

Quantifying the prevalence and impact of overreaching causal claims in social science

Corresponding Author: Mr Calvin Isch

Version 0:

Decision Letter:

22nd October 2025

Dear Mr Isch,

Thank you once again for your manuscript, entitled "Quantifying the prevalence and interpretive impact of causal claims within cross-sectional social science articles", and for your patience during the peer review process.

Your Article has now been evaluated by 3 referees. You will see from their comments copied below that, although they find your work of potential interest, they have raised quite substantial concerns. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version if you are willing and able to fully address reviewer and editorial concerns.

We hope you will find the referees' comments useful as you decide how to proceed. If you wish to submit a substantially revised manuscript, please bear in mind that we will be reluctant to approach the referees again in the absence of major revisions. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

To guide the scope of the revisions, the editors discuss the referee reports in detail within the team, including with the chief editor, with a view to (1) identifying key priorities that should be addressed in revision and (2) overruling referee requests that are deemed beyond the scope of the current study. We hope that you will find the prioritised set of referee points to be useful when revising your study. Please do not hesitate to get in touch if you would like to discuss these issues further.

Specifically, please address the following:

1. As all three reviewers suggest, the AI experiment part is timely and interesting, and your work would benefit from further development of it. Please respond to all comments/suggestions made by the reviewers on this aspect.
2. Reviewer #1 raised concerns regarding your position that cross-sectional studies cannot demonstrate causality. We believe that this should be carefully responded to and discussed.
3. Reviewers #2 and #3 both pointed out that the current work is Abstract based and that the conclusions might be overstated by omitting the full text information. Please address these concerns fully by providing finding-level evidence/analysis as Reviewer #3 suggests and discussing it as limitations.
4. Reviewer #3 made an important point by discouraging the use of "Silicon participants", we agree that the language should accurate, please revise your manuscript to describe the experiment in more appropriate language.

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration. To assist you in this process, I have attached a checklist that lists all of our requirements. If you have any questions about any of our policies or formatting, please don't hesitate to contact me.

If you wish to submit a suitably revised manuscript, we would hope to receive it within 4 months. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. When formatting this document, please respond to each reviewer comment individually, including the full text of the reviewer comment verbatim

followed by your response to the individual point. This response will be used by the editors to evaluate your revision and sent back to the reviewers along with the revised manuscript.

- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

Thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss the required revisions further.

Sincerely,



Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Science of science

Reviewer #2: Econometrics, experiments

Reviewer #3: NLP, Computational Social Science

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

Isch et al. present an impressive and panoramic diagnosis of the the extent and impact of employing causal claims in cross-sectional studies (studies without a quasi-experimental intervention, and where participants are not repeatedly sampled over time). They employ a human-in-the-loop AI workflow to label causal claims within papers, providing a descriptive snapshot of the landscape of causal claims. This is followed by experiments on both human and generative AI chatbots where a collection of abstracts with such causal claims are either flagged or modified, and the causal takeaways on participants recorded.

The paper is overall high quality and I recommend it for publication, conditioned on meeting four major points of feedback.

The first major point of feedback is that the authors do not expend enough effort to justify their assumption that causal claims cannot be drawn from cross-sectional studies. I actually disagree with the authors' cynicism on this point. The authors could surface their assumptions more, and directly articulate the polemical position that they are taking, rather assuming it as received community consensus. It seems sufficiently controversial to me that the authors are essentially positing that it is never valid to draw a causal claim from closing back-door paths in a DAG, but one should always look for "front-door paths" instead.

I fundamentally disagree with that pessimism for a few reasons. The authors articulate one reason in their Limitations section, which is that such an attitude would have delayed the scientific consensus that smoking causes lung cancer -- a *reductio ad absurdum* in my view. The second reason is that causal inference cannot be reduced to a set of mathematical operations that if followed yield "true causal effects." The statisticians insist on this, which is

why they always punt and say that the ultimate interpretation of causality remains beyond them, in the hands of the domain experts who draw the causal diagrams or pick the quasi-experimental interventions. Yet the authors are proposing a further punt -- they write in the Limitations section that the peer-reviewed literature should focus on the evidence, leaving the causal interpretations to be left to a mythical "decision-maker." That is a puzzling suggestion, since who is better equipped to interpret the strength of the causal evidence than the scientists in the peer review process? In the public domain, there would be any number of competing interests re-interpreting "evidence" to their advantage, and it seems improbable to me that the causal truth will emerge from post-publication public debate.

My belief is that a strong cross-sectional study with robustness tests *could* under certain circumstances provide a better "strict test" of a causal hypothesis (to use the language of Deborah Mayo) than a limited RCT that suffers from issues of generalizability.

Picking a completely arbitrary stimulus example from the OSF repository (which the authors make cleanly available), they find issue with two clauses in the abstract of "Breaking the Glass Ceiling: Local Gender-Based Earnings Inequality and Women's Belief in the American Dream" in the American Journal of Political Science (Newman, 2015). This is an article that I have not seen before, but at a glance, the cross-sectional design employed by Newman actually aggregates several polls across years, includes year fixed effects, and includes a thoughtful validation test that rules out alternative causal hypotheses. Even if it does not establish "gold standard" causal evidence, it also seems reasonable to me that a reader would update their priors about the causal mechanisms behind women's beliefs in meritocracy after reading this paper. The use of the words "explore" and "should" in the abstract seem like appropriate hedges, and I do not see this example as representative of problems in the scientific literature. I did not look through any other stimuli, since it seems onerous for this peer review to have to involve reviewing dozens of other papers...

That brings me to my second major point of feedback, which is that the paper overstates the importance of its findings. It isn't clear to me whether participants "infer causality" after reading the original vs. rewritten or warning-labeled abstracts, or whether they are simply referring to the language in the abstract to summarize and answer questions. But what is the significance of that if the reader does not internalize the actual definition of causality -- which is that the outcome would change under a hypothetical intervention on the treatment? Indeed, the second experiment falls short of showing that participants internalized any interventional takeaways from these tweaks of language.

I found the argument on L168 that attempted to salvage the result by claiming the effects were in the same direction to be less rigorous than the rest of the article, since it was not pre-registered nor labeled as such.

The third point is that the experiments on Generative AI have some under-developed potential, and could be potentially extended to make the paper more interesting. Just as probabilistic statements in scientific articles become more certain and less hedging as they are passed from journals to reporters [1], perhaps there is a way that the conditional causal statements carefully crafted by authors (which I don't see as a problem) become unconditional as they are summarized by machines, since those summaries are increasingly presented in search engines and within the chatbots themselves. This would increase the stakes of the empirical phenomena described by the authors.

The fourth point is that I would have liked to see more analysis of the "direct", "implied" and "conditional" throughout both the

descriptive analyses and the experiment, since to me, the distinctions seem quite interesting and important.

Minor notes: - The term "silicon participant" leads to an anthropomorphization of Generative AI. I suggest rewriting to avoid the word "participant" in that context (and ideally "silicon" too, since a system like GPT is far more than just "silicon" for all kinds of reasons). Something more directly accurate would be preferred, like "synthetic experiments using GPT".

