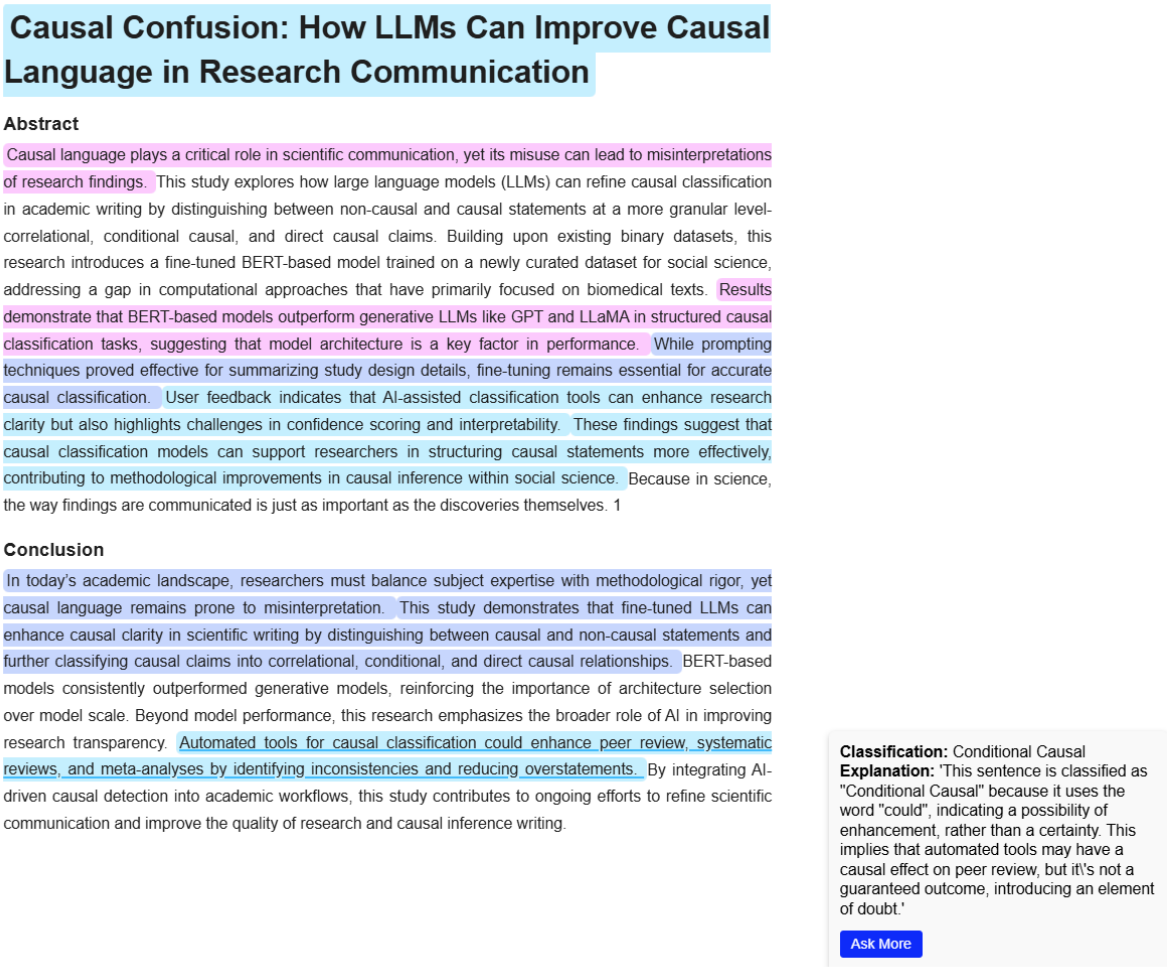
Causal Confusion

How LLMs Can Improve Causal Language in Research Communication

Tessa van Abkoude



Causal Confusion: How LLMs Can Improve Causal Language in Research Communication

Abstract

Causal language plays a critical role in scientific communication, yet its misuse can lead to misinterpretations of research findings. This study explores how large language models (LLMs) can refine causal classification in academic writing by distinguishing between non-causal and causal statements at a more granular level-correlational, conditional causal, and direct causal claims. Building upon existing binary datasets, this research introduces a fine-tuned BERT-based model trained on a newly curated dataset for social science, addressing a gap in computational approaches that have primarily focused on biomedical texts. Results demonstrate that BERT-based models outperform generative LLMs like GPT and LLaMA in structured causal classification tasks, suggesting that model architecture is a key factor in performance. While prompting techniques proved effective for summarizing study design details, fine-tuning remains essential for accurate causal classification. User feedback indicates that AI-assisted classification tools can enhance research clarity but also highlights challenges in confidence scoring and interpretability. These findings suggest that causal classification models can support researchers in structuring causal statements more effectively, contributing to methodological improvements in causal inference within social science. Because in science, the way findings are communicated is just as important as the discoveries themselves. 1

Conclusion

In today's academic landscape, researchers must balance subject expertise with methodological rigor, yet causal language remains prone to misinterpretation. This study demonstrates that fine-tuned LLMs can enhance causal clarity in scientific writing by distinguishing between causal and non-causal statements and further classifying causal claims into correlational, conditional, and direct causal relationships. BERT-based models consistently outperformed generative models, reinforcing the importance of architecture selection over model scale. Beyond model performance, this research emphasizes the broader role of AI in improving research transparency. Automated tools for causal classification could enhance peer review, systematic reviews, and meta-analyses by identifying inconsistencies and reducing overstatements. By integrating AI-driven causal detection into academic workflows, this study contributes to ongoing efforts to refine scientific communication and improve the quality of research and causal inference writing.

Classification: Conditional Causal
Explanation: 'This sentence is classified as "Conditional Causal" because it uses the word "could", indicating a possibility of enhancement, rather than a certainty. This implies that automated tools may have a causal effect on peer review, but it\'s not a guaranteed outcome, introducing an element of doubt.'

Ask More

Eindhoven University of Technology

Acknowledgments

I would like to express my gratitude to everyone who contributed to the completion of this Master's thesis in Human Technology Interaction at Eindhoven University of Technology.

First and foremost, I want to express my appreciation to my supervisor, Daniël Lakens, for his shared enthusiasm and continuous guidance throughout this thesis. Thank you for introducing me to meta-science and guiding me in identifying and improving research practices. You were always available when I needed your support, and I truly appreciate all the effort you put into this thesis. Fortunately, I could give back a little by teaching you a thing or two about LLMs along the way. Even when I became too focused on coding, you always helped me refocus on the core of the research and keep my activities meaningful.

Second, I am grateful to the technical experts Zeno van Cauter, Zhang Chaobo, Casper van Lissa, and Rasoul Norouzi for sharing their expertise. I would like to thank the participants of my interviews, whose valuable feedback helped shape the final interface. I would also like to thank Krist Vaesen for the opportunity to present my work as a seminar, which allowed me to gain new perspectives and helped shape this thesis.

Finally, I am deeply grateful to my mom and dad, family, friends, and partner for their support throughout my thesis, whether through encouragement, much needed distractions or simply being there. I especially want to thank my housemate, Midas Zegers, for being there from the introduction week all the way to the end of my master's. Your friendship and support made studying so much more enjoyable, with endless dinner conversations, late-night wine talks, and long Atlas study sessions.

Tessa van Abkoude
Eindhoven, February 2025

Abstract

Causal language plays a critical role in scientific communication, as it shapes public understanding, informs policy, and impacts healthcare decisions. When causal statements are ambiguous or misleading, they can lead to confusion and misinterpretation of research findings. This study explores how large language models (LLMs) can improve causal language in academic writing. It employs a two-step approach: first, distinguishing non-causal from causal statements, and second, classifying causal sentences as correlational, conditional causal, or direct causal. The models were fine-tuned on a blended dataset of general-purpose (news, web) and scientific (social science, biomedical) human-labeled sentences. The BERT-based classifier achieved a macro F1-score of 0.94 for detecting causal versus non-causal sentences, while SciBERT attained 0.83 in distinguishing correlational, conditional causal, and direct causal statements. To explore how these classifiers can be applied in practice, a tool was developed to analyze scientific papers and texts, offering personalized warnings and highlighting potential inconsistencies in causal reasoning. By providing researchers with a (visual) overview of causal strength and alignment with study design, the tool supports clearer, more precise communication of research findings. This study demonstrates how LLMs can enhance the clarity and precision of causal language in academic writing, offering a scalable approach to improving scientific communication.

Contents

1

Introduction

In today's academic world, researchers are expected to balance in-depth knowledge of their fields with advanced statistical skills, ensuring their work is both rigorous and reproducible. Given the rapid growth of research and the increasing complexity of methodologies, meeting these demands has become more challenging. Cognitive constraints limit how much one can learn in a lifetime, making it harder to manage the demands of both advanced statistical methods and specialized knowledge, especially in combination with the pressure to publish. A frequent challenge is the misuse of causal language, where researchers may incorrectly use terms like "causes" or "leads to," implying causality when only a correlational relationship has been established (Yu et al., 2019). As a result, errors are common in causal inference, where misinterpretations arise not only in statistical analyses but also in the use of causal language to describe findings, as well as in study designs that do not appropriately align with causal claims (Haber et al., 2022).

Causal inference is central to understanding relationships across scientific fields, from medicine to the social sciences. It allows researchers to explore cause-and-effect relationships that inform theory and practice. However, mastering causal inferences is complex, requiring a balance of subject-matter expertise and advanced statistical knowledge (Yu et al., 2019). Misunderstandings of concepts such as confounders, biases introduced by methodological choices, and the misuse of causal language in research outputs all contribute to a broader issue of scientific miscommunication (Haber et al., 2022). These issues are particularly consequential in fields like public health and social sciences, where findings often inform policy decisions. For instance, Cofield et al. documented widespread misuse of causal language in observational studies of obesity, leading to overstated claims that influenced both policy and public understanding.

1. Misuse of Causal Language in Research and Media

The misuse of causal language not only leads to scientific misinterpretations but also propagates incorrect conclusions both within the scientific community and the general public. Sumner et al. found that 33% of press releases from correlational research exaggerated causal claims, presenting associations as causal. In fields such as the social sciences and medicine, accurate causal reasoning is essential for both theoretical development and informed decision-making (Thapa et al., 2020). However, hedging language—terms like "may contribute to" or "could be linked to"—often creates ambiguity about the strength and direction of causal claims (Frankenhuis et al., 2023; Hedström & Ylikoski, 2010). Journalists, aiming to capture attention, may distort original findings by amplifying causal language to create a more appealing narrative (Sumner et al., 2014). Even in academic papers, abstracts and conclusions sometimes overstate causality to simplify the study's message (Thapa et al., 2020). Results sections may also present associations as if they were causal effects without sufficient methodological grounding, leading to misinterpretation (Cofield et al., 2010; Haber et al., 2022). The use of clickbait headlines (Carcioppolo et al., 2021) further distorts public understanding, especially when headlines imply causality where only correlation exists in the original research. Within scientific communities, researchers may also overstate findings to make their work more accessible to a broader audience,

inadvertently leading to unwarranted causal interpretations (Adams et al., 2017).

Although corrections can reduce the overstatement of causal claims in media reports, the original misinformation persists in people's minds, continuing to influence their reasoning. Irving et al. found that even after corrections, participants still referenced the initial claim, highlighting the challenge of fully correcting the impact of misinformation once it has been introduced through the media.

2. Interpretation of Causal Language

The public's understanding of causal language, particularly in media contexts, is often shaped by linguistic subtleties and contextual factors. Adams et al. conducted a series of experiments to examine how people interpret causal and correlational expressions in news headlines. The study recruited 71 participants from social media (ages 17–63) and 160 psychology undergraduates (ages 17–30) to assess how different backgrounds influence causal interpretation. Despite variations in science education levels, no significant differences were found in how participants understood causal versus correlational language. Participants consistently categorized terms like "makes" and "increases" as implying direct causation, while more cautious terms such as "might cause" or "linked" were interpreted less definitively. However, participants did not consistently distinguish between moderate causal and correlational terms, indicating a gap in their understanding of more nuanced distinctions (Adams et al., 2017).

Another common misinterpretation arises not from confusion between causal and correlational language but from the way statistical findings are framed. Michal and Shah observed that individuals without formal training in scientific or statistical reasoning, known as non-experts, tend to overestimate the importance of findings described with vague terms such as "more effective" or "better outcomes." This "practical significance bias" suggests that ambiguous language can lead readers to infer substantial real-world impact even when the measured effect sizes are minimal. Such biases highlight the necessity for scientific reporting to include clear contextual explanations, thereby reducing the risk of misinterpretation.

Moving from general misinterpretations to more academic contexts, Seifert and Hammond and List examined causal reasoning errors, where students infer causation from correlational evidence. Seifert studied psychology undergraduates in research methods courses, while List focused on undergraduates in the social and natural sciences. Both groups struggled to differentiate between correlational and causal evidence, often failing to recognize the methodological limitations of correlational data. Students in both studies frequently assigned similar quality ratings to causal claims supported by correlational evidence as those supported by experimental data, reflecting a misunderstanding of key methodological distinctions.

These errors were attributed to heuristic thinking, where students intuitively connected variables without critically analyzing evidence. For example, Seifert's students misinterpreted a study linking maternal control to childhood obesity, overlooking third-variable effects (e.g., socioeconomic status) or reverse causation. Similarly, List found students misjudged a correlation between optimism and longevity, concluding optimism directly increases lifespan while ignoring alternative explanations like healthier lifestyles. Errors were particularly pronounced when evidence lacked qualifiers about its limitations.

To address these biases, Seifert and Hammond introduced an intervention using example-based learning and guided exercises. Students analyzed cases of causal theory error and generated alternative explanations, such as reverse causality and third-variable effects. The intervention improved students' ability to critique causal claims, but nearly half continued to rate correlational studies as strong evidence for causation, showing the persistence of intuitive reasoning errors. While Seifert and Hammond's example-based learning showed promise in addressing biases, the persistence of these errors indicate the need for scalable solutions. With the rise of AI and computational methods, there is an opportunity to develop tools that can automatically identify and clarify causal claims, helping both researchers and the public navigate causal language and ensuring more accurate communication of scientific findings. While AI tools can help clarify causal claims and improve communication, their application must align with existing disciplinary guidelines on when causal language is justified.

3. Guidelines on When to Infer Causality

The question of when researchers can appropriately draw causal conclusions remains an ongoing debate across disciplines. In medicine, randomized controlled trials (RCTs) are widely considered the gold standard for establishing causation because randomization minimizes selection bias and unmeasured confounding. Consistent with this view, the *Journal of the American Medical Association* (JAMA) advises that causal language—such as "effect" and "efficacy"—be reserved for RCTs. Other study designs are instructed to describe findings in terms of association or correlation (Dahabreh & Bibbins-Domingo, 2024). This guidance aligns with the AMA Manual of Style, which states that: *"causal language (including use of terms such as effect and efficacy) should be used only for randomized clinical trials. For all other study designs … methods and results should be described in terms of association or correlation and should avoid cause-and-effect wording"* (Christiansen et al., 2020).

Not all fields, however, share the view that RCTs are the only way to establish causal inference. Hernán and Robins argue that observational studies can provide valid causal conclusions if key assumptions, such as exchangeability and the absence of unmeasured confounding, are satisfied. Their counterfactual framework demonstrates that rigorous measurement and careful adjustment for confounders can allow observational studies to approximate the causal insights typically derived from randomization. Under the right methodological conditions, observational designs can yield meaningful causal inferences (Hernán & Robins, 2020).

In psychology, researchers have similarly challenged the idea that causal claims require randomized experimental designs. Grosz et al. describe how the traditional aversion to causal language in nonexperimental psychology has been mitigated by advances in longitudinal studies and quasi-experimental designs. These methodologies address issues like reverse causation and confounding, enabling researchers to draw causal inferences under certain conditions. For example, careful design and robust statistical modeling in nonexperimental studies can yield valuable causal insights, demonstrating the potential for methods beyond RCTs to inform psychological science.

There is no consensus on when causal language is justified. In medicine, guidelines generally restrict causal claims to RCTs, while fields like epidemiology and the social sciences acknowledge that well-designed observational studies can sometimes support causal inferences. These disciplinary differences underscore the importance of aligning causal claims with field-specific norms, the robustness of the study design, and the assumptions underpinning causal inference. Observational studies that approximate experimental conditions through rigorous sampling, advanced statistical techniques, and thorough control of confounders can contribute valuable insights. Researchers must remain cautious in their use of causal language, clearly communicate the limitations of their methods, and base their claims on the strongest evidence available.

