~Haoliang_Wang1), Paiheng Xu (/profile?id=~Paiheng_Xu1), +9 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission450/Authors))

25 Nov 2024, 08:04 (modified: 02 Jan 2025, 08:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer gbyU, Commitment Readers

Revisions (/revisions?id=kj4ZL8TyFc)

Comment:

Thank you for the valuable comments. Below, we address your questions in detail.

> **W1**: The LLM-based causality part appears to be covered by the existing survey the authors are aware [1].

Since the references in the manuscript are not numbered, we are unsure which specific survey [1] refers to. For the efficiency of the discussion phase, we will include comparisons of the two possible surveys mentioned in our manuscript.

- If the LLM-based causality [1] survey refers to Wan et al. 2024 (Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions), we acknowledge that their survey covers a subarea in Section 4.2 with a specific focus on causal discovery methods from observational data. Section 4.2 also includes papers that discover causal relationships from text or documents.
- If [1] refers to Feder et al. 2022 (Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond), we follow their categorization of how causal framework can improve LLM applications. However, it is worth noting that while Feder et al. 2022 focus on embedding-based methods, our survey reflects the transformative impact of LLMs with strong generative capabilities. We supplement the categorization with additional discussions on LLM's reasoning capacity and multi-modality, capturing research trends driven by the enhanced capacities of LLMs.

> **W2**: Some work like [2] can be categorised as LLMs for causality, but it is put as causality for LLMs in this survey. I am wondering what the criteria for categorising are.

The criteria for evaluating whether a work contributes to a causal task or an LLM task focus on the nature of the causal task itself. These tasks include fundamental causal discussions (such as causal inference assumptions, statistical causal methods, and causal applications). However, due to the limited number of existing works, discussions around assumptions and purely statistical methods are relatively sparse. We have acknowledged in the limitations section that there are other possible ways to organize this paper. We are happy to discuss this in more detail or make any necessary adjustments in the revised version if you could clarify which paper [2] refers to. We will also include more discussion on the categorization criterion.

> **C1**: Line 223, equation 6 and 7, should be equation 2 and 3.

We will fix the typo. Thanks!

> **C2**: Selection criteria of papers

The intersection of causal inference and LLMs represents a rapidly evolving research frontier where significant developments appear across NLP/ML venues and preprints. Given this field's dynamic nature, our selection methodology prioritizes comprehensive coverage of key developments over exhaustive enumeration. For preprints, the authors of this paper manually reviewed them to assess their quality and relevance to the topic.

We focused on identifying key areas within this interdisciplinary domain and curating representative works highlighting patterns, trends, and open challenges in each area. This approach is consistent with other surveys in related fields (e.g., Feder et al., 2022), which similarly did not employ exhaustive paper selection strategies such as searching all papers with specific keywords across several venues.

We will clarify our paper selection methodology in the revised version to address this concern more explicitly.

Add: **Author-Editor Confidential Comment**

Official Review of Submission450 by Reviewer cYys

Official Review by Reviewer cYys 12 Nov 2024, 02:31 (modified: 20 Dec 2024, 12:23)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cYys, Commitment Readers

Revisions (/revisions?id=alaMkQVu2u)

Paper Summary:
This paper surveys causal inference as a critical tool for advancing the capabilities and responsible deployment of LLMs. By integrating causal inference, the authors seek previous research to address limitations in LLMs related to reasoning, fairness, safety, and interpretability. They explore various existing methods, including causal graphical models and counterfactual reasoning, that can potentially enhance LLMs' decision-making processes and improve robustness against biases and unintended behaviors. Additionally, the study investigates multi-modal applications, showing how causal reasoning can enrich LLMs' understanding across text, image, and video data. The paper also discusses future directions for combining LLMs with causal inference frameworks to overcome data limitations, address unobserved confounding factors, and further expand LLMs' application scope.

Summary Of Strengths:

1. The paper is well-structured with categories thoughtfully designed to survey causal inference in LLMs, clearly articulating how the interplay is addressed.
2. As an organized and accessible work, it allows even readers unfamiliar with the subject to easily understand the topics covered

Summary Of Weaknesses:

1. In section 3.1.3 on counterfactual reasoning, the research from lines 361-364 would benefit from clearer explanations of how LLMs can address issues through a causal lens.
2. For section 3.6 on evaluation and benchmark, a more substantial focus on the explanation and Table 1 would enhance the camera-ready version.
3. In section 4.1 on Treatment Effect Estimation, content from reference [1] should be incorporated.
4. The length of 8 pages makes it challenging to cover all the necessary content for a comprehensive survey. Although the paper presents previous studies based on solid materials, it falls short of providing clear insights or discussions that contribute to a deeper understanding.

Comments Suggestions And Typos:
mentioned in weakness.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3.5
Overall Assessment: 3 = Good: This paper makes a reasonable contribution, and might be of interest for some (broad or narrow) sub-communities, possibly with minor revisions.
Best Paper: No
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 5 = They could easily reproduce the results.
Datasets: 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
Software: 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)
Knowledge Of Or Educated Guess At Author Identity: Yes
Knowledge Of Paper: After the review process started
Knowledge Of Paper Source: Preprint on arxiv
Knowledge Of Paper Source Other: https://arxiv.org/abs/2403.09606 (https://arxiv.org/abs/2403.09606)
Impact Of Knowledge Of Paper: Not at all

Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Add: **Author-Editor Confidential Comment**

Official Comment by Authors

Official Comment

by Authors (Tianrui Guan (/profile?id=~Tianrui_Guan1), Fuxiao Liu (/profile?id=~Fuxiao_Liu1), Haoliang Wang (/profile?id=~Haoliang_Wang1), Paiheng Xu (/profile?id=~Paiheng_Xu1), +9 more (/group/info?id=aclweb.org/ACL/ARR/2024/October/Submission450/Authors))

25 Nov 2024, 08:13 (modified: 02 Jan 2025, 08:25)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cYys, Commitment Readers

Revisions (/revisions?id=9cvpFYtnay)

Comment:

Thank you for the valuable comments. Below, we address your questions in detail.

> **W1**: In section 3.1.3 on counterfactual reasoning, the research from lines 361-364 would benefit from clearer explanations of how LLMs can address issues through a causal lens.

> **W2**: For section 3.6 on evaluation and benchmark, a more substantial focus on the explanation and Table 1 would enhance the camera-ready version.

We will include more discussion on the two works mentioned from lines 361-364 and benchmarks in Section 3.6 in the future version!

> **W3**: In section 4.1 on Treatment Effect Estimation, content from reference [1] should be incorporated.

If [1] refers to Feder et al. 2022 (Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond), we will mention their categorization on estimating causal effects with text including settings where text is used as an outcome, treatment, or to address confounding.

However, we would appreciate further clarification on what [1] refers to or which specific content from [1] you think should be incorporated. This will help us address your point accurately.

> **W4**: The length of 8 pages makes it challenging to cover all the necessary content for a comprehensive survey. Although the paper presents previous studies based on solid materials, it falls short of providing clear insights or discussions that contribute to a deeper understanding.

Our survey provides insights mainly in the following two directions: 1. We systematically analyze the reciprocal relationship between causality and LLMs, revealing how each field can advance the other; 2. For each subfield we identify, we summarize the common trends and findings. We believe that these insights provide a deep understanding of the field and highlight open challenges and future research directions.

To address your concern, we plan to expand the discussion on common trends and findings both within and across subareas, if given the opportunity to include an additional page in the camera-ready version.

Add: **Author-Editor Confidential Comment**