References: [1] <https://aclanthology.org/2022.emnlp-main.117.pdf>

Reviewer #1 (Remarks on code availability):

I did not attempt to run the code, but it looks reasonably organized and replicable. They are simple Jupyter Python notebooks and the figures and statistics from the paper are clearly generated from these notebooks.

Reviewer #2 (Remarks to the Author):

Summary

The paper studies the prevalence and consequences of causal language in cross-sectional studies, a design that generally does not allow for causal inference. The authors document that the use of causal language has increased over time, with a notable rise around 2000 coinciding with the "causal revolution" in economics and other social sciences. They show that both human and AI ("silicon") participants classify studies as causal when such language is used, and that interventions can mitigate this tendency. The study suggests that causal language can shape perceptions of research credibility and interpretation.

General Appreciation

The paper addresses an important and timely issue. The use of causal language in non-causal research designs is widespread and has implications for how research is read, cited, and used in policy. The documentation of temporal trends and the experiments with readers are valuable contributions. At the same time, the paper could be strengthened by providing further contextualization (e.g., comparing cross-sectional studies to true experimental studies), by clarifying some methodological details (statistical tests, reporting practices), and by elaborating on the motivation for using AI participants.

Comments

1. Benchmarking prevalence – While the study carefully documents the prevalence of causal language in cross-sectional studies, it would be helpful to provide a benchmark from experimental or quasi-experimental papers. Such a comparison would help readers assess whether the prevalence in cross-sectional work is unusually high, or in line with broader research practice.
2. Role of incorrect causal statements – The discussion could acknowledge more explicitly that even incorrect or overstated causal claims can sometimes stimulate valuable research progress. For instance, the Cowles Foundation report on research in economics triggered substantial improvements in methods and standards. Highlighting this aspect could bring nuance to the normative perspective.
3. Abstract-based experiments – The experiments based on abstracts are insightful, yet they may overstate the real-world impact since information in the full text can moderate impressions created by the abstract or title. The paper already documents cases where causal language appears later in the text; it may be useful to also note the possibility of the reverse. The intervention results partly address this issue, but a more explicit discussion of these limitations would strengthen the interpretation.
4. Use of "silicon participants" – The inclusion of AI participants is an innovative idea, though the added value could be explained in more detail. The results show higher compliance by silicon agents, which may reflect model parameters such as "Temperature" and "Presence Penalty." Clarifying this point and discussing how such choices shape outcomes would increase transparency. One possibility would be to also assess silicon agents in settings with causal designs (e.g., randomized experiments or natural experiments). Contrasting outcomes in causal versus non-causal contexts could illustrate more clearly what these agents are capturing.
5. Figure 3 / time trends – It would be helpful to clarify whether the evidence in Figure 3 accounts for time effects, for example by ensuring that comparisons are made across journals within the same time period.
6. SM: One-tailed tests (p.15) – Please provide a brief justification for the choice of one-tailed rather than two-tailed tests.
7. SM: Transparency of results – The paper and supplementary material mention multiple-testing corrections, but standard errors and uncorrected p-values are not reported. Providing these would allow readers to better assess the robustness of the findings and verify the implementation of corrections.

Reviewer #3 (Remarks to the Author):

Thanks a lot for submitting this interesting article which clearly tackles an important problem.

Overall strength:

- (1) important and timely topic, and I think studying causal claims in cross-sectional social science articles is neat
- (2) using LLM + fine-tuned model to identify causal claims is reasonable and the model could be potentially used by other researchers
- (3) nice study on audience perception of causal claims

Key weaknesses and suggestions

(1) While studying cross-sectional social science articles is interesting, it is unclear how much these findings can be generalized to other types of articles. It would be meaningful to discuss the generalizability of the findings

(2) One key weakness of the study is that it focuses on abstracts instead of individual findings. A paper might have multiple conclusions and it's possible that they mostly used correlational language but use causal words in some. It would be nice to conduct finding-level analysis to understand whether specific scientific findings are presented in a causal way. Furthermore, I'm not sure if the current methodology could support the claim that "full texts are substantially more likely to contain causal language". It would be necessary to compare the same findings presented in the paper abstracts and the main text to better understand the dynamics behind this. There is a paper (<https://aclanthology.org/2022.emnlp-main.117.pdf>) with an associated package that allows you to easily identify pairs of scientific findings: <https://github.com/copenlu/scientific-information-change>. So basically, with clusters of findings presented in different sections, it should be fairly straightforward to study whether the authors convert a correlational claims to a causal claim in certain sections.

(3) the types of causal claims would also be an interesting thing to study. For example, would the authors hedge more when making causal claims? It would be nice to analyze the uncertainty of the claims and see whether the authors use hedges as a strategy to make causal claims for correlational findings. This also requires analysis at the finding level.

(4) The language the authors used to describe GPT evaluation results is problematic and can be very misleading. For example, "GPT was already hesitant to say" has a serious anthropomorphism issue. I also STRONGLY discourage the authors from using "Silicon participants" in a serious research article, as it may create wrong perceptions that AI can actually be treated as research subjects like human participants. Using AI to conduct experiments is fine, but it must be described in an appropriate way.

(5) I do like the GPT summary analysis as it provides additional evidence on how these inappropriate causal claims might become a bigger problem when everyone uses AI to access information. Please provide more details on how you use GPT to generate the summaries of the paper. To me, how AI models interpret the causal claims seems less exciting. It would be great if you could provide further empirical evidence on how the causality of the findings affects AI summaries. For example, if you provide 5 findings to AI, will AI be more likely to pick the causal one? This will further strengthen the implications of your findings.

Overall I think this is an interesting paper and congratulations on finding this neat angle to study the overuse of causal claims!

Version 1:

Decision Letter:

20th March 2026

Dear Mr Isch,

Thank you once again for your revised manuscript, entitled "Quantifying the prevalence and impact of overreaching causal claims in social science," and for your patience during the re-review process. I am communicating this decision to you instead of Dr Jiang, as she is currently out of the office.

Your manuscript has now been evaluated by the same reviewers who evaluated your original manuscript. All reviewer

feedback is included at the end of this letter. Although the reviewers found your manuscript to have improved during revision, they also raise some important outstanding concerns. We remain interested in the possibility of publishing your study in Nature Human Behaviour, but would like to consider your response to these outstanding concerns in the form of a revised manuscript before we make a decision on publication.

In your revision, please address all remaining concerns raised by Reviewers #1 and #3. Among the reviewer comments, we highlight the need to demonstrate generalization by using other, more recent LLMs, especially because GPT 4.1 is now outdated (Reviewer 3).

Finally, your revised manuscript must comply fully with our editorial policies and formatting requirements. Failure to do so will result in your manuscript being returned to you, which will delay its consideration.

In sum, we invite you to revise your manuscript taking into account all reviewer and editor comments. We are committed to providing a fair and constructive peer-review process. Do not hesitate to contact us if there are specific requests from the reviewers that you believe are technically impossible or unlikely to yield a meaningful outcome.