While different disciplines have varying standards for when causal language is justified, inconsistencies and misapplications of these standards remain common. Given the importance of accurately communicating causal claims and the frequent misuse of causal language, researchers have increasingly explored whether AI tools can help systematically identify and classify causal statements in scientific papers. This has led to growing interest in computational methods that assess the strength and appropriateness of causal claims.

4. Computational Approaches to Classifying Causal Claims

Identifying and categorizing the strength of causal claims in scientific communication has been a key focus of computational studies aiming to improve the rigor of causal language use. Sumner et al. laid the groundwork for this field by developing a seven-level taxonomy of claim certainty through manual content analysis. These levels ranged from "No statement" to "Unconditionally causal," with intermediate categories like "Correlational," "Ambiguous," and "Conditional causal."

Adams et al. examined how readers interpret causal language. They found that participants struggled to reliably distinguish between "Conditional causal" and "Correlational" statements. Instead, readers grouped causal claims into three broader categories:

- Direct cause statements (e.g., "drinking wine increases cancer risk")
- Can cause statements (e.g., "drinking wine can increase cancer risk")

- Moderate cause statements (e.g., “drinking wine is associated with cancer risk”)

The “Can cause” category included statements where causality was implied but not definitively asserted. This group consisted of claims using terms like “can lead to” or “may contribute to”, which suggest a potential causal relationship without making a definitive claim. While this distinction aligns with common hedging language used in scientific writing, the study found that participants did not reliably separate “Can cause” statements from “Moderate cause” (correlational) statements.

These findings reveal limitations in how humans perceive distinctions in causal language. Subsequent computational studies have adopted a more streamlined approach than the seven-level taxonomy proposed by Sumner and colleagues. Researchers refined the taxonomy to four categories that balanced interpretability for readers and computational efficiency for machine learning models (Tan et al., 2023; Wright & Augenstein, 2021; Yu et al., 2019), and used the categories: Direct causal, Conditional causal, Correlational, and No relationship.

This refined taxonomy has since been widely adopted in machine learning applications. It allows for effective automated classification of causal claims while retaining sufficient granularity to support meaningful distinctions. Table 1.1 outlines the taxonomy of causal strength alongside their descriptions and common language cues, while Table 1.2 provides annotated examples of sentences.

Label	Description and Language Cue
Correlational	The statement describes the association between variables, but causation cannot be explicitly stated. *Example cues:* association, associated with, predictor, at high risk of
Conditional Causal	The statement shows that one variable directly changes the other but includes an element of doubt. *Example cues:* increase, decrease, lead to, effect on, contribute to, result in. *Doubt cues:* may, might, appear to, probably
Direct Causal	The statement asserts that the independent variable directly alters the dependent variable. *Example cues:* increase, decrease, lead to, effective in, contribute to, reduce
No Relationship	No correlation or causation relationship is mentioned in the statement. *Example cue:* -

Table 1.1: Taxonomy of Causal Strength

By automating the identification of causal language, researchers and media practitioners can reduce ambiguities and prevent misrepresentation of results. This approach underscores the growing role of computational tools in promoting clarity and accuracy in scientific communication (Wright & Augenstein, 2021; Yu et al., 2019). These tools can function similarly to a spell checker for causal language, automatically flagging inconsistencies in claim strength and ensuring the appropriate use of causal terms. A tool could also provide interactive feedback, guiding users to correct or adjust imprecise language, making it easier for researchers and communicators to produce more accurate and transparent causal claims.

Building on this foundational work, we will use these four categories to classify causal claims made in sentences in scientific articles from the social science literature. Unlike existing methods, which often focus on isolated sections of a paper, our approach examines causal statements throughout the entire document.

Description	Example Sentence
Correlational	1. Lifelong brain stimulating habits linked to lower Alzheimer's protein levels
	2. Low ALT was associated with higher mortality risk
Conditional	3. Breastfeeding may prevent asthma.
Causal	4. Yet when group identification was high, both leader types appeared to be equally efficient
Direct Causal	5. We find that conflict increases cooperation within groups, while decreasing cooperation between groups.
	6. Stroke is a leading cause of death and disability worldwide.
No Relationship	7. The offers have been placed into our catalogue.
	8. More research is needed to better understand this association.

Table 1.2: Example Sentences Per Category

5. From Computational Tools to Large Language Models (LLMs)

Given the complexity and importance of causal inference, researchers require effective tools to help identify causal language misuse and improve the clarity of their findings. While traditional tools such as Statcheck (Nuijten et al., 2017) and JARS (Kazak, 2018) have been developed to assist with statistical analyses by automatically detecting certain common errors in statistical reporting or communicate reporting guidelines. Statcheck primarily focuses on identifying errors, such as misreported p-values, without addressing the nuances of language, particularly causal language. These tools also lack the ability to provide interactive, in-text, feedback that could guide researchers towards improving the clarity and accuracy of their causal claims. A LLM could flag vague terms like "may cause" or "could lead to" and recommend best practices such as use terms like "was found to be statistically significant" or "add context." This would enable researchers to avoid misleading causal claims, ensuring that causal language used in scientific publications is both accurate and transparent (Kim et al., 2023).

Given the complexity and importance of causal inference, researchers need tools that not only detect causal language misuse but also provide actionable guidance to improve clarity. Existing tools such as Statcheck (Nuijten et al., 2017) and JARS (Kazak, 2018) focus primarily on statistical reporting errors but do not address causal language writing. Statcheck, for example, identifies misreported p-values but lacks the ability to assess whether causal statements align with study design or whether hedging terms obscure causal claims. These tools also do not provide interactive, in-text feedback that helps researchers refine causal phrasing throughout their papers. Large Language Models (LLMs) could offer a more dynamic approach by not only identifying vague or misleading causal statements but also offering suggestions and personalized warnings. An LLM-based system could analyze an entire research paper, visually mapping causal strength, flagging inconsistencies between causal claims in different sections, and warning against excessive hedging. By integrating these capabilities, LLMs could help researchers align their causal language with best practices, ensuring that scientific findings are communicated with greater precision and transparency (Kim et al., 2023).

Large Language Models (LLMs), such as BERT, GPT, and LLAMA, have shown great potential in addressing the challenges of causal language misuse in scientific texts. These models are capable of understanding complex language patterns and contextual meaning, making them suitable for improving the accuracy of causal inferences in research.

BERT (Bidirectional Encoder Representations from Transformers) excels at tasks where the model must classify and understand the local context within sentences. It is particularly effective when fine-tuned on domain-specific data, enabling it to detect subtle distinctions in language, such as identifying causal relationships. For tasks requiring high accuracy on specific problems, BERT can outperform other models, provided there is enough training data Yu et al. However, BERT's performance depends heavily on the availability of labeled data, and while it can be highly effective for classification tasks, it lacks flexibility when dealing with diverse, less defined tasks. It is also more static, in the sense that it does not engage in interactive dialogue or provide explanations for its outputs.

On the other hand, GPT (Generative Pretrained Transformer) and LLAMA (Large Language Model Meta AI) are more adaptable and perform well even with minimal data through few-shot or zero-shot learning. Both GPT and LLAMA are well-suited for tasks that require flexibility and scalability, offering solutions with less training data. These models are designed for text generation and can adapt to a wide range of queries. Notably, they are also conversational models, providing an interactive user experience by explaining their reasoning and engaging in dialogue. This capability makes them particularly valuable for researchers who need clarification or further elaboration on causal language detection. For instance, GPT models can generate context-specific suggestions and explain the rationale behind their corrections, which adds a layer of transparency to the process (Kim et al., 2023). Another advantage of GPT and LLAMA is that they are suitable for deployment in different environments. While GPT is typically cloud-based, requiring access to platforms like OpenAI, BERT and LLAMA can be run locally or on a researcher's own server, offering more control over sensitive data—an important consideration in scientific contexts, especially when dealing with unpublished research. This gives researchers flexibility in choosing the best balance between privacy and performance when checking for causal inference errors in their work.

6. Existing Tools and Approaches

Several computational models have been developed to detect causal language misuse in scientific texts, each offering unique strengths and limitations based on the domain and task. We compare model performance using macro F1 scores, a common metric in classification tasks that balances precision (true positives vs. false positives) and recall (true positives vs. false negatives). It is calculated by computing the F1 score for each class separately and averaging them equally, regardless of class size. This provides a single number that reflects how well a model makes correct predictions across all categories.

In the work by Yu et al., BioBERT, a domain-specific version of BERT, was fine-tuned on over 3,000 PubMed abstracts, which are health-related in nature. BioBERT achieved a macro F1 score of 0.88 in identifying causal language, excelling in identifying correlational and direct causal relationships. However, BioBERT is specifically designed for biological and health-related texts and was trained only on abstracts, limiting its applicability to other parts of a paper (e.g., discussions or results) and its ability to generalize beyond the health domain.

There is a need to identify causal statements throughout an entire paper, rather than restricting the analysis to sections like the abstract. In social scientific writing, causal claims are often expressed in subtle ways across multiple sections, with varying contexts shaping how causality is presented. For instance, a title might imply a causal relationship that is cautiously qualified in the discussion, or causal language in the results section might contradict the methodological limitations described in the methods. Conducting a full-text analysis allows for the detection of such patterns and inconsistencies, offering a more comprehensive understanding of how causality is conveyed.

Norouzi et al. explored the use of BERT in social science contexts on full papers, where the model was fine-tuned using a mix of public general-purpose datasets that were not primarily scientific literature. Their dataset included approximately 10,000 sentences labeled for non-causal or causal language. Fine-tuning models on a domain-specific dataset, such as the social science training set, increased performance. BERT showed a 15% improvement in the F1 macro average metric when classifying causal and non-causal sentences from the social science test set. This demonstrates how even a relatively small amount of domain-specific data (1,058 examples) can improve the detection of causal language in social science texts.

Kim et al. tested GPT-3 and LLAMA on 2,076 sentences from PubMed abstracts and EurekAlert! press releases, labeled for direct causal, conditional causal, correlational, and no relationship. While these models showed strong few-shot performance, GPT-3 initially struggled with conditional causal language (macro F1 of 0.164). Chain-of-Thought prompting improved GPT-3's performance to 0.631. LLAMA also showed promise in zero-shot tasks across multiple domains, but it experienced difficulties with subtle hedging terms.

Overall, BioBERT seems to perform the best for health-related texts with a macro F1 score of 0.90, whereas BERT excels when fine-tuned for texts from the social sciences (macro F1 score of 0.89).

GPT-3 and LLAMA provide flexibility and interactive feedback but require targeted tuning for specific domains.

7. Improving Causal Language in Scientific Communication

Efforts to improve causal language focus on making explanations clear, precise, and accessible for both scientific audiences and the public. Several studies, including those by (König et al., 2023).(Bott et al., 2019),(Adams et al., 2017), (Seifert & Hammond, 2022), provide valuable insights into how to present causal relationships more effectively.

König et al. emphasized that well-organized, jargon-free communication significantly improves public understanding of scientific findings. Their review outlined approaches such as using neutral language, clearly indicating uncertainty, and referencing expert sources. This is particularly crucial during crises, like the COVID-19 pandemic, where the public's trust in science directly impacts behavior.

Bott et al. including caveats ("further research is needed," "correlational, not causal") to maintain engagement while encouraging critical appraisal. Educational interventions using causal diagrams also enhance students' ability to evaluate evidence critically (Seifert & Hammond, 2022).

Strategies for Better Causal Communication

Building on these findings, researchers and communicators can adopt a comprehensive set of strategies, grouped into key categories, to improve how causal relationships are presented:

Clear and Accurate Presentation of Causality

Differentiate Between Causation and Correlation
Specify whether findings show causation or correlation. Adams et al. found that readers often confuse phrases like "causes" and "linked to." For instance, replace "A causes B" with "A is associated with B, but this does not imply causation." Adding statements like "correlation does not equal causation" reinforces this distinction and ensures clarity.

Explain Study Design and Limitations
Clearly describe the type of study and its implications. Adams et al. and Bott et al. both stress that readers should understand whether a study is experimental, observational, cross-sectional, or longitudinal. For example, "This was a randomized controlled trial, providing strong evidence for causality," or "This observational study cannot determine causation." These explanations are critical in helping readers appropriately interpret the scope of the findings.

Provide Quantitative Context
Include numerical data and statistical measures to prevent overinterpretation. Michal and Shah noted that numerical details help readers understand findings more accurately. For instance, "Green tea drinkers showed a 10% reduction in blood pressure" is more informative than vague claims like "green tea improves health."

Engaging and Transparent Communication

Use Simple and Accessible Language
Avoid technical jargon and prioritize clear terms. König et al. found that accessible language improves trust and understanding. Replace terms like "statistically significant" with plain language explanations such as "This result is unlikely to be due to chance." Simple phrasing helps reduce confusion and builds credibility.

Write Accurate and Balanced Headlines
Headlines shape first impressions. Adams et al. emphasized the importance of accuracy, advocating for headlines like "Study finds association between X and Y" instead of "X causes Y." Balanced headlines reduce the risk of misinterpretation and build trust with readers.

Include Caveats to Frame Findings
Add caveats to provide context and caution. Bott et al. found that phrases like "further research is

needed" effectively communicate limitations without disengaging readers. For instance, when discussing observational studies, include disclaimers such as "This study cannot establish causality." These disclaimers enhance transparency without making the content less engaging.

Enhancing Critical Engagement and Understanding

Provide Alternative Explanations
Encourage critical thinking by suggesting other possible causes. Seifert and Hammond, 2022 demonstrated that generating alternative explanations reduces reasoning errors. For instance, instead of stating, "Optimism leads to better health," write, "Optimism may contribute to better health outcomes, but other factors like lifestyle could also play a role." This encourages readers to evaluate evidence more critically.

Use Visual Aids to Clarify Relationships
Visual tools like flowcharts or causal diagrams help clarify relationships and distinguish correlation from causation. Seifert and Hammond found that diagrams improved students' understanding of confounding factors. For example, a flowchart illustrating potential confounders can make complex ideas easier to grasp.

The issues outlined above, from the misuse of causal language to inconsistencies in its application across different sections of a paper, point to the need for systematic, scalable solutions. Studies like Adams et al. provide valuable insights into interpreting causal language and offer tips for clearer communication, particularly in distinguishing correlation from causation. However, not everyone reads these papers or is familiar with the best practices, leading to frequent mistakes. An automatic warning or checker tool could serve as a timely reminder, helping researchers ensure more accurate communication of causal relationships and reduce the likelihood of misinterpretation.

This research aims to contribute to the literature by evaluating whether large language models (LLMs) can be used to automatically detect causal statements, classify their level of causality, and provide personalized warnings for better causal language writing. Beyond classification, this study explores whether LLMs can serve as an interactive tool that warns researchers about potential misalignment between causal claims in different sections of a paper, flags excessive hedging or overstatement, and offers actionable suggestions for clearer causal inference communication. By investigating both the feasibility and effectiveness of such a tool, this research provides insights into how automated approaches can support more precise and transparent causal reasoning in scientific writing.

2

Method

To evaluate the effectiveness of LLMs in causal classification, this study compares the performance of BERT, GPT, LLAMA, SciBERT, and RoBERTa across both binary (causal vs. non-causal) (RQ 1.1) and three-level (correlational, conditional causal, direct causal) classification tasks (RQ 1.2). The models are fine-tuned using a diverse dataset consisting of both general-purpose texts (news, web) and scientific literature, including social science and biomedical research articles. Additionally, the study examines how well these models generalize to social science texts (RQ 1.3). Since current causal classification models are primarily trained on biomedical literature, this study evaluates whether they still perform well on social science writing, even though only a small portion of the training data comes from this domain. This helps determine whether these models can be applied to different fields or require further fine-tuning for different academic fields.