We hope to receive your revised manuscript within 4-8 weeks. I would be grateful if you could contact us as soon as possible if you foresee difficulties with meeting this target resubmission date.

With your revision, please:

- Include a "Response to the editors and reviewers" document detailing, point-by-point, how you addressed each editor and referee comment. If no action was taken to address a point, you must provide a compelling argument. This response will be used by the editors and reviewers to evaluate your revision.
- Highlight all changes made to your manuscript or provide us with a version that tracks changes.

Please use the link below to submit your revised manuscript and related files:

Link Redacted

Note: This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We look forward to seeing the revised manuscript and thank you for the opportunity to review your work. Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

Sincerely,

[Redacted Signature]

Nature Human Behaviour

on behalf of

[Redacted Signature]

Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Science of science

Reviewer #2: Econometrics, experiments

Reviewer #3: NLP, Computational Social Science

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The paper is much improved, and I only have a few additional minor suggestions, mostly about narrative and presentation:

1. The figures are not colorblind accessible; the bars should have fill patterns (especially when it is just red and green), and the lines should have line styles.
2. Have the authors run their own detection method on this manuscript? The authors should report whether the tool flagged any causal claims that were based on their cross-sectional analysis. This could make for an interesting discussion caveat, if it turns out to be hard to avoid slipping into a causal-implying mode, even for the authors.
3. I don't like the Stroop task analogy. It implies a particular mechanism (cognitive load) that isn't actually born out by any evidence here. It could be that authors introduce causal language deliberately and strategically, not because they're tired. (Proving once again that narrative license is hard to avoid!)
4. The Journal of Quantitative Description example is really interesting. To me this brings up the key counterexample to the whole paper, which is that I find it completely plausible for there to be a causal effect between the journal's policies and the extent of causal language used in its pages, and for this to be absolutely admissible evidence in favor of a policy intervention (that journals should restrict causal language). Don't you? I think this issue deserves a more prominent discussion.
5. Relatedly, there is no magic bullet technique that grants someone narrative license, not even RCTs. I see Yarkoni cited in the introduction, but it would be useful to revisit that point in the discussion along with something like Deaton & Cartwright (2018): <https://www.sciencedirect.com/science/article/pii/S0277953617307359>.
6. "Prior work on statistical misperceptions [80] has found that experts are no less susceptible than lay participants" overstates that particular finding. Experts are _also_ susceptible, but they may be (and one certainly hopes they are) less susceptible than lay participants.

Conditioned on addressing these points, I am happy to recommend this manuscript for publication.

Reviewer #2 (Remarks to the Author):

The revised version of the paper addresses all my comments from the previous review round.

Reviewer #3 (Remarks to the Author):

Thank you so much for revising the manuscript and the updated manuscript looks great!

Below are some suggestions and comments

1. For AI experiments, while I do believe the results on GPT 4.1 could be generalized to other models, it would be nice to add more LLMs since GPT 4.1 is only one (and outdated) model. The authors should use the latest LLMs when preparing the final version. This will help strengthen the paper, especially when it is being discussed by the public.
2. Thanks a lot for the effort on studying the connections between uncertainty and causality expressions. While I did suggest this in earlier reviews, I find the current results a bit confusing and does not add too much to the overall narrative. One way to improve it is to directly look at claim-level uncertainty and causal expressions. Please consider using <https://pypi.org/project/certainty-estimator/> when assessing the certainty of scientific claims, as it is a fine-tuned model designed specifically for measuring the uncertainty of findings. It could also help to scale up the experiment on uncertainty and causality, and it could actually be a fairly straightforward experiment, given that you already have the model for causal expressions.

Please also consider adding more references regarding hedges and uncertainty expressions in science communication:

Hyland, Ken. "Writing without conviction? Hedging in science research articles." *Applied linguistics* 17.4 (1996): 433-454.

Friedman, Sharon M., Sharon Dunwoody, and Carol L. Rogers. *Communicating uncertainty: Media coverage of new and controversial science*. Routledge, 2012.

Gustafson, Abel, and Ronald E. Rice. "The effects of uncertainty frames in three science communication topics." *Science Communication* 41.6 (2019): 679-706.

While I do think a refined analysis on this could help, I also feel it is okay to leave this out of the main text to make a coherent narrative.

3. I'm not sure if Figure 6 does the best job of presenting the findings on LLMs' summarization. It might be helpful to have a river plot highlighting how many correlational claims of a paper got translated into causal claims through LLM summarization. It would also help to add a more illustrative example highlighting how LLM summarization pick/present causal claims. Another idea would be ask LLM to do deep research on some controlled topics and see how it picks and

summarizes the results. I really think the LLM summarization experiment presents a potentially very impactful finding when so many people are using AI to do deep research (basically a deep summarization of literature). And I would really hope the authors could further strengthen this result.

Again, I would like to highlight that I really like this paper and I would love to see it being published. I hope the authors could consider my suggestions and further strengthen it to generate a larger impact in the AI world.

Version 2:

Decision Letter:

Our ref: NATHUMBEHAV-25083731B

11th May 2026

Dear Dr. Isch,

Thank you for submitting your revised manuscript "Quantifying the prevalence and impact of overreaching causal claims in social science" (NATHUMBEHAV-25083731B). It has now been seen by the original Reviewer #3 and their comments are below. As you can see, the reviewer finds that the paper has improved in revision. We will therefore be happy in principle to publish it in Nature Human Behaviour, pending minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements within two weeks. Please do not upload the final materials and make any revisions until you receive this additional information from us.

Please do not hesitate to contact me if you have any questions.

Sincerely,

A black rectangular box redacting the signature of the editor.

Nature Human Behaviour

Reviewer #3 (Remarks to the Author):

Thanks a lot for revising it! I have no further questions and congratulations on this great work!

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Reviewer expertise:

Reviewer #1: Science of science

Reviewer #2: Econometrics, experiments

Reviewer #3: NLP, Computational Social Science

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

Isch et al. present an impressive and panoramic diagnosis of the the extent and impact of employing causal claims in cross-sectional studies (studies without a quasi-experimental intervention, and where participants are not repeatedly sampled over time). They employ a human-in-the-loop AI workflow to label causal claims within papers, providing a descriptive snapshot of the landscape of causal claims. This is followed by experiments on both human and generative AI chatbots where a collection of abstracts with such causal claims are either flagged or modified, and the causal takeaways on participants recorded.

The paper is overall high quality and I recommend it for publication, conditioned on meeting four major points of feedback.

Response: We thank the reviewer for carefully considering the paper, and the useful advice for improving its quality. In the revised version we have made a number of changes that we hope address the concerns raised.

The first major point of feedback is that the authors do not expend enough effort to justify their assumption that causal claims cannot be drawn from cross-sectional studies. I actually disagree with the authors' cynicism on this point. The authors could surface their assumptions more, and directly articulate the polemical position that they are taking, rather assuming it as received community consensus. It seems sufficiently controversial to me that the authors are essentially positing that it is never valid to draw a causal claim from closing back-door paths in a DAG, but one should always look for "front-door paths" instead.

I fundamentally disagree with that pessimism for a few reasons. The authors articulate one reason in their Limitations section,

which is that such an attitude would have delayed the scientific consensus that smoking causes lung cancer -- a *reductio ad absurdum* in my view. The second reason is that causal inference cannot be reduced to a set of mathematical operations that if followed yield "true causal effects." The statisticians insist on this, which is why they always punt and say that the ultimate interpretation of causality remains beyond them, in the hands of the domain experts who draw the causal diagrams or pick the quasi-experimental interventions. Yet the authors are proposing a further punt -- they write in the Limitations section that the peer-reviewed literature should focus on the evidence, leaving the causal interpretations to be left to a mythical "decision-maker." That is a puzzling suggestion, since who is better equipped to interpret the strength of the causal evidence than the scientists in the peer review process? In the public domain, there would be any number of competing interests re-interpreting "evidence" to their advantage, and it seems improbable to me that the causal truth will emerge from post-publication public debate.

My belief is that a strong cross-sectional study with robustness tests *could* under certain circumstances provide a better "strict test" of a causal hypothesis (to use the language of Deborah Mayo) than a limited RCT that suffers from issues of generalizability.

Picking a completely arbitrary stimulus example from the OSF repository (which the authors make cleanly available), they find issue with two clauses in the abstract of "Breaking the Glass Ceiling: Local Gender-Based Earnings Inequality and Women's Belief in the American Dream" in the American Journal of Political Science (Newman, 2015). This is an article that I have not seen before, but at a glance, the cross-sectional design employed by Newman actually aggregates several polls across years, includes year fixed effects, and includes a thoughtful validation test that rules out alternative causal hypotheses. Even if it does not establish "gold standard" causal evidence, it also seems reasonable to me that a reader would update their priors about the causal mechanisms behind women's beliefs in meritocracy after reading this paper. The use of the words "explore" and "should" in the abstract seem like appropriate hedges, and I do not see this example as representative of problems in the scientific literature. I did not look through any other stimuli, since it seems onerous for this peer review to have to involve reviewing dozens of other papers...

Response 6: We appreciate this thoughtful insight, and have substantially revised our paper in response.

First, we appreciate the reviewer's point on the nuance of causal claims from cross-sectional studies using backdoor adjustment. We agree that it is too strong to assert that "it is never valid to draw a causal claim from closing back-door paths in a DAG"; such inference is possible in theory if all relevant confounders are measured and properly adjusted for and if it can be established that changes in the independent variable occur prior to changes in the dependent variable. In practice, however, especially in cross-sectional, non-experimental settings that lack repeated measures, these assumptions are rarely testable and are often likely to be violated. Therefore, our study design is not grounded in the absolute impossibility of causal inference from these designs, but rather in the pragmatic assessment that their evidentiary strength is weak compared to designs leveraging exogenous variation (e.g., IV, DiD, RD) or longitudinal designs (e.g., Granger causality), which we exclude using our first classifier. Our classification strategy thus focuses on studies where causal claims rest solely on assumptions of conditional independence in observed covariates and about temporality—a widely criticized and frequently insufficient identification strategy. This focus on method is primarily a pragmatic choice, allowing scalability and interpretability; however, deeper qualitative insights into specific studies are also useful, as we will demonstrate below.

Additionally, our causal language classifier was not designed to target claims about correlational findings framed as consistent with theory. We believe such language can rightly indicate theoretical plausibility rather than data-based causal evidence. For example, the claim "We found that social media use was correlated with depression. This aligns with models suggesting social platforms impact mental-well-being." is coded as non-causal by our classifier (It is possible that some such statements will be captured within the "Conditional" or "Implied" causal categories, but our review of the labeling suggests that such cases constitute only a small minority, and in a small, tailored sample of several statements like this, the large majority were classified as non-causal.). Under our design restrictions, we argue that causal interpretation should be framed like this example: as associational findings, consistent with theory, but contingent on assumptions (i.e., clearly stating the design); not as a causal finding, the type we sought to classify. We now clarify this position in the manuscript and reference canonical sources defining the conditions under which causal inference from observational designs may be credible.

Second, we agree that "a strong cross-sectional study with robustness tests could under certain circumstances provide" a better test of a hypothesis than a poorly executed RCT. Our emphasis on the limitations of cross-sectional designs may have unintentionally suggested that we privilege experimental designs across the board; however, we share the concern that execution problems in RCTs can undermine inference, particularly for generalizability. We have expanded the discussion to clarify this important point.

Third, we agree that our reference to leaving interpretations to a "mythical decision-maker" was unhelpful, and we have removed it. Our intended recommendation is to encourage clearer articulation of methodological choices, assumptions, measurement, and uncertainty, a point we now make with greater care.

Fourth, we appreciate the opportunity to clarify the smoking/Fisher example and the AJPS example highlighted. In the case of smoking, our understanding is that cross-sectional studies did indeed supply important evidence in favor of the causal effects of smoking. We note, however, that the eventual causal consensus did not derive from cross-sectional studies alone, but from triangulation across prospective cohorts, longitudinal evidence, mechanistic toxicology, biomarker data, and animal models. Cross-sectional comparisons of smokers vs. non-smokers were only a part of this broader evidentiary base. Thus, although we agree with (and appreciate) the point that we also must also be careful and

precise in our use of language, we do not believe that the smoking example refutes our position that cross-sectional, nonexperimental data does not, on its own, constitute direct evidence for causal claims.

Regarding the AJPS article, we agree that its main causal claim that gender gaps in pay cause women to be less optimistic about the American Dream seems entirely plausible. We also agree that its use of causal language reflects our own impression of what is typical of quantitative social science papers, including those published in top venues, and is in many places (particularly in the abstract) appropriately hedged. Indeed, we did not choose it because we thought it was especially egregious; rather, it was one of 28 papers drawn from a stratified random sample of elite journals (we retained all papers containing any causal claim to increase external validity for our experiment). Nonetheless, we continue to think that the paper does communicate to the reader that the evidence contained in the paper is itself causal when it should be framed as correlational, though aligning with their causal theory. We think this impression is actually stronger for the paper read as a whole than it is for any individual sentence; however, there are numerous examples of individual sentences that make strong causal claims about the results. For example, on p. 1022 the author concludes “The findings presented here reveal that not only are women’s beliefs in the American Dream responsive to local variation in the economic resources of women, they are highly responsive to the resources of women relative to men.” Importantly, this sentence does not say that the evidence presented, although correlational, is consistent with other causal evidence or that together with other evidence it presents a plausible picture of causality—it plainly asserts that the evidence in the paper itself is causal. According to the classification scheme detailed above, this claim would be coded as overreach. (side note: although the author does include fixed effects for the year in which each survey was conducted, the analysis is not in any sense longitudinal; i.e. the author is not estimating women’s stated beliefs in year T conditional on economic inequality in previous years. Thus, although the data was collected over time, the analysis treats all data as simultaneous, hence its rating as cross-sectional.)