Beyond model performance, this research explores the practical application of LLMs for improving causal inference in scientific writing. Specifically, it investigates how prompt engineering can enable large models, such as LLAMA 3.3 70B, to generate structured study design summaries and provide explanations for causal classifications (RQ 2.2). Since certain causal statements are only justified after specific study designs, yet they are often presented without explicitly linking them to the underlying methodology, leading to potential misuse and overstatments. This study investigates whether an LLM can accurately summarize study design details and contextualize causal statements to help researchers ensure their claims align with their research methods. Additionally, this study explores how researchers prefer to receive automated feedback on causal writing (RQ2.1). Different researchers and disciplines may have distinct needs regarding feedback presentation. Some may prefer direct, inline annotations that highlight problematic causal statements, while others may find high-level summaries or structured warnings more useful. To assess which formats best support accurate and effective causal reasoning, this study evaluates multiple ways of feedback, including interactive explanations, summaries and charts. Furthermore, the study explores researchers' preferences regarding classification model strictness, specifically whether they prioritize minimizing false positives (Type I errors) or false negatives (Type II errors) in causal language detection (RQ2.3). Overstated causal claims can have significant consequences, including misleading policymakers or misrepresenting scientific findings. As a result, some researchers may prefer a conservative approach that avoids false positives, ensuring that only well-supported causal claims are flagged. Others may prioritize minimizing false negatives, as missing a valid causal statement could lead to overlooked insights. This study examines how researchers weigh these trade-offs and whether they prefer customizable confidence thresholds that allow them to adjust strictness levels based on their specific needs.

1. Research Questions

To investigate these issues, this study addresses the following research questions:

> **Main Research Question:**
> Which large language model (BERT, GPT, LLAMA) perform best in causal classification tasks, and how can their outputs be effectively used to provide feedback for academic writing?

1.1. Sub-questions:

RQ 1.1 Which large language model (BERT, GPT, LLAMA) performs best in two-level causal classification (causal vs. non-causal)?

RQ 1.2 Which large language model (BERT, SciBERT, RoBERTa, GPT, LLAMA) performs best in three-level causal classification (correlational, conditional causal, direct causal)?

RQ 1.3 How well do these LLMs generalize to social science texts in two-level causal classification?

RQ 2.1 How do researchers prefer to receive feedback on causal writing (e.g., through conversational LLMS, in-text annotations, warnings, or high-level overviews)?

RQ 2.2 How can generative models like LLAMA 3.3 70B, through prompt engineering, support causal inference writing, such as summarizing study designs or explaining causal classifications?

RQ 2.3 When evaluating causal statements, do researchers prioritize minimizing false positives (Type I errors) or false negatives (Type II errors), and why?

Open Science Statement: We openly share the dataset and code underpinning this research at https://github.com/Tessavana/Causal-Clarity[1]. For ease of use and reproducibility, we provide convenient Kaggle links for fine-tuning and trained models within the repository. Both the dataset and code are released under the Apache License 2.0, encouraging use and modification with appropriate attribution.

2. Exploring Questionable Research Practices

At the beginning of this thesis, the initial focus was on identifying questionable research practices (QRPs) and examining the potential of automated or computational tools to mitigate them. QRPs are alarmingly common in science, with over half of researchers in the Netherlands in the social and behavioral sciences admitting to practices like selective reporting and p-hacking, and a small fraction even confessing to data fabrication (Gopalakrishna et al., 2022). These behaviors contribute to the replication crisis, where many studies fail to replicate, undermining trust in scientific findings (Schlegelmilch et al., 2015). To address these issues, the initial phase of this research explored QRPs and their potential for detection through computational tools.

At the *Perspectives on Scientific Error* conference (6th edition), held at TU Eindhoven (February 29–March 2 2024), Daniël Lakens and I hosted a hackathon where researchers began by analyzing psychological papers using text-mining tools and training their own BERT models through simplified Python notebooks. This hands-on session was followed by a brainstorming discussion where participants, primarily professors or PhD candidates, were asked what additional QRPs could be identified using computational and automated tools. They shared examples of QRPs commonly encountered in peer reviews or student work, including inadequate reporting standards, statistical missteps, and misuse of causal language. These insights were grouped into broader categories (see Figure 2.1), helping identify causal inference as a promising area for further study.

2.1. Narrowing the Focus: Causal Inference

Causal inference was selected as the focus of this thesis for several reasons. First, causal claims have significant theoretical and practical implications in fields like medicine, psychology, and the social sciences, where they influence policy and public understanding. Second, causal inference poses a distinct linguistic challenge, as subtle shifts in phrasing, such as “may contribute to” versus “results in”, can dramatically change the interpretation of findings. These nuances make it an ideal application for large language models (LLMs) which excel at analyzing language. As these models are becoming

[1] https://github.com/Tessavana/Causal-Clarity

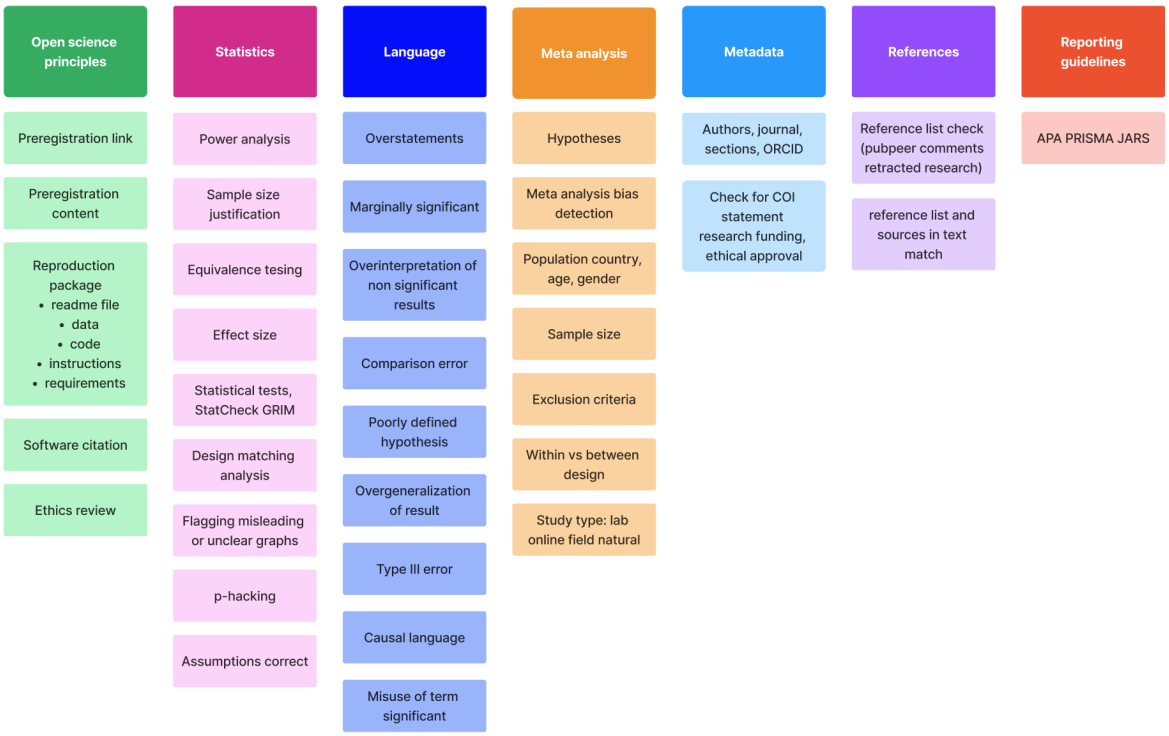


Figure 2.1: Questionable Research Practices Categories Overview

increasingly smarter, this study sets out to explore its current capabilities. An additional reason for focusing on causal inference was the availability of existing literature and datasets, which provided a valuable benchmark. This reduced the need for extensive manual labeling, a task that would have been infeasible within the timeframe of this one-semester thesis.

3. Creating a Causal Dataset

The aim of constructing this dataset is to classify causal sentences in scientific articles while leveraging the widest range of available human labeled data. To achieve this, we adopted a two-step classification approach, allowing us to utilize datasets with binary labels (causal vs. non-causal) and those with more granular annotations (correlational, conditional causal, direct causal, and non-causal). This strategy ensured that the model could be trained on a broad range of data sources and later fine-tuned to distinguish between different types of causal language.

The decision to use this approach was also informed by the performance of classifiers during the initial phase. Models that were trained with three classes (non-causal, correlational, and causal) performed better than those trained with four classes (non-causal, correlational, conditional causal, and direct causal). Specifically, the differences between non-causal and correlational sentences were often difficult for the models to detect, which led to lower accuracy in distinguishing these categories. Given that the binary classifiers (causal vs. non-causal) were highly accurate, it was determined that a two-step approach would be the best way forward.

3.1. General-Purpose Datasets

We drew upon earlier work that introduced six datasets for analyzing causal sentences (Tan et al., 2023). However, as the Penn Discourse Treebank (PDTB) dataset is not openly available, it was excluded from our study. The following is a summary of the datasets that were utilized to construct the General Purpose Dataset.

AltLex
AltLex focuses on alternative lexicalizations of causal expressions within single sentences, primarily sourced from news articles. By emphasizing explicit causal markers, such as “caused by” or “due to,” AltLex provides valuable insights into direct causation. However, its small size and restriction to intra-sentence relations limit its applicability for modeling more complex causal structures across broader contexts (Chambers & Jurafsky, 2007).

BECAUSE 2.0
BECAUSE 2.0 employs a Construction Grammar framework to annotate spans of cause, effect, and connectives across a variety of sources, including congressional hearings and the New York Times Annotated Corpus. While it captures explicit causal relationships with high precision, its dataset size remains modest, and it lacks inter-sentence annotations, reducing its coverage of discourse-level causality (Dunietz et al., 2020).

CausalTimeBank (CTB)
Sourced from the TempEval-3 corpus, CausalTimeBank annotates explicit causal relationships between events within news articles. Its event-centric focus allows for inter-sentence annotations, yet the broader discourse context is often absent. This makes CTB well-suited for tasks requiring event-level causal reasoning but less applicable for comprehensive discourse analysis (Mirza et al., 2014).

EventStoryLine (ESL)
EventStoryLine extends the scope of causal event annotations by including both explicit and implicit relations between events. This dataset permits inter-sentence connections and implied causality, providing a richer context compared to CTB. However, its event-focused structure may not fully generalize to non-event-based causal claims found in scientific texts (Caselli et al., 2017).

SemEval 2010 Task 8
SemEval 2010 Task 8 focuses on semantic relations between noun phrases, including cause-effect, drawn from various web sources. It excels at capturing causation in an entity-to-entity manner but lacks broader contextual details, as it is restricted to inter-sentence relations. This limitation makes it less ideal for analyzing the nuanced causal language used in scientific writing (Hendrickx et al., 2010).

3.2. Causal Definition
The definition of the label causal in the general-purpose datasets for causal text mining, as described in the UniCausal paper and related sources, takes a broad approach that extends beyond direct causality. Sentences are labeled as causal if they contain linguistic cues that imply a causal relationship, even if the causation is not directly stated or empirically validated. For instance, in the AltLex dataset, causality is identified by explicit markers like “because,” “caused by,” or “due to,” signaling a cause-effect relationship. The BECAUSE 2.0 corpus annotates text to identify causal relationships, marking spans that represent causes, effects, and the connecting language, including instances where these relationships are conditional or hypothetical. CausalTimeBank and EventStoryLine adopt an event-centric focus, labeling explicit and implicit causal links between events as causal. Lastly, SemEval 2010 Task 8 applies causal labels to sentences featuring semantic relationships between entities (e.g., “X causes Y”) without requiring extensive contextual evidence. This broader interpretation captures a diverse range of causal expressions, allowing models to learn from implied or hypothetical causality as well as explicit statements.

3.3. Scientific Datasets
To better reflect the language of scientific writing, we integrated four additional datasets: PubMed, Haber, EurekAlert!, and Social Science. The first three come from the medical domain, which has its own distinct writing style. To ensure broader applicability, we also included a social science dataset from Norouzi et al., allowing the final model to perform well on social science papers and hopefully generalize better. This expansion was particularly important as we aim to test the model on psychological science texts.

Dataset	Source	Example
AltLex	News articles	”Finland did not accept the Soviet call for its land, so it was attacked in November 1939”
BECAUSE 2.0	Congressional Hearings, Penn Treebank, New York Times Annotated Corpus	”Roe vs. Wade, the 1973 case legalizing abortion, made fetal viability an important legal concept”
CausalTimeBank (CTB)	TempEval-3 corpus (news articles)	”The truck maker said the significant drop in net income will result in lower earnings for the fiscal year”
EventStoryLine (ESL)	Event Coreference Bank+	”Police are continuing to investigate a fire at a Waitrose supermarket in Wellington”
SemEval 2010 Task 8	Various web sources	”The tides caused by the sun follow the exact same methods as those by the moon”
EurekAlert	Medical related press releases	”Better use of lighting in hospital rooms may improve patients’ health.”
PubMed	Biomedical abstracts	”Apelin may not be directly involved in the regulation of maternal insulin sensitivity.”
Haber	Observational Health Research (full text and abstracts)	”Our findings provide supportive evidence that lifetime exposure to PM2.5 increases risk in infant mortality.”
Social Science	The Cooperation databank (full text)	”Leniency and forgiveness seem to be motivated by strategic concern rather than social preferences.”

Table 2.1: Overview of The Datasets

PubMed
This dataset consists of biomedical abstracts that includes technical causal statements often framed with cautious language. For example, a typical statement: “Apelin may not be directly involved in the regulation of maternal insulin sensitivity.” It consists of biomedical abstracts annotated with four labels (correlational, conditional causal, direct causal, and non-causal).

Haber
The Haber dataset systematically examines the use of causal and associational language in observational health research. It evaluates how medical and epidemiological studies describe causal relationships between exposures and outcomes. The dataset categorizes causal language into four levels (none: no causal implication, Weak: possible causal implication, but unclear, Moderate: likely causal implication, Strong: clear causal implication. By analyzing 1,170 articles from top medical and epidemiology journals (2010–2019), the dataset includes how causal language is often implied even when explicit causal claims are avoided.

EurekAlert!
This collection of press releases captures causal claims communicated to the public. By contrasting with academic texts, EurekAlert! sentences include four labels and emphasize contrasts between academic precision and media oversimplification. The dataset is a collection of press releases that capture how causal claims are communicated to the public. Unlike academic writing, which tends to be cautious and precise, media reports often oversimplify scientific findings, sometimes overstating causal links. Each sentence in the dataset is labeled according to how strongly it implies causality (correlational, conditional causal, direct causal, and non-causal).

Social Science Dataset

This dataset, curated by Norouzi et al., emphasizes causal relations in social science research, drawing from The Cooperation Databank (CoDa) (Spadaro et al., 2022), a large repository of studies on human cooperation. The dataset consists of 529 causal and 529 non-causal sentences, manually labeled to train machine learning models for extracting causal claims from full-text academic papers. Unlike traditional causal datasets, this dataset specifically addresses the subtle and ambiguous language used in social science. Causal relationships in this domain are often indirect, conditional, or implied, making them harder to detect.

4. Dataset Pre-processing

4.1. Binary Classification

For datasets such as PubMed, Haber, and EurekAlert!, that initially had four distinct labels, the labels were put together to align with binary classification requirements. Sentences annotated as correlational, conditional causal, or direct causal (or weak, moderate, and strong in the case of Haber) were consolidated under a single "causal" label, while non-causal sentences retained their original label. This transformation allowed for consistent integration with general-purpose datasets and the Social Science dataset.

Pre-processing

To finalize the datasets for training, validation, and testing, preprocessing was applied. After merging datasets from the nine sources, 19,020 duplicate sentences were removed to reduce redundancy and prevent bias in classification. The largest reduction occurred in the ESL dataset, where 15,423 duplicates were identified and removed, and the CTB dataset, where 1,432 duplicates were eliminated. Removing these duplicates was essential to prevent the model from overfitting to frequently repeated sentence structures. Initially, the dataset contained 25,794 sentences, including 15,982 non-causal and 9,812 causal sentences. After preprocessing, including duplicate removal and label corrections, the final dataset was cleaned and ready for classification. Label inconsistencies were addressed, particularly in datasets like PubMed, Haber, and Press Release, where multi-level causal labels were consolidated into a binary classification system.

Train-Test Split

Once data cleaning and label standardization were completed, the dataset was split into training, validation, and test sets. The training is 70% percent of the total dataset was used for training, 10% for validation, and 20% for testing, with undersampling ensuring an even distribution of causal and non-causal sentences. The final training set has 6,868 causal and 6,868 non-causal sentences, the validation set contains 981 causal and 981 non-causal sentences, and the test set has 1,963 causal and 1,963 non-causal sentences. Additionally, a separate social science test set was created with 96 causal and 96 non-causal sentences, ensuring a fair evaluation of classification accuracy within this domain.

4.2. Four-Level Classification

For the four-level classification data set, the focus was on aligning the labels in data sets such as PubMed, Haber, EurekAlert! and Social Science to a unified framework. This alignment ensured consistency in labeling sentences as one of four causal categories: correlational, conditional causal, direct causal, or non-causal.

Both the Haber and PubMed datasets are labeled according to four levels of causal strength. Despite differences in terminology, both systems aim to capture the degree of causal certainty expressed in a sentence, ranging from weak associations to definitive causal claims. Haber's framework includes *None*, *Weak*, *Moderate*, and *Strong*, while PubMed uses *Non-Causal*, *Correlational*, *Conditional Causal*, and *Direct Causal*. The alignment between these categories becomes clear when examining the linguistic cues employed in both datasets. For instance, Haber's *Weak* (e.g., "affect") aligns with PubMed's *Correlational* (e.g., "is associated with"), as both suggest tentative connections without asserting causation. Similarly, Haber's *Moderate* (e.g., "might lead to") maps to PubMed's *Conditional Causal* (e.g., "could result in"), indicating causality that is context-dependent or inferred. Finally, Haber's *Strong* (e.g., "causes" or "leads to") directly corresponds to PubMed's *Direct Causal*, where causation is explicitly as-

Source	Label	Train	Validation	Test	Combined
AltLex	Causal	289	44	82	415
	Non-Causal	258	25	65	348
BECAUSE 2.0	Causal	230	28	63	321
	Non-Causal	42	10	10	62
CausalTimeBank (CTB)	Causal	195	36	45	276
	Non-Causal	814	139	197	1150
EventStoryLine (ESL)	Causal	786	119	245	1150
	Non-Causal	484	49	135	668
SemEval 2010 Task 8	Causal	929	137	261	1327
	Non-Causal	4016	575	1198	5789
Press Release	Causal	1115	194	280	1589
	Non-Causal	207	22	65	294
PubMed	Causal	1210	152	340	1702
	Non-Causal	577	84	164	825
Haber	Causal	1742	226	535	2503
	Non-Causal	243	42	74	359
Social Science	Causal	372	45	112	529
	Non-Causal	227	35	55	317

Table 2.2: Sentence Distribution per Source for Train, Validation, Test, and Combined Datasets (Causal vs. Non-Causal)

serted. As these label definitions are relatively similar, we hypothesized that they could be combined into one dataset.

The Social Science dataset originally had binary labelled sentences: causal or non-causal. To make it compatible with the four-level classification, we manually re-labeled all causal sentences into one of three categories: correlational, conditional causal, or direct causal. This process used a coding scheme based on Yu et al. (2019) definitions from the PubMed dataset (see Appendix A). This process involved reviewing ambiguous cases in collaboration with Daniël Lakens to ensure consistency. Notably, four sentences that were initially marked as causal were reclassified as non-causal.

Pre-processing

The preprocessing and train-test split process begins with the removal of any duplicate rows in the dataset, ensuring that each data point is unique and eliminating any redundant information. After duplicates are removed, the dataset is filtered to exclude any samples labeled as ‘0‘ (non-causal), as these are not relevant. The two-level classifier will already classify a sentence into causal or non-causal. Therefore, this dataset leaves only the samples with labels 1, 2, and 3, ensuring the model will be trained solely on correlational, conditional causal, and direct causal labeled sentences.

Train-Test Split

Following this, the dataset is split into training, validation, and test sets. This split uses stratified sampling, a technique that ensures the proportions of each label (1, 2, and 3) are preserved across both the training and the remaining subsets. After the initial split, the next step involves undersampling

to address any potential class imbalance. The undersampling approach specifically focuses on the non-social science data (SSC) (Haber, PubMed, Press Release), reducing the number of non-SSC samples for each label (1, 2, and 3) to match the smallest class size. Importantly, all SSC samples are preserved in the final dataset, regardless of label distribution. This ensures that the model will work effectively on both the medical and social science text, with the SSC data fully represented.

The final dataset used for fine-tuning the three-level classification models consisted of PubMed, press releases, and social science texts. The Haber source was initially included; however, it was later excluded as its classifications were inconsistent across models (BERT, GPT, LLAMA) and did not align with our framework of correlation, conditional causal, and direct causal relationships. This misalignment introduced discrepancies in the training data, making it difficult for models to generalize effectively. Models fine-tuned with Haber achieved an F1 score of only 0.65, whereas excluding it and training solely on PubMed, press releases, and social science texts improved BERT-based model performance to an F1 score of 0.83.

Source	Label	Train	Validation	Test	Combined
Haber	Correlational	539	68	153	760
	Conditional Causal	753	112	194	1059
	Direct Causal	345	48	89	482
Press Release	Correlational	353	62	116	531
	Conditional Causal	184	31	69	284
	Direct Causal	260	42	80	382
Pubmed	Correlational	273	39	65	377
	Conditional Causal	153	9	51	213
	Direct Causal	465	66	132	663
Social Science	Correlational	42	3	11	56
	Conditional Causal	117	20	31	168
	Direct Causal	137	16	44	197

Table 2.3: Sentence Distribution per Source for Train, Validation, Test, and Combined Datasets (Three-Level Classification)

5. Training Models

5.1. Initial Experimentation: One-Shot and Few-Shot Prompting

Before starting supervised fine-tuning, one-shot and few-shot prompting were tested using GPT-4, following the approach described by Brown et al. (2020). In the one-shot paradigm, a single annotated example illustrating a specific causal label (e.g., correlational) was provided, while the few-shot paradigm used a handful of annotated examples. The goal was to evaluate whether GPT-4 could accurately classify sentences across four levels of causality (“no relationship,” “correlational,” “conditional causal,” and “direct causal”) with minimal training data. Only the Haber dataset was used for this.

Results indicated that while GPT-4 could handle straightforward sentences, it struggled with more nuanced cases. Correlational and conditional causal sentences were frequently misclassified. Moreover, ambiguous phrasing in observational studies proved challenging for GPT-4’s context-limited capabilities under few-shot conditions. These results indicated that a supervised approach is crucial for achieving higher accuracy.

5.2. Fine-Tuning Large Language Models

Fine-tuning is a process where a pre-trained large language models are further trained on domain-specific data to optimize its performance for a given task. It involves updating the model's parameters using labeled data, enabling it to generate more accurate and contextually relevant predictions. Fine-tuning is essential for adapting general-purpose language models to specialized applications, such as causal sentence classification in our case.

The fine-tuning process varies across the three models due to differences in their architectures and computational demands. For BERT, fine-tuning was conducted using the PyTorch framework (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf et al., 2020). The model was trained on labeled datasets specifically designed for causal classification. Given the extensive GPU hours required for fine-tuning, Kaggle's cloud infrastructure (Kaggle, n.d.) was utilized, providing the necessary computational resources to run the training process efficiently. A maximum of six epochs was selected to allow the models sufficient time to learn from the data. Since large models are prone to overfitting when trained on smaller datasets, additional epochs were not necessary. Early stopping was implemented to halt training if the F1 score did not improve for two consecutive epochs. In this case, training stopped, and the epoch with the highest F1 score was selected. To further mitigate overfitting, both validation loss and training loss were monitored when choosing the best model. If the validation loss began to increase again, those epochs were disregarded, as this indicated a risk of overfitting.

For LLAMA 3.2 1B, fine-tuning presented additional challenges due to the model's large parameter size. Training such large-scale models requires significant computational power, often exceeding the capabilities of standard GPU setups. To address this, Quantized Low Rank Adaptation (QLoRA) was employed, a parameter-efficient fine-tuning technique that enables training large models on a single GPU. QLoRA works by backpropagating gradients through a frozen, 4-bit quantized pre-trained model while updating a smaller set of Low-Rank Adapters (LoRAs) (Dettmers et al., 2023). This reduces memory usage without significantly sacrificing model performance. The bitsandbytes library (Dettmers et al., 2022), along with the Hugging Face PEFT framework, was used to implement this technique, allowing LLAMA 3.2 to be fine-tuned with limited computational overhead.

For GPT-4o-mini, fine-tuning was performed using OpenAI's API service, which enables users to adapt models to specific tasks without relying on extensive local computational resources. This cloud-based approach allowed for efficient model adaptation while utilizing OpenAI's optimized training environment. For these bigger autogenerative models it was chosen to only train one epoch which is common practice. We also saw that these models quickly learned the data, with both training and validation losses decreasing significantly in the early steps and stabilizing at low values. Since the losses have converged and show no signs of overfitting, training for more epochs would be redundant and unlikely to yield further improvements.

Throughout the fine-tuning process, Weights & Biases (W&B) was used to log training progress and monitor key metrics. W&B facilitated real-time tracking of loss curves, model checkpoints, and hyperparameters, providing valuable insights into the optimization process. The seamless integration between Hugging Face and W&B ensured efficient experiment logging, with results saved locally and synced to the W&B cloud platform. All code and datasets used in this research are available in the following GitHub repository[2].

5.3. Three Candidate Models

We compared three transformer-based language models: BERT (Devlin et al., 2019), LLAMA 3.2 (AI, 2024), and GPT-4o mini(OpenAI, 2024). Each model uses one of two pre-training methods: Masked Language Modeling (MLM) or Autoregressive Language Modeling. These methods determine how the models learn language representations before being fine-tuned for specific tasks.

BERT

BERT follows the MLM objective, in which tokens within a sentence are randomly masked, and the model is trained to predict these missing tokens using surrounding context (Devlin et al., 2018). This bidirectional learning process enables BERT to develop a deep understanding of language structure and relationships, making it particularly well-suited for classification tasks where capturing fine-grained

[2] https://github.com/Tessavana/Causal-Clarity

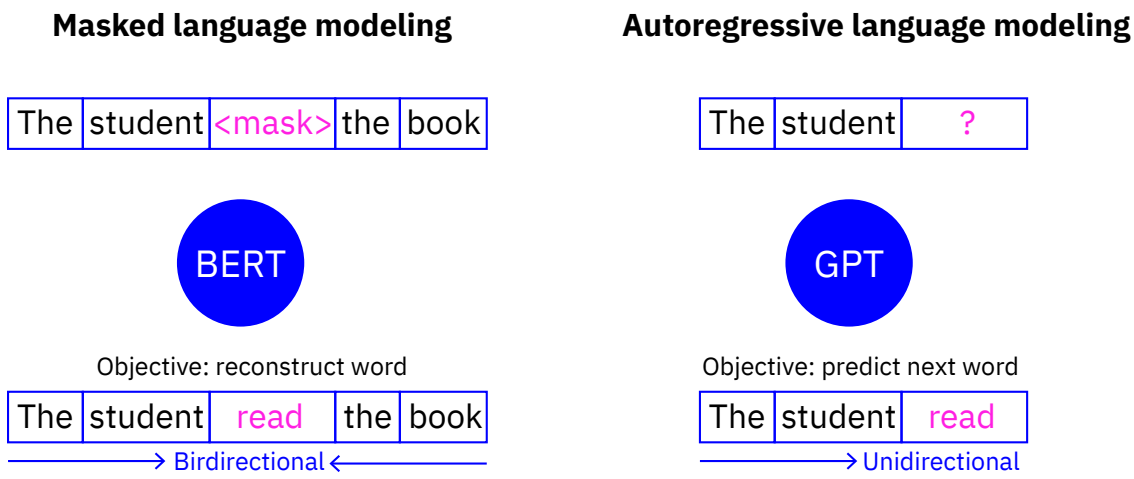


Figure 2.2: Masked Language Models vs Autoregressive Language Models

linguistic nuances is important. Previous causality extraction studies (Tan et al., 2023) have shown BERT's strong performance, making it a reliable tool for sentence classification. Also, when fine-tuning BERT models, a specific fine-tuned classification layer is added, making BERT models very suitable for classification tasks where a fixed response is desired. SciBERT (Beltagy et al., 2019) is introduced to enhance performance in three-level classification by leveraging its pretraining on scientific texts. Since the training data consists of scientific sources, SciBERT is a logical choice due to its domain-specific vocabulary and linguistic patterns, which could improve causal classification. Its exposure to structured academic writing is expected to help distinguish between correlational, conditional causal, and direct causal relationships more effectively than standard BERT.

RoBERTa (Liu et al., 2019) was also tested in three-level classification due to its optimized training process, which enhances masked token learning. Unlike SciBERT, RoBERTa is trained on a significantly larger and more diverse dataset, improving its ability to generalize across different contexts. Given these strengths, SciBERT and RoBERTa were introduced to assess whether their respective enhancements could further improve causal classification accuracy for three level classification.

LLAMA 3.2 and GPT-4o-mini
LLAMA 3.2 and GPT-4o-mini, on the other hand, are built on the autoregressive language modeling paradigm. Unlike MLM, autoregressive models involves predicting the next token in a sequence given all preceding tokens (Touvron et al., 2023). This unidirectional approach is advantageous for text generation tasks but can also be adapted for classification. LLAMA 3.2 1B features an optimized architecture that enhances both efficiency and performance, making it a strong choice for handling longer contextual dependencies. The 1-billion parameter version was selected to reduce the computational resources required for fine-tuning. Meanwhile, GPT-4o-mini represents a scaled-down version of the GPT-4 model, optimized for efficient deployment while maintaining high performance in generative and classification tasks. GPT-4o-mini was selected to minimize costs and training time, with the fine-tuned models now costing less than 3 euros for training and validation.

6. Model Evaluation

In evaluating our fine-tuned models, we used standard performance metrics widely recognized in machine learning including: precision, recall, F1-score, and primarily the macro-average F1-score. Precision is defined as the number of true positive predictions divided by the total number of observations. Recall is the ratio of true positive predictions to the total number of true positive observations. The F1 score is the harmonic mean of precision and recall, measuring a model's ability to balance both metrics. The macro-average F1 score is calculated by computing the F1 score for each class separately and then averaging these scores equally, regardless of class size. This metric is widely used in machine learning research because it provides a detailed per-class performance assessment, ensuring that all categories are weighted equally, even when some are more difficult to classify. Even when

class distributions are perfectly balanced, the macro-average F1 score remains useful as it captures subtle performance differences across categories, such as correlational, conditional causal, and direct causal relationships (Lipton et al., 2014).

In addition to standard evaluation metrics, we conducted a misclassification analysis to identify common errors. This analysis highlighted cases where the model misclassified correlational statements as causal or confused conditional and direct causal claims. Understanding these error patterns is critical for guiding further refinements. Additionally, we examined the distribution of misclassifications across different sources and analyzed the occurrence of Type I and Type II errors for each source.

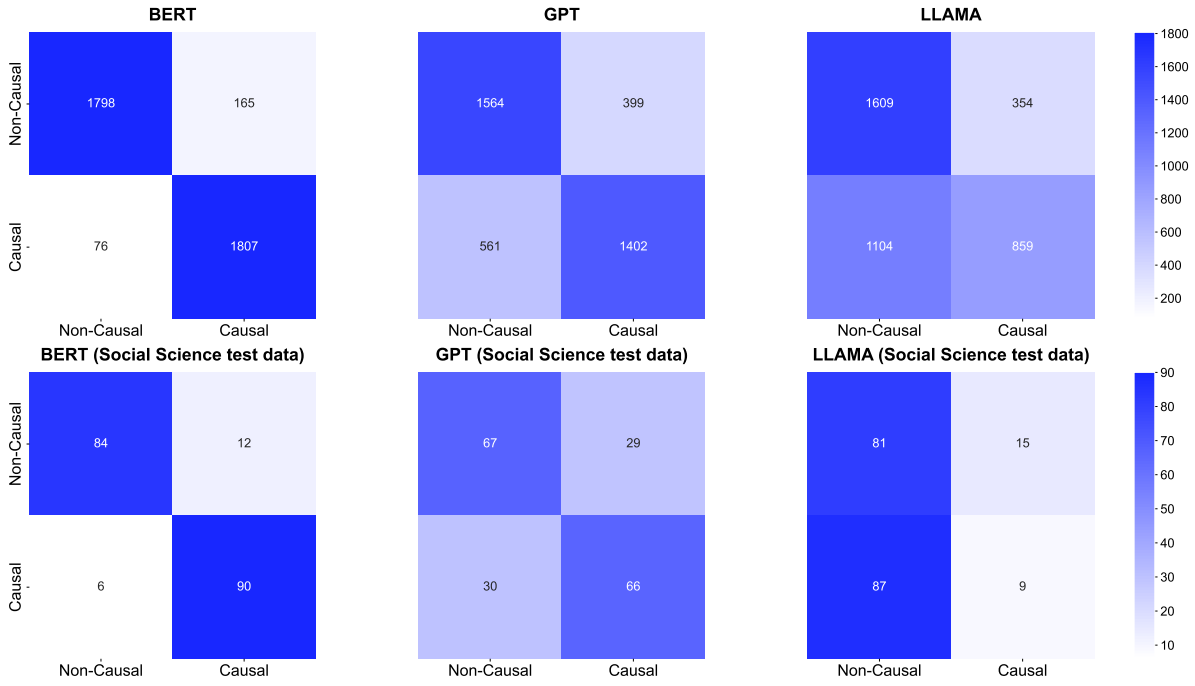
6.1. Social Science Evaluation

Moreover, we also evaluated the models on social science data exclusively. Although our training data primarily originates from medical and general scientific texts, one of our goals is to deploy these models on social science papers. Social science literature often features distinct linguistic patterns and terminologies that differ from those found in other domains Healy. By separately assessing model performance on social science data, we can determine whether the models generalize well to this domain or if further domain-specific fine-tuning is needed. Norouzi et al. found strong evidence of a domain shift, where models trained on general-purpose datasets performed significantly worse on social science text. Their results showed that fine-tuning on even a small portion of social science data led to notable improvements in model performance, showing the importance of domain adaptation. This separate evaluation is performed to determine if we can replicate this effect in our models.