Again, it is not our contention that the causal claims in this article are necessarily wrong or that the data presented do not offer any evidence in support of them. To the contrary, the claims are highly plausible and the data appear to be generally consistent with them. A reasonable conclusion of the analysis, therefore, might be that it should somewhat increase the confidence that one might already have had in the conclusions based on theoretical arguments or other empirical evidence. Returning to our central claim, however, the paper goes well beyond this “reasonable” conclusion, instead making direct and in some places quite strong causal claims on behalf of the evidence presented—evidence that, on its own, does not justify such claims (another example from p. 1014: “As can be seen, women’s belief in meritocracy is significantly influenced by local variation in the gender pay gap, but men’s attitudes are not.”). For this reason, we think the paper is an appropriate example of the misuse of causal language that we wish to highlight—a misuse that is especially likely to pass undetected when the claims are plausible. We are grateful to the reviewer for bringing this paper to our attention and forcing us to articulate our argument more carefully.

In addition to this paper, we have also reviewed other examples in our dataset and reaffirmed that they also contain the types of causal claims that are illustrated in the random sample shown in Table 1. We understand that reviewing many examples would be onerous, but we invite the reviewer to look at additional cases if they remain skeptical (e.g., from the 100 articles used in our new GPT-4.1 experiment; see **Response 23**). Our measurement strategy is not perfect, and some weak examples and even false positives are inevitable. Yet our extensive qualitative review and empirical validation suggest that the models reliably identify the causal claims at scale, and most examples exhibit strong, unhedged, and insufficiently qualified causal language (as summarized in **Response 9**).

In sum, we have softened the language that implied a universal scholarly agreement with our assumptions, stated our scope more precisely, pointed to alternative perspectives in causal inference, and retained our core claim. Here are several places where we modified our text:

In the *abstract*, we changed the following lines:

- Previous: "...scholarly articles often employ cross-sectional designs that reveal correlations but cannot establish causality"
- Modified to: "...cross-sectional designs that reveal associations but are generally unable to support direct causal claims."
- Previously: "Experiments with both AI and human participants showed that audiences frequently infer causal conclusions from associational studies, but methodological labels and cautious phrasing can reduce such overinterpretation."
- Modified to "To examine the effects of such language, we conducted a human-subjects experiment, finding that readers frequently indicate abstracts with this phrasing provide causal evidence but that methodological labels and associational wording reduce this tendency."

In the *main text*, we modified our description of the cross-sectional sample.

- Previously Paragraph beginning on line 40: "Here, we focus on single-wave, non-intervention, cross-sectional studies..."
- Now: "To simplify the problem, we focus on empirical articles using cross-sectional, nonexperimental designs, excluding those with longitudinal components. By definition, such designs cannot establish the temporal ordering of cause and effect (i.e., Bradford Hill's temporality criterion^{25, 26}), and nonexperimental data, lacking randomization, is generally susceptible to confounding variables and reverse causality. We recognize, of course, that quasi-experimental methods such as instrumental variables and regression discontinuity designs allow for causal conclusions to be drawn from observational snapshots when investigators make strong, explicit assumptions and articulate a credible identification strategy^{27–30}. We therefore exclude designs of this sort, such that the remaining corpus of studies rely on purely cross-sectional variation, rendering their causal claims sensitive to untestable identification assumptions. We further note that associational evidence, even if not causal on its own, could reasonably be interpreted as adding support to causal claims derived from theory, models, or other empirical evidence. For this reason, we designed a classifier to code causal claims as overreaching only when they were made about the data under consideration, not targeting claims presented as correlational, though "aligning" with causal theories. We note finally that this rule-based system for identifying causal claims supports interpretable, large-scale quantification, including assessment of longitudinal trends."
- We added: "We note, however, that conditional claims are relatively rare: of the articles classified as causal by human coders, only 3% contained only conditional claims while the remaining 97% contained either direct (54%) or implied (43%) claims (where articles contained more than one type of causal claims we classified them by the strongest form present)."

- Previously: "...cross-sectional designs are not suited to establishing causality..."
- Now: "As we have just argued, cross-sectional designs are poorly suited to support such claims..."

To address the concern about the "mythical decision maker" we changed the following:

- We removed: "Decision-makers will inevitably need to act under conditions of uncertainty, often drawing on correlational findings; however, such judgements are best made outside the confines of peer-reviewed literature."
- Previously: "While this concern is valid, we maintain that causal claims should not be made within scholarly papers themselves, unless the research methods can support such a link."
- Now: "Third, in this paper we treated all causal claims about novel empirical results from cross-sectional studies as overreach. However, some such correlations do reflect genuine causal relationships, and ignoring these causal signals can also be harmful, as illustrated by Ronald Fisher's skepticism about the smoking–cancer link, which was rooted in the observational nature of the evidence^{74, 75}...Although we continue to argue that caution when making causal claims is warranted in science communication, future work should consider how to balance the harms caused by careless rhetoric against the harms caused by excessive skepticism."

Finally, we added the following paragraph to our discussion in response to the reviewer's concern about the generalizability of experiments, which we share:

- "Second, our focus on cross-sectional studies, although helpful for creating clear criteria for identifying instances of causal overreach, does not imply that other designs are less susceptible to narrative license. For example, carefully executed cross-sectional evidence can, in some settings, be more informative for theory testing than, say, laboratory experiments with limited external validity. Our aim is not to criticize any single design or to single out any one manifestation of overclaiming, but rather to call attention to the broader phenomenon of narrative license and to introduce methods for identifying it. Future work should therefore examine other forms of narrative license within other research designs, including (quasi-)experimental designs, simulation studies, mathematical modeling, and ethnographic research. If systemic pressures such as funders' emphasis on "impact" or publication incentives are driving these trends, we would expect parallel increases in other modes of overstatement cataloged in Table 2.

We appreciate the reviewer's prompt to clarify our argument and for raising such interesting and informative examples.

That brings me to my second major point of feedback, which is that the paper overstates the importance of its findings. It isn't clear to me whether participants "infer causality" after reading the original vs. rewritten or warning-labeled abstracts, or whether they are simply referring to the language in the abstract to summarize and answer questions. But what is the significance of that if the reader does not internalize the actual definition of causality --

which is that the outcome would change under a hypothetical intervention on the treatment? Indeed, the second experiment falls short of showing that participants internalized any interventional takeaways from these tweaks of language.

I found the argument on L168 that attempted to salvage the result by claiming the effects were in the same direction to be less rigorous than the rest of the article, since it was not pre-registered nor labeled as such.