3

Results

1. Two level classification results



1.1. Performance Metrics

From our results, BERT achieves the highest classification performance, maintaining an F1-score of 0.94 on the full test set, with a precision of 0.92 and a recall of 0.96. The high recall indicates that the model is effective at reducing false negatives, meaning it rarely misses actual causal sentences. The consistency of these metrics suggests that BERT accurately captures causal relationships while minimizing classification errors.

GPT-4o-mini underperforms compared to BERT, achieving an F1-score of 0.70, with both precision and recall around 0.70–0.71. Since precision and recall are nearly the same, this suggests that the model does not favor either false positives or false negatives. LLAMA-3.2.1-B demonstrates the weakest performance, with an F1-score of 0.68, precision of 0.61, and recall of 0.77. The significantly higher recall relative to precision suggests that LLAMA-3.2.1-B over-predicts causality, leading to an excessive number of false positives. While it detects causal relationships, its low precision indicates its inability to differentiate causal and non-causal statements effectively.

Model	Test Data	Precision	Recall	F1-score	Macro F1-score
BERT	Full Test Set	0.92	0.96	0.94	0.94
	Social Science Test Set	0.88	0.94	0.91	0.91
GPT-4o-mini	Full Test Set	0.70	0.71	0.70	0.70
	Social Science Test Set	0.69	0.69	0.69	0.69
LLAMA-3.2.1-B	Full Test Set	0.61	0.77	0.68	0.68
	Social Science Test Set	0.38	0.09	0.15	0.38

Table 3.1: Performance Metrics for Different Models on Full and Social Science Test Sets

1.2. Social Science

When testing the models on the social science test set we see that in general the performance metrics are a lower than the average for the social science sentences. The biggest drop in performance can be seen with the LLAMA model dropping from 0.68 to 0.38 macro F1-score. This indicates that LLAMA-3.2.1-B misclassifies nearly all causal statements as non-causal, demonstrating a complete inability to transfer learned causal knowledge to the social science domain. The more minor performance decline of BERT and GPT suggests that they effectively adapted to linguistic variations in social science texts, maintaining strong feature transfer across domains.

1.3. Misclassifications analysis

The normalized misclassification rate figure 3.1 adjusts the misclassification rates based on the frequency of each dataset appearing in the training and validation sets, ensuring a fair comparison across the different sources. A key observation is the notably high misclassification rate for the Haber dataset, particularly for the LLAMA model. This suggests that the Haber dataset may include more complex or ambiguous examples that the model struggles to classify correctly. Furthermore, datasets such as Social Science and PubMed exhibit relatively high misclassification rates across multiple models. This indicates that scientific and domain-specific texts, which often contain more complex structures and terminology, are more difficult for these models to classify accurately. In contrast, datasets like ctb and semeval show relatively lower misclassification rates, suggesting that these datasets are easier for the models to process and classify.

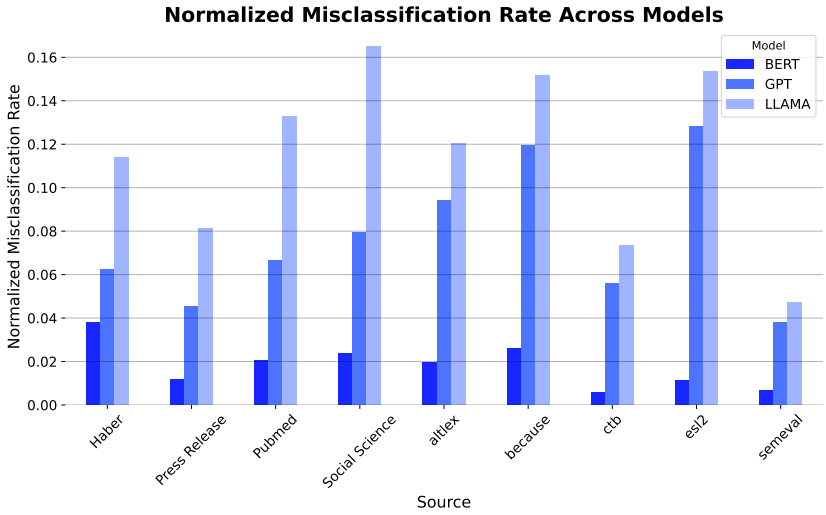


Figure 3.1: Misclassifications 2-level models per source normalized

The type I vs. type II Errors Across Models figure 3.2 shows that different datasets exhibit distinct tendencies toward false positives (Type I errors) or false negatives (Type II errors). This figure represents

errors aggregated across all models. In many datasets type II errors are more frequent than Type I errors, indicating that the models are more likely to miss causal relationships rather than falsely identifying them. This suggests that the models for these datasets tend to be overly conservative, favoring specificity (avoiding false positives) at the cost of recall (missing true causal relationships). In contrast, dataset ctb shows the opposite pattern, where Type I errors exceed Type II errors, meaning that the models are more prone to incorrectly classifying non-causal relationships as causal. This variation across datasets suggests that certain types of texts or linguistic structures may make the models more cautious in some cases and more prone to overgeneralization in others. showing that each domain has its own specific causal language.

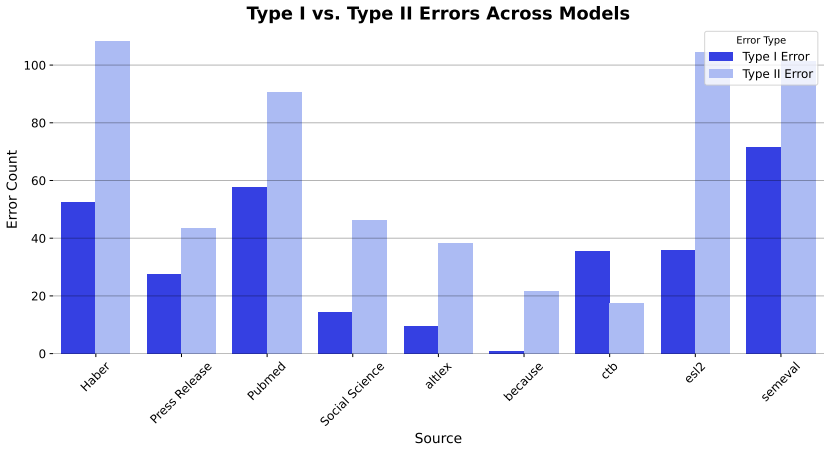
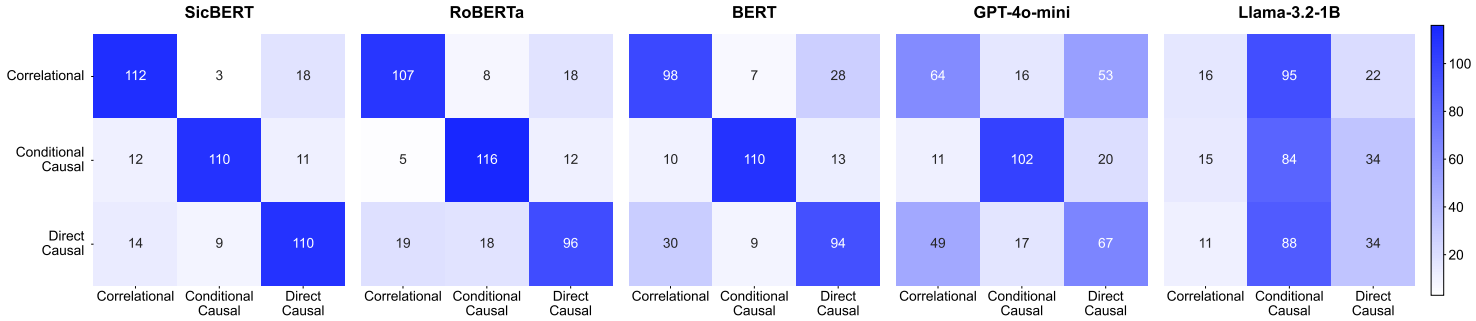


Figure 3.2: Type I and Type II errors per source across models

2. Three level classification



	Correlational			Conditional Causal			Direct Causal			
Model	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Macro F1-score**
SciBERT	0.81	0.84	0.83	0.90	0.83	0.86	0.79	0.83	0.81	0.83
RoBERTa	0.82	0.80	0.81	0.82	0.87	0.84	0.76	0.72	0.74	0.80
BERT	0.71	0.74	0.72	0.87	0.83	0.85	0.70	0.71	0.70	0.76
GPT-4o-mini	0.48	0.85	0.61	0.76	0.78	0.77	0.50	0.50	0.50	0.63
Llama-3.2-1B	0.12	0.52	0.20	0.63	0.45	0.52	0.25	0.25	0.25	0.33

Table 3.2: Performance Metrics for Different Models on Full Test Set for Three-Level Classification (Including Macro F1-score)

2.1. Performance Metrics

SciBERT achieves the highest overall performance, with a macro F1-score of 0.83. It maintains high recall (0.84) for correlational classifications while achieving the highest precision (0.90) in conditional causal classification. RoBERTa follows with a macro F1-score of 0.80, showing strong performance in conditional causal classification (recall = 0.87) but lower precision (0.76) in direct causal classification, where it tends to misclassify some conditional causal instances as direct causal. While the perfomance of these models is less accurate, causal classification models do not necessarily need to be flawless to be useful. Even with some misclassifications, they can still provide valuable feedback by flagging potentially unclear or inconsistent causal claims, allowing researchers to refine their language. The utility of these models lies in their ability to assist rather than replace human judgment, offering structured insights that help improve the clarity of scientific communication.

BERT achieves a macro F1-score of 0.76. It performs worse in correlational classification (F1-score = 0.72) and shows lower recall (0.71) in direct causal classification compared to SciBERT and RoBERTa. GPT-4o-mini and Llama-3.2.1B exhibit lower overall performance. GPT-4o-mini achieves a macro F1-score of 0.63, with lower precision and recall in correlational and direct causal classification. Llama-3.2.1B has the lowest macro F1-score (0.33), failing to distinguish causal relationships effectively. The confusion matrices show mainly misclassifications between direct causal and correlational categories, with direct causal often being misclassified as correlational and vice versa. This suggests that the model may be struggling to differentiate between these two categories, despite their conceptual differences, possibly due to overlapping linguistic features.

2.2. Misclassification Analysis

Based on the analysis of the normalized misclassification rate figure Figure 3.3, we can conclude that all models perform quite well on the Social Science dataset. The low misclassification rates across all models suggest that the labeling in this dataset is relatively consistent and learnable. This indicates that the models are able to effectively capture the causal relationships present in the social science texts. However, looking at the best performing SciBERT model you do see it has the most misclassifications in Social science compared to the other sources. On the other hand, PubMed, which contains scientific abstracts, exhibits the highest misclassification rates, particularly with models like GPT and LLAMA. This suggests that the complexity of scientific language and domain-specific terminology creates challenges for the models.

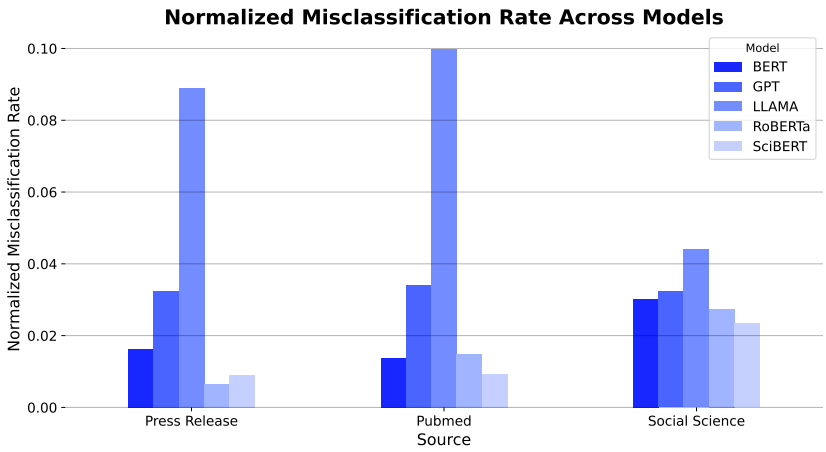


Figure 3.3: Misclassifications 3-level models per source normalized

Figure 3.4: Type I and Type II errors per source across models

2.3. Social Science

After reviewing the misclassified social science sentences, three key patterns emerged that likely contributed to errors in classification. These patterns involve sentence complexity, specialized terminology, confusion between conditional and direct causality, and overlooked conditional markers.

Overly Complex, Multi-Clause Sentences

Many misclassified sentences contain multiple clauses, conditional statements, or references to different concepts, making them difficult for the model to interpret. When multiple conditions are present, the core causal claim can become buried, leading to misclassification.

For example, the sentence:

> *Otherwise, if the accumulated gain explains the behavior of the birds but the one-trial gain does not, then the birds are making decisions based on the four-trial reward and are playing a mutualism game.*

was originally labeled as conditional causal but was misclassified as correlational. The phrase ”if the accumulated gain explains the behavior” introduces a clear condition for the effect, but the model may have struggled with the multiple clauses and references to different variables, making it difficult to extract the primary causal relationship.

Specialized or Ambiguous Terminology

Certain misclassified sentences contain domain-specific jargon or ambiguous phrasing, leading to confusion in classification. Words such as ”feedback,” ”activation,” or ”reinforcement” may have precise

technical meanings in specific disciplines that differ from everyday usage. If the model is not trained to recognize these distinctions, it may fail to interpret them correctly in context.

For example, the sentence:

> *Consequently, our results indicate that emoticons mainly affect behavior through the communication channel provided with it (feedback as pre-play communication).*

was originally labeled as direct causal but was misclassified as correlational. The phrase ”affect behavior through” suggests a causal mechanism, yet the model likely misinterpreted ”feedback as pre-play communication” as a neutral association rather than an explanatory factor. This suggests that the model struggles to differentiate between causal mechanisms and general descriptive explanations in specialized terminology.

Confusion Between Conditional and Direct Causality
The model frequently misclassifies conditional causal statements as direct causal, likely because it does not give enough weight to conditional markers such as ”if,” ”provided that,” or ”depends on.” Instead of recognizing that an effect depends on a specific condition, it often interprets the relationship as a direct cause-and-effect link.

For example, the sentence:

> *If participants believe that others will cooperate, then they are more likely to contribute to the public good.*

was originally labeled as conditional causal but was misclassified as direct causal. The phrase ”if participants believe” sets a clear condition for cooperation, meaning the causal relationship is not absolute but contingent on the given belief. However, the model likely overfocused on ”more likely to contribute,” misinterpreting the sentence as stating a direct causal effect rather than a conditional one.

Overlooked Conditional Markers
Another common misclassification occurs when the model fails to detect key conditional phrases that distinguish conditional causality from direct causality.