Response 7: We appreciate this concern about potential overstatement of our findings and agree that our original wording did not clearly distinguish between participants endorsing a causal interpretation and internalizing an interventionist (counterfactual) definition of causality. We have revised the manuscript to describe our outcomes with greater precision: our primary measure assesses whether participants judged a study as providing causal evidence, and our secondary measures code causal phrasing in participant summaries and implied causal inference items akin to the hypothetical intervention suggested. While the latter outcomes better reflect an interventionist understanding of causality, we believe that the primary outcome is informative in its own right. If the phrasing of results changes whether readers state that a study provides causal evidence, that represents a meaningful shift in interpretation, even if it does not immediately alter their underlying causal theory. For this reason, and to stay consistent with our pre-registration, we keep this item as our primary outcome, though we change the language throughout the text to more clearly describe what was measured.

In the *abstract*:

- Previously: “Experiments with both AI and human participants showed that audiences frequently infer causal conclusions from associational studies, but methodological labels and cautious phrasing can reduce such overinterpretation.”
- Modified to “To examine the effects of such language, we conducted a human-subjects experiment, finding that readers frequently indicate abstracts with this phrasing provide causal evidence but that methodological labels and associational wording reduce this tendency. Experiments with GPT-4.1 revealed that its summaries of these articles can amplify causal overstatement, removing hedges and introducing causal claims where articles used strictly associational phrasing”

In the *full text*, we made several edits to our language that align with these examples:

- We modified our Research Question and Hypotheses to reflect this more precise language, while aligning with our pre-registered versions.
 - “RQ: To what extent do participants judge that a study provides causal evidence after reading a cross-sectional abstract that uses causal language to describe its results?”
 - “H1-H3: Each of the three interventions—a rewritten abstract using purely associational language, a methodological note clarifying study design and limitations, and “AI” feedback highlighting interpretive language alongside a similar note—will reduce the amount that participants report a study provides causal evidence.”

- Previously: “Exposure to causal wording in the abstracts markedly increased participants’ tendency to infer causality.”
- Now: “First, Fig. 5a shows that in the original abstract condition, participants tended to agree that the studies provided causal evidence...In relative terms, participants who read a rewritten version with only associational language showed lower agreement...”

Additionally, we agree that our exploratory analyses were less informative regarding the second outcome, which was meant to target more subtle or implied beliefs about causality through a hypothetical manipulation of the IV. We thus removed the following:

- We removed “...an exploratory analysis using Fisher's combined-probability test indicated joint significance across the outcomes ($p = 0.022$). Thus, this result may reflect smaller effects (β 's range from -0.12 to -0.18) than our experiment was powered to identify.”

The results now state that there was “no significant difference observed across conditions” for that outcome.

The third point is that the experiments on Generative AI have some under-developed potential, and could be potentially extended to make the paper more interesting. Just as probabilistic statements in scientific articles become more certain and less hedging as they are passed from journals to reporters [1], perhaps there is a way that the _conditional_ causal statements carefully crafted by authors (which I don't see as a problem) become unconditional as they are summarized by machines, since those summaries are increasingly presented in search engines and within the chatbots themselves. This would increase the stakes of the empirical phenomena described by the authors.

Response 8: We appreciate this suggestion, which echoes points from Reviewers #2 and #3. We have now added an extensive new experiment with GPT-4.1 to test precisely this mechanism, whether conditional claims are rendered unconditional by model summarization, as well as several related questions.

The full set of results are available in the revised manuscript and in **Response 23** below. Briefly, for this new experiment, we drew a stratified random sample of articles by claim type (direct, implied, conditional, descriptive, based on the strongest form found in the abstract), manually identifying 25 for each category. We then had GPT-4.1 summarize each article under several prompting strategies that approximate real user behavior: (i) a “Basic Summary” requesting a short summary of the article, (ii) a “Simplified Summary” written at an 8th-grade reading level, (iii) “Practical Implications” describing both the result and what it entails, and (iv) a “Careful Summary” emphasizing limitations from a critical perspective. We analyzed these summaries for causal claims using our existing classifier and a new validated classifier developed for this simpler output.

Across prompting strategies, GPT-4.1 summaries frequently omitted careful conditional phrasing and sometimes introduced new unhedged causal claims for papers that used only associational language.

Specifically, among articles whose abstracts contained only conditional causal claims, nearly half of GPT's summaries included an unhedged causal statement (48.0%, 95% CI = [41.1%, 54.9%]). Even for abstracts using exclusively associational or descriptive language, nearly one in five summaries contained a direct causal claim (19.5%, 95% CI = [14.0%, 25.0%]). Although the degree of overstatement varied somewhat across prompting strategies, it persisted even under the prompts explicitly designed for cautious summarization.

We believe these additions sharpen the focus of our paper on LLMs, raising important concerns about these models as they become more integrated into research workflows.

The fourth point is that I would have liked to see more analysis of the "direct", "implied" and "conditional" throughout both the descriptive analyses and the experiment, since to me, the distinctions seem quite interesting and important.

Response 9: We agree that the distinctions among direct, implied, and conditional claims are substantively important. In response, we added descriptive analyses by claim type, examining both longitudinal trends and patterns by journal impact factor. Because this multi-class coding task is more demanding, intercoder reliability was moderate and GPT alignment with human codes was modest, so these descriptives contain more measurement error than our main analyses. To conserve space and avoid overinterpretation, we summarize the key patterns in the main text and present full details in the SI.

Please also see **Response 21**, where we introduce a claim-level analysis that further develops our full-text findings. With the claim-level classification, we were able to integrate different levels of causal claim strength (separating conditional from stronger direct claims), finding a high prevalence of unhedged causal claims within these articles. This claim level approach was also incorporated into our second experiment with GPT-4.1, as is described in more detail in **Response 23**.

Modifications to the Descriptive Analysis:

We added the following to the *Main Text*: “Exploratory analyses that separate claims into different types show that increases over time and with impact factor are driven mainly by an increase in articles with stronger direct and implied claims rather than by more careful, conditional phrasing (see SI, “Results by Causal Type”).

And to the methods: “Although conditional statements are technically justified, we included them because we believe that they may similarly lead to excessive causal inference compared to more careful phrasing. We note, however, that conditional claims are relatively rare: of the articles classified as causal by human coders, only 3% contained only conditional claims while the remaining 97% contained either direct (54%) or implied (43%) claims (where articles contained more than one type of causal claims we classified them by the strongest form present). Note: correlational claims framed as aligned with causal hypotheses were not targeted by our classifier and were therefore generally marked as associational. Given the low prevalence of conditional claims and the modest agreement of human coders when using this four-level coding scheme

(Krippendorff's $\alpha = 0.67$), our primary analysis relies on the simpler binary outcome: whether any form of causal language was present. For this binary classification, human coders demonstrated strong agreement (Krippendorff's $\alpha = 0.86$)."

In the SI, we added a section on "Results by Causal Type":

"To assess whether descriptive results vary by causal-claim type, we analyzed prevalence by type over time and across journal impact factor quintiles. Intercoder agreement for this task was moderate (Krippendorff's $\alpha = 0.67$), and GPT alignment with the majority label was likewise modest (macro-F1 = 0.53; precision = 0.52; recall = 0.62, accuracy = 0.60). These results should therefore be interpreted cautiously. This analysis was conducted on the 23,626 cross-sectional articles exported from TDM Studio for which we have LLM claim-type labels.