For example, the sentence:

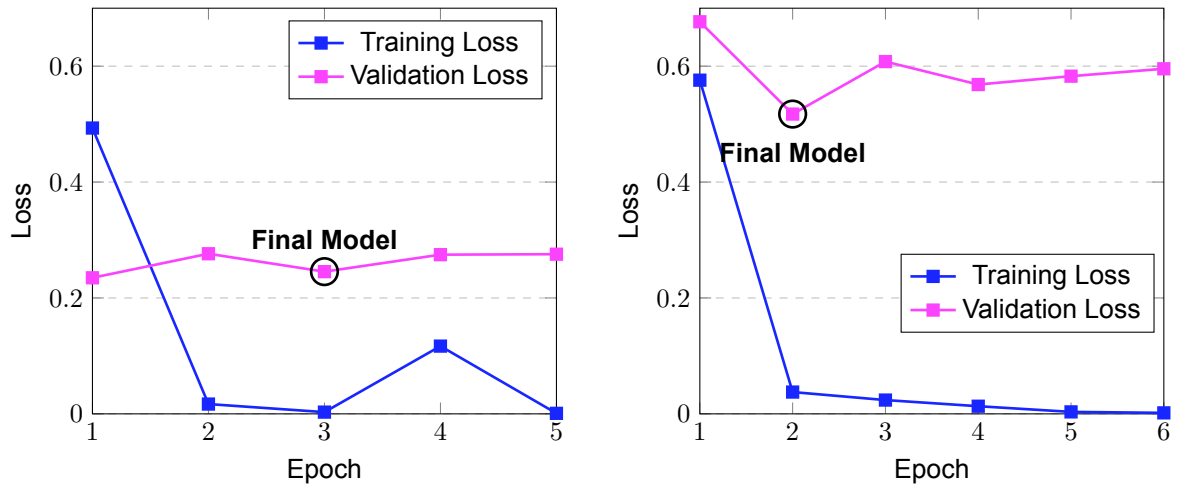
> *But since the payoff for mutual defection is sometimes low, the model suggests that if players have the opportunity to cooperate, they might choose to do so.*

was originally labeled as conditional causal but was misclassified as direct causal. The phrase ”if players have the opportunity to cooperate” clearly introduces a condition, meaning the effect (choosing to cooperate) depends on a specific situation. However, the model likely ignored the conditional phrasing and instead focused on ”they might choose to do so,” leading to an incorrect classification.

Implications for Causal Inference in AI Models
Based on this analysis, we can identify strategies to improve causal inference in both human interpretation and large language models. Complex sentence structures with multiple clauses often obscure causal meaning. Breaking them into shorter, more digestible statements can improve clarity. Using explicit causal markers such as ”causes,” ”results in,” and ”leads to” ensures that causal relationships are clearly stated rather than implied. Conditional causality should be explicitly signaled using structured phrases such as ”only when” or ”under the condition that” to prevent confusion between conditional and direct causality. Domain-specific terms should be clearly defined in training data to ensure models correctly interpret their causal implications. Different academic fields may define causality differently, and models must be trained to account for these variations. For instance, in psychology, ”reinforcement” implies a causal effect on behavior, while in economics, ”correlation” may often be explicitly non-causal. Training models to recognize these discipline-specific nuances can reduce misclassifications.

3. Learning Curves of the two best performing models

By looking at the training and validation loss both models show effective early learning, followed by stabilization and potential overfitting in later epochs. For BERT, training loss drops significantly by



(a) BERT 2-level: Epoch 3 was chosen as the final model.

(b) SciBERT 3-level: Epoch 2 was chosen as the final model.

Figure 3.5: Learning curves of the final best performing models

epoch 2 and remains low, while validation loss stays nearly constant across epochs. SciBERTs mainly horizontal validation loss curve suggests that the model's generalization performance remains stable, meaning further training does not lead to significant improvements but also does not introduce overfitting. The model has likely already captured the key patterns in the dataset, making epoch 3 the optimal stopping point based on the highest F1 score. In contrast, SciBERT (3-level) shows a sharp drop in both training and validation loss in the first two epochs, with validation loss reaching its lowest point at epoch 2. However, unlike BERT, SciBERT's validation loss starts to increase slightly in later epochs, suggesting overfitting. This pattern indicates that while early training effectively improved generalization, continuing training beyond epoch 2 led the model to fit too closely to the training data, reducing its ability to generalize to unseen examples. The final models for both were selected based on the highest F1 score, ensuring an optimal balance between minimizing loss and preventing overfitting.

4

Tool Development and Implementation

1. Tool Architecture

After fine-tuning the models, we developed a multi-step pipeline to analyze scientific manuscripts and provide actionable feedback on causal statements. The process follows several key steps to make our tool. A simplified demo of the tool can be found via this link[1].

The first step, **document preprocessing**, uses GROBID to convert PDFs into XML, stripping out figures, tables, and references ("GROBID", 2008–2024). A custom Python script then extracts clean text from key sections, including the abstract, methods, results, and discussion, along with metadata such as the title, authors, and DOI. The introduction is excluded because researchers often reference prior work rather than making direct claims, which could lead to false classifications.

Next, the **causal sentence classification** step involves a fine-tuned BERT classifier that identifies whether a sentence is causal or non-causal. Sentences classified as causal undergo a secondary classification step to determine the type of causality, such as correlational, conditional causal, or direct causal. The system then highlights causal sentences in different colors based on classification, while non-causal sentences remain unhighlighted. To provide users with information about the models, an explanatory document has been created, as detailed in Appendix C1.

To enhance interpretability, the tool incorporates **summarization** features. A stacked bar chart presents the distribution of causal statements across manuscript sections. Additionally, LLAMA-3.3-70B generates a method-section summary, capturing key methodological aspects that define the study's causal validity.

Since different domains have specific guidelines for causal language, users can select a **guidelines mode** that best aligns with their research context, as outlined in Appendix C2. Two modes are available:

- **Strict mode:** Only RCTs can use causal language, while non-RCTs (observational studies) must use the term "association" instead.
- **Moderate mode:** Allows well-controlled observational studies to make causal claims if they account for confounders.

To improve accuracy, the system applies **personalized warnings** through rule-based triggers. These warnings flag potential issues, such as mismatches between causality type and the study title or abstract. The rules, derived from best practices in the literature (see Appendix C3), provide targeted alerts to help users refine their causal claims.

For users who wish to further educate themselves on clear causal inference writing, several supporting documents have been created. One of these includes a list of the most precise words to use for each category, ensuring that readers interpret sentences as intended. This includes clear correlation terms like "association" and distinct causal terms, as outlined in Appendix C4. Another document highlights

[1] https://huggingface.co/spaces/Tessava/Causal-Clarity

four recommended papers that provide essential insights into causal inference, as detailed in Appendix C5.

The **sidebar** on the right shows classification labels (e.g., “conditional causal”) and flags cases where classification confidence falls below 0.90. Users can also request explanations from LLAMA-3.3-70B for specific classifications and ask follow-up questions to clarify uncertainties. Finally, the **confidence score slider** allows users to adjust the strictness of displayed classifications, filtering out lower-confidence predictions if desired.

This pipeline provides researchers with an intuitive tool to evaluate causal claims, ensuring clarity and consistency in scientific writing.

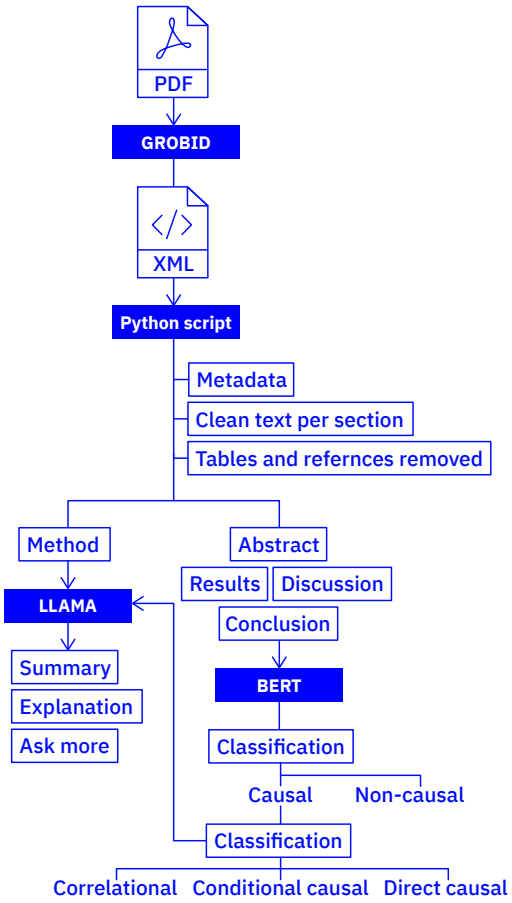


Figure 4.1: Tool architecture

1.1. GROBID

GROBID (GeneRation Of BIbliographic Data) (“GROBID”, 2008–2024) is a machine learning library specifically designed to extract and parse information from scholarly documents, such as PDFs, and convert them into structured XML/TEI formats. It uses a combination of Conditional Random Fields models and deep learning architectures, trained on high-quality, manually annotated datasets, to process the structural and textual elements of scientific papers. By analyzing both the textual content and its layout features, GROBID is able to accurately extract metadata, such as titles, authors, affiliations, and abstracts, as well as segment documents into sections like introduction, methods, results, and discussion. Additionally, it provides robust reference parsing capabilities, which are critical for bibliographic tasks.

In this thesis, GROBID will be used to streamline the preprocessing of scientific papers by performing three critical tasks: removing references and tables, collecting metadata, and categorizing text into sections based on the section titles detected by GROBID. This segmentation is particularly important,

as the research focuses on isolating specific sections—such as results and discussion—where causal claims are most relevant, while excluding sections like the introduction or methods, where such claims are unlikely to appear. Additionally, GROBID ensures that irrelevant content, such as tables, is excluded from the analysis. By enabling a precise focus on the relevant sections of a paper and outputting clean sentences, GROBID not only enhances the efficiency of the research process but also improves the accuracy and specificity of causal claim identification, which is critical for achieving the study's goals.

1.2. Python Script

To refine GROBID's XML output for analysis, a Python script [2] was developed to ensure the accurate extraction and structuring of scientific papers. Using libraries such as `lxml`, `nltk`, and `pandas`, the script processes XML files to extract relevant sections (e.g., Abstract, Methods, Results) and metadata (e.g., title, authors, DOI, keywords). The script classifies sections by parsing `<tei:div>` elements and recognizing headers via `<tei:head>` tags, aided by predefined keyword matching (e.g., "abstract," "methods," "results"). It also aggregates content within these sections using `nltk` to tokenize and count words.

A key feature of the script is its handling of multi-study papers, where methods and results are often presented sequentially for each study (e.g., "Study 1: Methods, Results; Study 2: Methods, Results"). In these cases, all "Methods" sections are combined into a single cohesive section, as are all "Results," ensuring no fragmentation and maintaining logical consistency. Additionally, tables and references are systematically removed to focus solely on textual content. By integrating these steps, the script produces a clean, consistent dataset optimized for causal claim detection, addressing the complexity of multi-study papers while filtering irrelevant content.

1.3. Groq API

In this research, the Groq API is used to generate structured summaries of the method sections from social science research papers using the latest Llama models. The Groq API was chosen over the more widely known OpenAI API because it offered free access, is faster, and has the latest large language models available for use. The API processed text with the following prompt:

```
prompt = f"""
You are a helpful assistant. Please analyze the **Method** section from a social

science or behavioral research paper.
**Respond** in a structured format with headings exactly like this (be very concise):
## 1. Summary (max 100 words)
[Your concise summary of the methodology and study design.]
## 2. Study Type
[One of: "Randomized Controlled Trial (RCT)", "Cohort Study", "Case-Control Study",
"Cross-sectional Study",  "Quasi-experimental", "Observational", etc.]
## 3. Randomization
["Yes" or "No"]
## 4. Groups or Comparisons
[Max 4 groups, separated by commas]
## 5. Mediators / Confounders
[List any mediators/confounders, look for phrases like: mediated by... controlled
for... If none, "Not stated". Comma-separated.]
## 6. Data Collection Method
[List data collection methods. If none, "Not stated". Comma-separated.]
## 7. Causal Inference Implications
[Explain what type of causality (correlational, conditional causal, or direct causal)
is allowed with this study design]
**Method Text**:
method_text
"""
```

[2] https://github.com/Tessavana/Causal-Clarity/blob/main/Tool/Qualitative_study.ipynb

The API's output summarized key aspects, including study design, confounders, randomization, data collection methods, and the causal inference implications.

1.4. Tool Design

To implement fine-tuned models, a tool [3] was created to visually display the classifications. The architecture described above powers these features, and the entire interface is coded in Python.

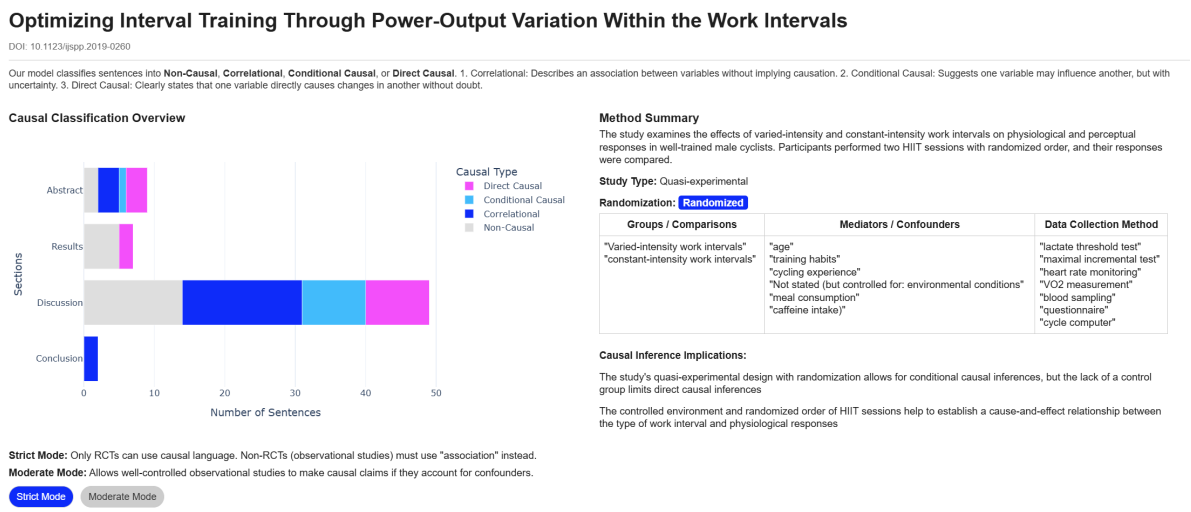


Figure 4.2: The overview chart and method summary

The design uses bright colors to give the tool a modern, engaging look and to create different shades within a single color. Bright colors are applied to charts and action buttons. Interaction elements are highlighted in bright dark blue to clearly indicate that they are clickable, making it easy for users to identify these areas. Medium brightness is used to highlight sentences within the paper, while the lightest color is used to indicate low-confidence predictions from the model, signaling that the model is less certain about these predictions. When you hover over a low-confidence sentence, a wavy underline appears, and a "LOW CONFIDENCE" message is displayed in the sidebar.

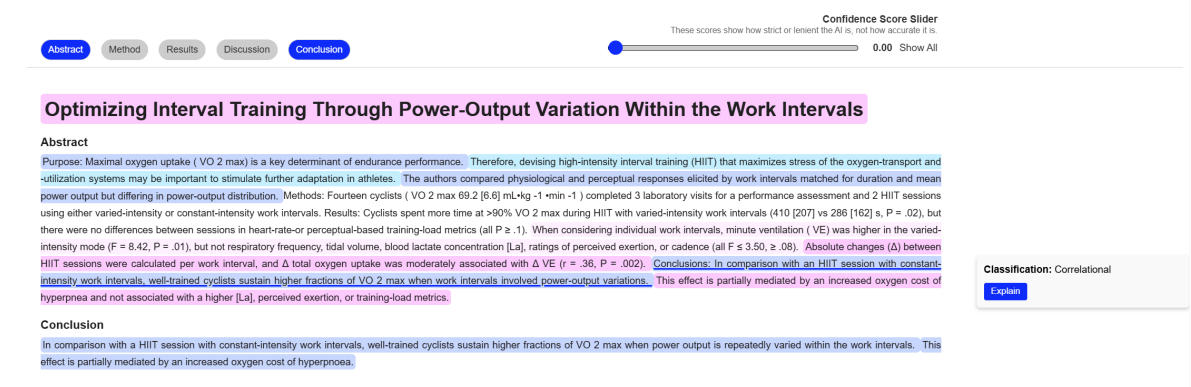


Figure 4.3: Classified sentences

The blue section buttons allow you to toggle sections like the abstract and conclusion in and out of view. This makes it easy to quickly analyze whether your causal statements match between these sections or to check for overstatements in the conclusion or even in the title.

When hovering over any sentence, a bright underline shows which sentence you are analyzing. Clicking on a sentence "locks" it in the sidebar, meaning the details stay open for that sentence until you click

[3] https://github.com/Tessavana/Causal-Clarity/blob/main/Tool/Qualitative_study.ipynb

another sentence or click the same one again. This feature helps you focus on a specific sentence while interacting with the LLAMA model.

The interface is designed to be clean, with the right sidebar providing additional information without obstructing the main content. This layout ensures that the paper you are analyzing remains clear and readable. Warnings are displayed in gray with blue highlights to stand out against the mostly white interface. These warnings contain important best practices and tips for improving the clarity of causal writing. By clicking “view details,” you can see a list of sentences related to each warning.

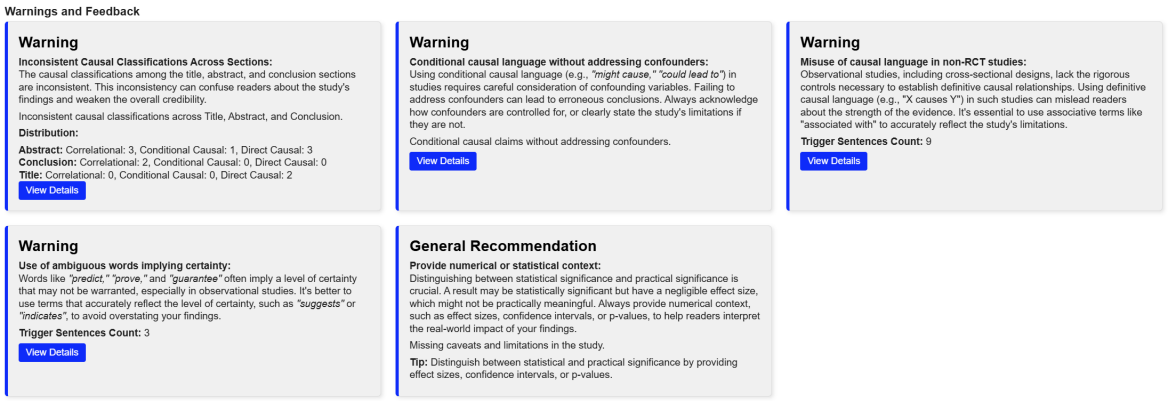


Figure 4.4: The warnings

2. Tool Evaluation

Five PhD candidates participated in this study, each with academic backgrounds in for instance human-computer interaction (HCI), philosophy, and meta-science. They provided verbal consent to participate. Participants reviewed their own research papers using the tool. The papers covered experimental, perception, and usability studies in sports science, psychology, AI, and HCI, as well as an opinion piece on scientific coordination. This diverse selection ensured the tool was evaluated across different research methodologies and writing styles.

The study followed a semi-structured interview format (see Appendix B for details). At the start of the session, participants were introduced to the tool and its intended use cases:

1. For researchers: to refine causal claims in their own writing.
2. For peer reviewers: to assess causal language in submitted papers.

To capture initial impressions, participants were asked to engage in a think-aloud exercise while exploring the interface. This allowed them to verbalize their thought processes and provide immediate feedback on usability and clarity.

Next, participants were encouraged to test specific functionalities, including:

- Color-coded highlights for different strengths of causal language.
- Leniency vs. strictness in causal classification.
- Confidence scores and low-confidence warnings to assess trust in the model’s predictions.
- Feedback mechanisms, including personalized warnings, explanations, and conversational options.

Participants provided feedback on the usefulness, interpretability, and clarity of the tool’s output. The semi-structured format allowed for in-depth discussions about preferences for feedback style (e.g., conversational explanations vs. direct warnings), the balance between leniency and strictness, and potential improvements.

Finally, participants were asked about future applications of the tool in their research workflow and what features they would like to see improved or expanded.

2.1. Qualitative study results

The interviews with five PhD candidates provided an understanding of their experiences using the tool. They explored its usability, effectiveness in identifying causal statements, and the clarity of its feedback. Participants reflected on how the tool influenced their writing process, whether it aligned with their expectations, and how it could be improved. Their insights reveal both the tool's potential benefits for academic writing and the challenges that need to be addressed to enhance its functionality.

Causal inference

A finding from the interviews is that early-career researchers often do not actively think about how they frame causal claims, but as they gain experience, they become more aware of causal language use in their writing. All participants indicated that over they years they became more aware of their causal language use. This indicates that AI-assisted causal classification tools could serve as educational resources, helping researchers refine their causal statements and avoid misrepresentations.

General Perception of the Tool

Overall, the tool was well received, with most participants finding it intuitive and beneficial for improving their causal writing. They appreciated its ability to highlight causal claims, provide structured feedback, and encourage reflection on their language choices. Some noted that the tool's color-coded classification system was particularly useful in drawing attention to specific areas of concern. A common misconception among users was that every highlighted sentence required review or was potentially incorrect. Some users felt compelled to check all highlighted sentences, assuming they indicated errors. However, the highlights merely classify sentences based on causal strength and do not assess their correctness. This led to some participants to be overwhelmed with the amount of highlights. While the majority found the highlighting effective, a few mentioned that additional context or explanations would help clarify why certain sentences were flagged.

Observations indicated that users typically began by reviewing the elements at the top, such as the chart and study summary, before moving down to the warnings. However, many skimmed over the warnings or skipped them entirely, often going straight to the highlighted text. Two participants noted that the warnings were not prominent enough and felt too text-heavy and theoretical, making them less effective in capturing attention. Some participants expressed a desire for the tool to include introductions, as they often contain hypotheses and other relevant statements they would like to review. Ideally, they wanted the tool to differentiate between claims made by the paper's author and those derived from external literature.

Strictness of Classification and Trust in Model Decisions

Participants held mixed opinions regarding how strict or lenient the tool should be in flagging causal language. Most preferred a lenient approach that highlighted a broad range of sentences, allowing them to critically assess their claims before making revisions. They would rather have false positives than missing an overstatement or understatement they might have made in their paper. Especially in sections as the conclusion or abstract they would want the tool to be lenient. Others wanted a stricter mode that only flagged sentences with a high likelihood of causal misstatement, particularly for peer review purposes or the results section.

One key issue raised was the trustworthiness of the model's classification decisions. Several participants noted that once they encountered a few misclassifications, they became less confident in the tool's overall reliability. They expressed a need for better explanations of why a sentence was flagged, particularly when the classification was uncertain. One participant stated that they would be more inclined to trust the tool's suggestions if it provided a clearer explanation for its decisions, such as highlighting the specific words or phrases that triggered the classification.

Perceptions and Improvements for the Warning System

The tool's warning system was one of the most debated features in the interviews. While participants generally found it helpful, they felt that its effectiveness depended on how the warnings were presented. A common request was for more actionable and targeted warnings. Participants suggested that instead of triggered alerts, the tool should provide specific recommendations for revision, including examples of how to rephrase sentences to make causal claims clearer. One participant proposed that warnings

should be linked directly to flagged sentences, allowing users to immediately navigate to the relevant part of the text and understand the issue in context.

Another concern was the visual presentation of warnings. Several participants felt that warnings were not sufficiently prominent and could easily be overlooked. They suggested that the warnings should be more visually distinct to ensure they grab the user's attention. One participant noted that the ability to dismiss or acknowledge warnings after reviewing them would be a nice feature.

Confidence Score and Low-Confidence Warnings

The confidence score was appreciated for the flexibility it provided, allowing them to adjust the strictness of classifications according to their needs. However, others found it unclear what a "low confidence" classification actually meant. One participant noted that they initially assumed a lighter highlight indicated a less important issue rather than a lower-confidence prediction. They suggested that the tool should provide a short explanation when users hover over a low-confidence classification, clarifying whether it signals uncertainty in the model's decision or inherent ambiguity in the sentence.

Users had difficulty understanding the confidence scores in general, and when it was explained by the interviewer, some became overly focused on it, seeking the confidence scores for all classifications as if they could derive additional meaning from them. One participant, a PhD candidate in explainable AI, pointed out that it might be better to let the researcher set the threshold and only display classifications that meet that threshold, rather than involving users in this process. This would help avoid overemphasis on the confidence scores, which are relative rather than absolute.

Additionally, some participants expressed a desire for greater control over how confidence scores influenced the display of results. One participant suggested that users should be able to set a threshold for low-confidence warnings, ensuring that only classifications above a certain confidence level were displayed. This would allow users to tailor the tool's behavior to their preferences and avoid being overwhelmed by unnecessary warnings.

Feature Requests and Suggested Improvements

Participants proposed several refinements that could enhance the tool's usability and effectiveness. One of the most common requests was improving explanations for flagged sentences. Instead of merely highlighting a sentence, participants wanted the tool to explicitly state which words or phrases contributed to the classification. This, they argued, would make it easier to understand the rationale behind each decision and facilitate quicker revisions.

Furthermore, two participants pointed out that the tool could be improved by checking for consistency across different sections of a paper. They noted that sometimes, a strong causal claim in the abstract was later hedged in the discussion section, creating potential inconsistencies. They suggested that the tool should highlight these mismatches, helping authors maintain coherence throughout their manuscript.

Use Cases and Integration into Research Workflow

Most participants envisioned using the tool at two key stages in their research process. First, they would use it before submitting a paper or the final version that you sent to your supervisor, to refine causal statements and ensure their writing was clear and rigorous. Second, they saw value in using it during literature reviews, as it could help them analyze causal claims in existing research and quickly see the study designs used. Additionally, some participants suggested that the tool could be useful for peer reviewers, helping them assess whether a study's causal claims were well-supported by its methodology. One participant noted that peer reviewers often lack time to check causal language in detail, and a tool like this could assist in flagging potential issues efficiently.

5

Discussion

This research aimed to evaluate the performance of large language models (BERT, GPT, and LLAMA) in two-level (RQ1.1) and three-level (RQ1.2) causal classification, distinguishing between correlational, conditional causal, and direct causal relationships. Additionally, we investigated how well these models generalize to social science texts (RQ1.3). The results show that BERT-based models significantly outperform generative models like GPT and LLAMA, particularly in distinguishing between correlational, conditional causal, and direct causal claims. While prompting techniques proved useful for extracting study design details, they were insufficient for causal classification, emphasizing the need for fine-tuned models.

This research also examines how researchers prefer to receive feedback on causal writing (RQ2.1), exploring different formats such as conversational LLMs, in-text annotations, warnings, or overviews. Additionally, we investigated how generative models like LLAMA 3.3 70B can be leveraged through prompt engineering to support causal inference writing, including tasks such as summarizing study designs or explaining causal classifications (RQ2.2). Finally, we analyzed whether researchers prioritize minimizing false positives (Type I errors) or false negatives (Type II errors) when evaluating causal statements, and the reasoning behind these preferences (RQ2.3). The qualitative study indicated that user feedback showed that AI-assisted tools for causal classification can improve research communication, though clearer confidence indicators and classification controls are necessary for usability and trust.

1. Key Findings

Addressing RQ1.1, our results show how different model architectures impact performance in two-level causal classification. The study's two-step framework separates causal from non-causal statements and further categorizes causal claims into correlational, conditional causal, and direct causal relationships. A key finding is that BERT-based models consistently outperformed generative models such as GPT and LLAMA, suggesting that model architecture plays a more significant role than scale in structured classification tasks. This aligns with prior research showing that bidirectional attention mechanisms, like those in BERT, are better suited for structured classification tasks compared to autoregressive models (Kim et al., 2023; Norouzi et al., 2024; Tan et al., 2023). While larger models may seem appealing, their weaker performance in this context suggests that scale alone does not guarantee superiority. Instead, model selection should be guided by performance metrics, resource constraints, and domain-specific requirements, rather than assumptions that bigger models inherently perform better. Additionally, human verification remains essential to mitigate potential biases and inaccuracies in automated causal classification (Van Dis et al., 2023).

The two-level classification task, which involved distinguishing between causal and non-causal statements, achieved an F1-score of 0.94, surpassing previous benchmarks such as prior work by Norouzi et al. and Tan et al. The blended dataset approach played a key role in this success, enabling the model to recognize causal statements across various subject areas. However, not all models per-

formed equally well as LLAMA and GPT models struggled to exceed an F1-score of 0.70. By integrating multiple datasets (both general purpose and scientific), the model was able to distinguish causal from non-causal statements with greater accuracy than previous approaches. This suggests that training on a variety of sources improves classification robustness, even in sources where causal markers are less explicit. These results show the importance of dataset composition in developing effective causal classification tools that can be applied across different research domains.

Addressing RQ1.2, the results reveal large differences in model performance across causal classification tasks. SciBERT achieved the highest macro F1-score (0.83), followed by RoBERTa (0.80), BERT (0.76), GPT-4o-mini (0.63), and LLAMA 3.2 (0.33). Although all models were fine-tuned on the same dataset, their pretraining influenced their ability to classify causal language, particularly in social science texts (RQ1.3). In general, the models performed worse on biomedical data than on social science sentences. However, if you look at the final model, SciBERT's most misclassifications happen in the biomedical sentences. One possible explanation is that biomedical data primarily consists of sentences from abstracts and press releases, where claims are carefully formulated. In contrast, social science data comes from full papers, where causal language is distributed throughout the text and often expressed in a more implicit or context-dependent manner. This makes causal classification more challenging in social science writing. However, for a classification tool to be truly useful, it must be able to detect causal claims throughout an entire paper rather than just in isolated sections. Given that social science writing is highly complex and nuanced (Healy, 2017), models must be capable of capturing these intricacies to ensure accurate classification. During the annotation process, frequent disagreements arose about whether a sentence should be categorized as correlational or causal, reflecting the complexity of real-world causal inference. If human annotators struggle with classification, current AI models will face similar limitations in differentiating between weak and strong causal relationships.

The confusion matrices reveal a recurring misclassification pattern, where direct causal statements are frequently misclassified as correlational and vice versa. This suggests that despite their conceptual distinctions, these categories share overlapping linguistic features that make them difficult for models to differentiate. An interesting secondary insight from these misclassifications is that if classification is incorrect, it could suggest that the original sentence was not clearly written enough to convey its intended causal meaning. This implies that causal classification tools could also assist researchers in improving the clarity of their causal statements, regardless of their accuracy.

Effectiveness of Prompting for Study Design Extraction

Addressing RQ2.2, this study examined how generative models like LLAMA 3.3 70B can support causal inference writing through prompt engineering. Specifically, the effectiveness of zero-shot and few-shot prompting in extracting key methodological details from research papers was evaluated. Prompts were used to have a LLAMA 3.3 70B model determine whether a study used random assignment of participants, identify potential confounding variables that could bias results, distinguish the different groups or conditions analyzed in the study, and describe how data was collected. The participants who reviewed these AI-generated study design summaries found them clear, useful, and accurate, helping them remain mindful of the study design. While this suggests that LLMs could assist in evidence synthesis, their most immediate benefit may be in helping researchers process and summarize large volumes of literature more efficiently, rather than fully automating systematic reviews or meta-analyses.

However, prompting alone was not sufficient for causal classification. While LLMs were effective at summarizing methodology, they failed to accurately classify causal claims in a structured manner without additional fine-tuning. This suggests that while autoregressive LLMs (GPT or LLAMA) are useful for text summarization tasks, structured causal classification is better suited for BERT-based models. This aligns with research showing that zero-shot and few-shot prompting have limitations in specialized classification tasks (Van Dis et al., 2023).

User Experience

To answer RQ2.3, which asks whether researchers prioritize minimizing false positives (Type I errors) or false negatives (Type II errors) when evaluating causal statements, the findings indicate that researchers tend to prioritize minimizing false positives over false negatives. Participants expressed a preference for seeing all classifications made by the tool—even those with low confidence—rather than risking missing any potential classifications. False negatives were seen as less harmful because

they result in missed insights, whereas false positives were viewed as more problematic, as they could lead to incorrect claims or over-extrapolation of findings, which could undermine the credibility of the research. This preference was reflected in participants' reactions to the confidence score slider. While some appreciated the flexibility it offered, others found it unclear and expressed confusion about what a "low confidence" classification really meant. One participant, a PhD candidate in explainable AI, suggested that it might be better to allow researchers to set a threshold for confidence so that only classifications above that level were displayed.

The findings for RQ2.3 suggest that researchers prefer lenient causal classification tools that display all identified statements rather than strict tools that filter out lower-confidence classifications. Participants prioritized minimizing false positives (Type I errors), as overstated causal claims could mislead readers and undermine research credibility. They still wanted access to all classifications, even those with lower confidence, rather than risk missing potential causal statements. This indicates that causal classification tools should prioritize transparency and provide researchers with control over classification strictness, rather than enforcing rigid filtering.