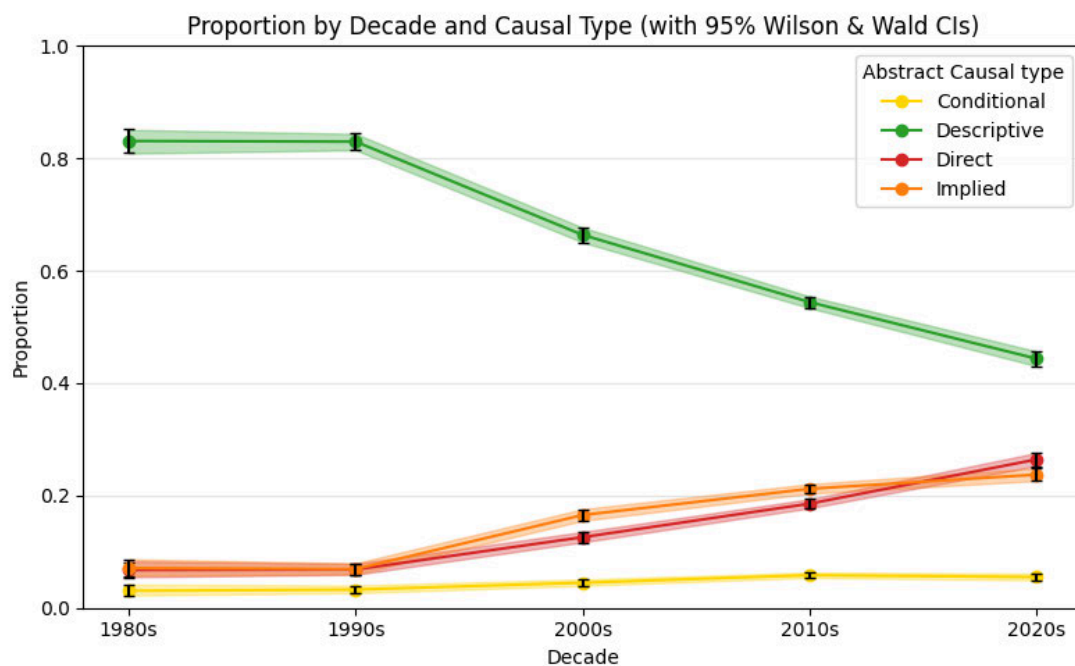


Figure S3. Decade proportions of cross-sectional articles by claim type. The decline in purely descriptive language is driven primarily by increases in direct and implied causal statements.

Figure S3 shows how the proportion of causal claim types changed over time. In 1980, 83% of cross-sectional articles were classified as descriptive; among those with causal language, 6.8% were direct, 7.1% implied, and 3.1% conditional. Beginning in the 1990s, causal phrasing increased as descriptive language declined. By 2020, descriptive classifications fell to 44%; direct (26.4%) and implied (23.7%) claims became much more common, while conditional claims remained comparatively rare (5.5%).

Figure S4 shows the breakdown of causal claim type by journal impact factor, split into quintiles for this sample. It reveals that higher impact factor journals have substantially more causal claims overall, as reported in the main text. In the lowest impact-factor quintile (0–1.69), descriptive language remains a majority (58.9%), with direct (17.9%), implied (19.0%), and conditional

(4.2%) claims less common. In the highest quintile (5.34–19.2), descriptive falls to 41.6%, while direct (28.7%) and implied (24.2%) are over-represented; conditional claims remain a small share (5.6%).

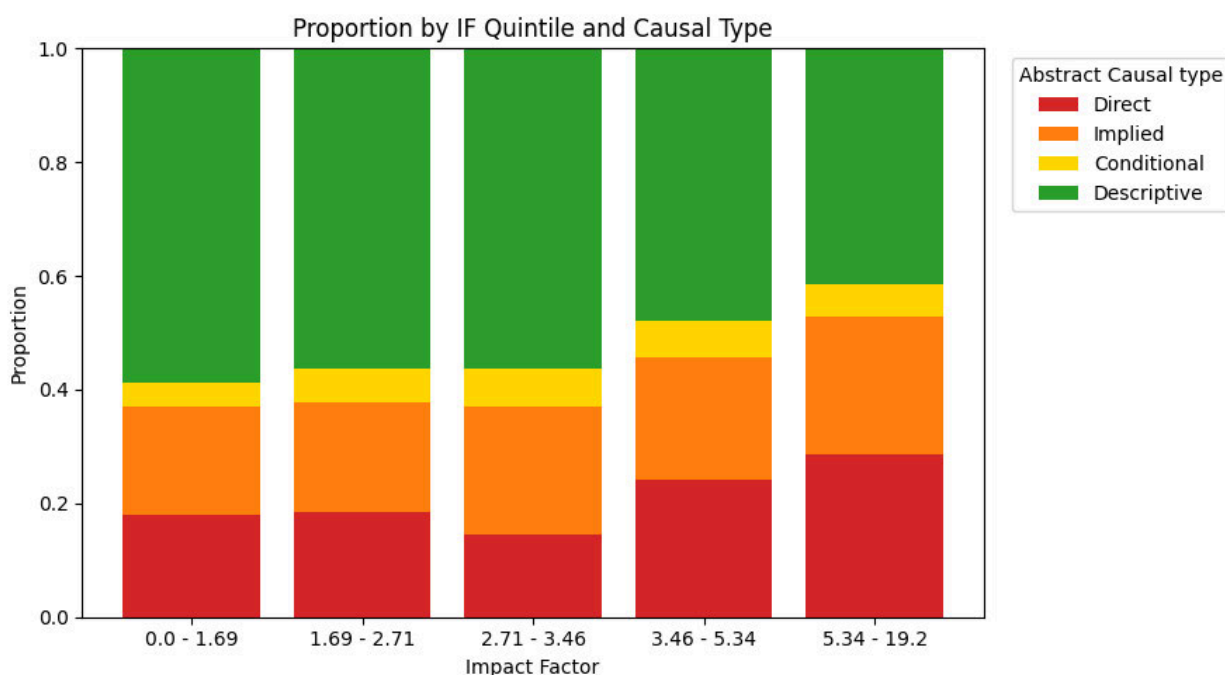


Figure S4. Claim-type distribution by journal impact-factor quintile. Higher-impact journals exhibit more causal language overall, particularly direct and implied claims.

Minor notes: - The term "silicon participant" leads to an anthropomorphization of Generative AI. I suggest rewriting to avoid the word "participant" in that context (and ideally "silicon" too, since a system like GPT is far more than just "silicon" for all kinds of reasons). Something more directly accurate would be preferred, like "synthetic experiments using GPT".

Response 10: We thank the reviewer for this suggestion which was echoed by the other reviewers and the editor. We switched our language to “experiments on GPT-4.1” throughout the paper and carefully reviewed our description of the results to avoid anthropomorphizing these models. We made numerous modifications, many examples of which can be seen in **Response 4** to the editor, above.