In answering RQ2.1, which examines how researchers prefer to receive feedback on causal writing, the study found that participants favored a combination of in-text annotations and overviews rather than an overload of warnings. Many appreciated the tool's ability to highlight causal claims directly in the text, as this helped them pinpoint areas needing revision. However, while warnings about causal overstatements were generally well received—especially in critical sections like the abstract and conclusion, participants often skimmed over the warnings because they felt too separate from the text rather than being integrated into the review process. Instead of appearing as a static element, participants wanted the warnings to be directly connected to the flagged text, making them easier to notice and act upon. Additionally, participants wanted more actionable feedback, preferring specific revision suggestions rather than generic alerts. One participant proposed linking warnings directly to flagged sentences, allowing for immediate context rather than requiring users to scroll through a separate panel. Others expressed interest in high-level summaries that provide an overview of causal writing patterns, which could help them reflect on broader trends in their text rather than focusing on individual flagged instances.

2. Implications

The present study introduces a specialized dataset[1] for causal language extraction in social science, an area that has been overlooked in existing computational approaches. While granular causal classification datasets have been developed for biomedical and health-related texts, no equivalent dataset has been curated for the social sciences. This research addresses this limitation by relabeling causal sentences from prior work (Norouzi et al., 2024) into more fine-grained categories: correlational, conditional causal, and direct causal, allowing for a more precise classification of causal statements within social science literature. By fine-tuning models on this structured dataset, the study expands the methodological toolkit for causal inference in the social sciences, enabling researchers to better distinguish between different types of causal claims.

Furthermore, the fine-tuned BERT model developed in this study outperformed existing benchmarks in accurately distinguishing causal from non-causal statements in social science texts. By moving beyond prior work that focused only on abstracts and titles, this study demonstrates the importance of classifying full-text manuscripts. Causal claims can appear anywhere in a paper: the introduction often reflects causal assumptions based on prior literature, hypotheses may suggest causal links in a tentative manner, and results or discussion sections may reinforce causal claims with supporting evidence. By capturing causality throughout an entire manuscript, this research offers broader applicability across academic disciplines.

Another important implication of this study is the role of automated causal classification tools in improving transparency and accountability in scientific communication. These tools can provide clear insights into how research findings are framed, helping both authors and reviewers ensure that claims are stated accurately. If integrated into academic publishing processes, such tools could improve the peer review process by identifying potential overstatements or inconsistencies between claims and study design.

[1] https://github.com/Tessavana/Causal-Clarity/blob/main/Datasetcreation/datasets/labeled_ssc_data.csv

This could be especially helpful for early-career researchers or those with less experience in academic writing, offering them guidance in presenting their causal claims more clearly and accurately. Furthermore, journalists writing articles about recent research findings could benefit from using such tools to verify which claims they are permitted to make in their reports, ensuring that their coverage reflects the research's true conclusions, as it has been found that average readers have a hard time distinguishing levels of causality (Adams et al., 2017).

3. Limitations

This study has several limitations to consider. First, the classification models operate at the sentence level, potentially missing causal claims that span multiple sentences or paragraphs (Yang et al., 2022). Many causal relationships in scientific writing are expressed across multiple sentences, requiring contextual understanding beyond isolated statements. Since identifying multi-sentence causality demands a different preprocessing approach, this study focuses only on single-sentence classification. Future work could explore document-level classification techniques to address this limitation.

Second, a limitation of the three-class annotated social science dataset is that inter-rater agreement was not formally assessed using Fleiss' Kappa index, meaning no Kappa score was calculated to measure annotation consistency. One of the authors categorized all sentences as correlational, conditional causal, or direct causal, with ambiguous cases reviewed in collaboration with Daniël Lakens. To enhance the dataset's reliability, incorporating an additional reviewer would be valuable. Another limitation is the imbalance in class distribution, with correlational statements underrepresented compared to conditional causal and direct causal statements. This imbalance negatively impacts model training, as models tend to favor majority classes, making it harder to accurately classify correlational statements. While undersampling can help balance the dataset, it reduces the total amount of training data, and oversampling increases the risk of overfitting.

Another limitation is the challenge of human annotation inconsistencies. As highlighted in the annotation process, there were sometimes disagreements about whether a sentence should be classified as correlational, conditional causal, or direct causal. This reflects the complexity of real-world causal inference, where even human experts struggle to agree on classifications. If human annotators find causal distinctions ambiguous, AI models will inevitably inherit some of these ambiguities.

Additionally, the study focuses on fine-tuning rather than zero-shot or few-shot learning approaches. While fine-tuning offers high accuracy, it requires large amounts of labeled data and computational resources. Few-shot learning, which allows models to generalize to new tasks with minimal labeled data, could be explored as an alternative for domains where extensive labeled datasets are unavailable. Given recent advancements in models like GPT-o3, DeepSeek, and Claude, which demonstrate improved reasoning abilities, future developments may further enhance causal classification. Perhaps, with continued progress, these models will eventually be capable of accurately distinguishing causal relationships.

Lastly, the user experience evaluation highlighted challenges with confidence scores and interpretability. Some users found the confidence scores ambiguous and were unsure how to interpret low-confidence classifications. Others suggested that AI-generated causal classifications could be more transparent in explaining why a sentence was classified in a particular way. Future work could improve AI interpretability by providing more intuitive explanations for classification decisions, such as highlighting key linguistic features that influenced the model's decision.

4. Future Work

Future work could focus on LLMs to identify who is making a causal claim and the context in which it is made. For example, the model could differentiate between claims made by the authors themselves and claims attributed to other studies or literature (e.g., in the introduction). This would ensure that causal statements are not only classified accurately but also understood within the appropriate context, whether the claim reflects the paper's findings, is based on previous research, or simply describes a procedure in the methods section. These improvements would provide better feedback on the use of causal language.

Additionally, expanding the dataset to include papers from various scientific fields, beyond just social science and biomedical literature, could improve the model's generalizability and performance. A more diverse dataset would allow the model to capture the different ways causal language is expressed across disciplines, making it a more robust tool for researchers across multiple fields.

Incorporating explainable AI methods, such as attention heatmaps, would also be a valuable next step. These methods could provide insight into why the model flags certain claims as causal, helping users better understand the reasoning behind its classifications. This transparency would build trust in the model, ensuring researchers can confidently use it to analyze their work, particularly when determining the correct use of causal claims.

Another direction for future research involves leveraging the available code to conduct large-scale metascience studies. By analyzing hundreds of papers, the models could extract key information, such as method summaries and causal claims from abstracts, enabling the study of trends and patterns across entire scientific domains. This could contribute to a better understanding of how causal claims are made and interpreted in the scientific literature.

6

Conclusion

Researchers today must balance subject expertise with methodological rigor, yet it is impossible to be an expert in everything. Cognitive constraints, combined with the pressure to publish, mean that researchers may not always have the time to carefully check whether their causal claims align with their study design and evidence. Scalable solutions are needed to support researchers in verifying their claims efficiently, ensuring their causal language accurately reflects their methodology and findings. This study demonstrates that fine-tuned LLMs can improve causal language in scientific writing by distinguishing between causal and non-causal statements and further classifying causal claims into correlational, conditional, and direct causal relationships. BERT-based models consistently outperformed generative models, reinforcing the importance of architecture selection over model scale. Beyond model performance, this research emphasizes the broader role of LLMs in improving research transparency. Automated tools for causal classification could enhance peer review, systematic reviews, and meta-analyses by identifying inconsistencies and reducing overstatements. By demonstrating how LLMs can provide a scalable solution for analyzing causal claims in scientific writing, this work advances ongoing efforts to improve research clarity and prevent misinterpretations. This study contributes to improving the use of causal language in research papers, ultimately supporting more precise scientific communication.

References

Adams, R. C., Sumner, P., Vivian-Griffiths, S., Barrington, A., Williams, A., Boivin, J., Bott, L., & Chambers, C. D. (2017). How readers understand causal and correlational expressions in news headlines: A replication study. *Public Understanding of Science*, *26*(8), 911–921.

AI, M. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*. https://arxiv.org/abs/2407.21783

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3615–3620.

Bott, L., Sumner, P., & Adams, R. C. (2019). Caveats in science-based communication: How hedging helps maintain accuracy without sacrificing engagement. *Science Communication*, *41*(5), 594–613. https://doi.org/10.1177/1075547019860648

Carcioppolo, N., Jensen, J. D., Lindke, J., Torres, N. J., & Fujita, K. (2021). Exaggerated and questioning clickbait headlines and their influence on media learning. *Psychology & Marketing*, *38*(6), 917–930. https://doi.org/10.1002/mar.21473

Caselli, T., Vossen, P., & Stevenson, S. (2017). Eventstoryline: Extending causal event annotations with implicit relations [Includes both explicit and implicit causal links between events, enhancing context richness]. *Proceedings of the 2017 Conference on Natural Language Processing*, 210–217.

Chambers, N., & Jurafsky, D. (2007). Altlex: A lexical resource for causality [Demonstrates the use of alternative lexicalizations to signal causal relations]. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 123–130.

Christiansen, S., Iverson, C., & Flanagin, A. (2020). *Ama manual of style: A guide for authors and editors (11th ed.)* Oxford University Press.

Cofield, S. S., Corona, R. V., & Allison, D. B. (2010). Use of causal language in observational studies of obesity and nutrition. *Obesity Facts*, *3*(6), 353–356.

Dahabreh, I. J., & Bibbins-Domingo, K. (2024). Causal inference about the effects of interventions from observational studies in medical journals. *JAMA*, *331*(21), 1845–1853. https://doi.org/10.1001/jama.2024.7741

Dettmers, T., Lewis, M., Shleifer, S., & Zettlemoyer, L. (2022). 8-bit optimizers via block-wise quantization [arXiv preprint. Retrieved from https://arxiv.org/abs/2110.02861].

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms [arXiv preprint. Retrieved from https://arxiv.org/abs/2305.14314].

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Dunietz, J., Burnham, G., Bharadwaj, A., Rambow, O., Chu-Carroll, J., & Ferrucci, D. (2020). To test machine comprehension, start by defining comprehension. *arXiv preprint arXiv:2005.01525*.

Frankenhuis, W. E., Panchanathan, K., & Smaldino, P. E. (2023). Strategic ambiguity in the social sciences. *Social Psychological Bulletin*, *18*(1), 1–25.

Gopalakrishna, G., Wicherts, J. M., Vink, G., Stoop, I., van den Akker, O. R., Ter Riet, G., & Bouter, L. M. (2022). Prevalence of responsible research practices among academics in the netherlands. *F1000Research*, *11*.

Grobid. (2008–2024).

Grosz, M. P., Rohrer, J. M., & Thoemmes, F. (2020). The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, *15*(5), 1243–1255. https://doi.org/10.1177/1745691620921521

Haber, N. A., Wieten, S. E., & Rohrer, J. M. (2022). Causal language and strength of inference in academic and media articles shared in social media (claims): A systematic review. *PLOS ONE*, *13*(5), e0196346. https://doi.org/10.1371/journal.pone.0196346

Healy, K. (2017). Fuck nuance. *Sociological Theory*, *35*(2), 118–127.

Hedström, P., & Ylikoski, P. (2010). Causal mechanisms in the social sciences. *Annual Review of Sociology*, *36*, 49–67.

Hendrickx, I., Kim, S., & Poibeau, T. (2010). Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals [Focuses on semantic relations including cause-effect, with a limitation to entity-to-entity relationships]. *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, 94–99.

Hernán, M. A., & Robins, J. M. (2020). *Causal inference: What if*. Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/

Irving, S. H., Flynn, T., & Ward, S. J. (2022). Statistical misinformation in media coverage: Challenges in undoing causal misperceptions. *Communication Research*, *49*(3), 495–518. https://doi.org/10.1177/0093650220931472

Kaggle. (n.d.). Kaggle: Your machine learning and data science community [Website. Retrieved from https://www.kaggle.com/].

Kazak, A. E. (2018). Journal article reporting standards.

Kim, Y., Guo, L., Yu, B., & Li, Y. (2023). Can chatgpt understand causal language in science claims? *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, 379–389.

König, L. M., Altenmüller, M. S., Fick, J., Crusius, J., Genschow, O., & Sauerland, M. (2023). How to communicate science to the public? recommendations for effective written communication derived from a systematic review. *Zeitschrift für Psychologie (in press)*. https://doi.org/10.31234/osf.io/cwbrs

Lipton, Z. C., Elkan, C., & Narayanaswamy, B. (2014). Thresholding classifiers to maximize f1 score [arXiv preprint. Retrieved from https://arxiv.org/abs/1402.1892].

List, A. (2024). The limits of reasoning: Students' evaluations of anecdotal, descriptive, correlational, and causal evidence. *The Journal of Experimental Education*, *92*(1), 1–31.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.

Michal, A., & Shah, S. (2024). A practical significance bias in laypeople's evaluation of scientific findings. *Public Understanding of Science*. https://doi.org/10.1177/09636625231234567

Mirza, P., Tonelli, S., & Schlangen, D. (2014). Causaltimebank: Annotating causal relations in news events [Emphasizes inter-sentence annotations for event-based causal reasoning]. *Proceedings of the Workshop on Temporal and Event Analysis*, 78–85.

Norouzi, R., Kleinberg, B., Vermunt, J., & van Lissa, C. (2024). Capturing causal claims: A text mining model for social science papers. *Tilburg University Working Papers*.

Nuijten, M. B., van Assen, M. A., Hartgerink, C., Epskamp, S., & Wicherts, J. M. (2017). The validity of the tool "statcheck" in discovering statistical reporting inconsistencies.

OpenAI. (2024). Gpt-4o mini: Advancing cost-efficient intelligence. *OpenAI Blog*. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library [Retrieved from https://arxiv.org/abs/1912.01703]. *Advances in Neural Information Processing Systems*, *32*, 8024–8035.

Schlegelmilch, R., et al. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716.

Seifert, C. M., & Hammond, L. (2022). The limits of reasoning: Students' evaluations of anecdotal, descriptive, correlational, and causal evidence. *Educational Psychology Review*, *34*(4), 1351–1372. https://doi.org/10.1007/s10648-021-09604-1

Spadaro, G., Tiddi, I., Columbus, S., Jin, S., Ten Teije, A., Team, C., & Balliet, D. (2022). The cooperation databank: Machine-readable science accelerates research synthesis. *Perspectives on Psychological Science*, *17*(5), 1472–1489.

Sumner, P., Vivian-Griffiths, S., Boivin, J., Williams, A., Bott, L., Chambers, C. D., & Adams, R. C. (2014). The association between exaggeration in health-related science news and academic press releases: Retrospective observational study. *BMJ*, *349*, g7015. https://doi.org/10.1136/bmj.g7015

Tan, Y., Zhang, J., & Li, Y. (2023). Unicausal: A universal dataset for causal language understanding. *Transactions of the Association for Computational Linguistics*, *11*, 1123–1139.

Thapa, D. K., Visentin, D. C., Hunt, G. E., Watson, R., & Cleary, M. L. (2020). Being honest with causal language in writing for publication.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., & Jegou, H. (2023). Llama: Open and efficient foundation language models [arXiv preprint. Retrieved from https://arxiv.org/abs/2302.13971].

Van Dis, E. A., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. H. (2023). Chatgpt: Five priorities for research. *Nature*, *614*(7947), 224–226.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., & Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45.

Wright, D., & Augenstein, I. (2021). Semi-supervised exaggeration detection of health science press releases. *arXiv preprint arXiv:2108.13493*.

Yang, J., Han, S. C., & Poon, J. (2022). A survey on extraction of causal relations from natural language text. *Knowledge and Information Systems*, *64*(5), 1161–1186. https://link.springer.com/10.1007/s10115-022-01665-w

Yu, B., Li, Y., & Wang, J. (2019). Detecting causal language use in science findings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4664–4674.

Causal Confusion

How LLMs Can Improve Causal Language in Research Communication

by

Tessa van Abkoude

In partial fulfillment of the requirements for the degree of Master of Science in Human-Technology Interaction

Scientific findings shape public understanding, policy, and healthcare decisions. Yet, the way research findings are communicated can lead to confusion, especially when vague or misleading language blurs the distinction between association and causation. For example, a statement that something is 'linked to' an outcome can easily be misinterpreted as 'causes,' leading to overstatements that misrepresent the strength of evidence. Large Language Models (LLMs) have the potential to improve how research findings are communicated, ensuring clarity and precision in causal claims. This study examines whether LLMs can effectively classify causal statements and help researchers convey their findings clearly

Student number: 1312715
Project duration: September 2, 2024 – February 18, 2025
Thesis committee: Dr. Daniël Lakens, TU Eindhoven
Dr. Chris Snijders, TU Eindhoven
Dr. Antal Haans TU Eindhoven

TU/e