References: [1] <https://aclanthology.org/2022.emnlp-main.117.pdf>

Response: We appreciate this interesting citation which we missed in our literature review. We utilized the associated package for claim-level analysis presented in **Responses 21 and 23** below.

Reviewer #1 (Remarks on code availability):

I did not attempt to run the code, but it looks reasonably organized and replicable. They are simple Jupyter Python notebooks and the figures and statistics from the paper are clearly generated from these notebooks.

Response: Two authors independently reproduced the results. For this revision, we incorporated a GitHub package (associated with the provided reference above) whose dependencies conflicted with those used in our main analysis. To avoid versioning issues, we created a separate Python environment and analysis pipeline for the updated analyses. All scripts and environment specifications for these revision-only analyses are available in the “Revision_Analyses” folder within the same OSF project.

Reviewer #2 (Remarks to the Author):

Summary

The paper studies the prevalence and consequences of causal language in cross-sectional studies, a design that generally does not allow for causal inference. The authors document that the use of causal language has increased over time, with a notable rise around 2000 coinciding with the “causal revolution” in economics and other social sciences. They show that both human and AI (“silicon”) participants classify studies as causal when such language is used, and that interventions can mitigate this tendency. The study suggests that causal language can shape perceptions of research credibility and interpretation.

General Appreciation

The paper addresses an important and timely issue. The use of causal language in non-causal research designs is widespread and has implications for how research is read, cited, and used in policy. The documentation of temporal trends and the experiments with readers are valuable contributions. At the same time, the paper could be strengthened by providing further contextualization (e.g., comparing cross-sectional studies to true experimental studies), by clarifying some methodological details (statistical tests, reporting practices), and by elaborating on the motivation for using AI participants.

Response: We thank the reviewer for their careful consideration of the paper and useful advice for improving its quality. We have made a number of modifications and hope we have addressed each concern adequately, through our methodological clarifications and expanded experiments with AI models.

Comments

1. Benchmarking prevalence – While the study carefully documents the prevalence of causal language in cross-sectional studies, it would be helpful to provide a benchmark from experimental or quasi-experimental papers. Such a comparison would help readers assess whether the prevalence in cross-sectional work is unusually high, or in line with broader research practice.

Response 11: We agree that benchmarking against experimental and quasi-experimental articles can help clarify whether the prevalence of causal language in cross-sectional work is unusually high. Accordingly, we added a comparison to the prevalence of causal claims within articles that our GPT classifier labeled as randomized controlled trials or quasi-experimental studies (obtained from the smaller training and evaluation sample articles that were exported from TDM Studio, the only ones for which we had detailed labels aside from the binary “cross-sectional” or not). Due to word limit constraints, we primarily report these in the SI.

In the main text, we summarize and point readers to the more detailed results:

As an additional baseline, we also compared this rate with the prevalence of causal claims in a separate sample of (quasi-)experimental studies, finding that the latter is higher; however, this discrepancy has diminished over time (see SI, “Comparison to Experimental Studies”).

We added a section to the SI, “Comparison to Experimental Studies”:

“To place the proportion of causal language in additional context, we also separately computed the prevalence of causal claims in the abstracts of 1,538 articles that were among the 40,000 articles that we exported from TDM Studio in stage 1 of our pipeline but were subsequently classified (in stage 2) as RCTs (random controlled trials) and quasi-experimental methods. As these designs are tailored for causal identification, we would expect causal claims to be more prevalent in both categories than in cross-sectional designs. As expected, for articles classified as RCTs the overall average prevalence was 71.1% (N = 218; 95% Wald CI = [65.1, 77.1]) while for quasi-experimental studies it was 77.0% (N = 1,313; 95% Wald CI = [74.7, 79.3]). As with cross-sectional designs, these rates are also increasing over time; however they do so at a slower rate. For example, in the interval 2000-2010 (N=225), the combination of RCTs and quasi-experimental designs exhibited an average prevalence of 71% compared with 35% for cross-sectional designs; for 2010-2020, the corresponding rates are 78% (N=664) and 48% respectively; and for 2020-2025 83% (N=498) and 56%. Thus, whereas causal claims among experimental and quasi-experimental methods were once more than twice as common as among cross-sectional designs, they are now only about 50% more common.”

2. Role of incorrect causal statements – The discussion could acknowledge more explicitly that even incorrect or overstated causal claims can sometimes stimulate valuable research progress. For instance, the Cowles Foundation report on research in economics triggered substantial improvements in methods and standards. Highlighting this aspect could bring nuance to the normative perspective.

Response 12: We appreciate this important historical point and have revised the discussion to explicitly address the catalytic role that incorrect causal claims can sometimes play. We now acknowledge that some incorrect or overstated claims have stimulated valuable methodological debates and follow-up research (e.g., the Cowles report). However, we also clarify a key distinction that underpins our normative position: such post hoc catalytic benefits should be treated as fortunate side effects rather than a guiding standard for scientific practice. Even if errors do incidentally trigger progress, it can still lead to misallocated research effort, confusion among policymakers and experts, and potentially risks to public

trust. For example, Daryl Bem's infamous 2011 "ESP study" is often credited with directing much-needed attention to the impact of researcher degrees of freedom on false positive findings. By stimulating reforms such as pre-registration and replication studies, the Bem paper arguably did make a valuable scientific contribution, but it also raised serious questions about generations of prior work that embraced similar methods.

On balance, therefore, we would argue that social and behavioral science would be better off today had we identified these problems, and corrected them, much sooner. In any case, we would certainly not view the Bem example as something to be emulated just because it led to some positive outcomes. Similarly, unethical and even fraudulent research such as the Stanford Prison Experiment or Milgram's obedience studies has also historically stimulated productive follow-up work, but we would still not wish to repeat the misconduct associated with the original work. In summary, our view is that the scientific community should always strive to improve its practices—in this case by insisting on careful and transparent communication of results—even if it has benefitted from earlier, problematic practices.

To clarify this view we have now added:

- "Furthermore, overstated claims can occasionally accelerate progress, by provoking methodological reform or rapid follow-up work, as in the Cowles Foundation's econometric contributions or the LK-99, ESP, and cold-fusion episodes^{76–78}. We also note that humans are, at their core, storytelling creatures^{63, 79}, who understand and retain information more effectively when it is embedded in narrative form. Eliminating narrative altogether from science communication is therefore not only likely impossible but also likely counterproductive. Although we continue to argue that caution when making causal claims is warranted in science communication, future work should consider how to balance the harms caused by careless rhetoric against the harms caused by excessive skepticism"

3. Abstract-based experiments – The experiments based on abstracts are insightful, yet they may overstate the real-world impact since information in the full text can moderate impressions created by the abstract or title. The paper already documents cases where causal language appears later in the text; it may be useful to also note the possibility of the reverse. The intervention results partly address this issue, but a more explicit discussion of these limitations would strengthen the interpretation.

Response 13: We agree that our abstract-based experiments may overstate real-world impact if information in the full text moderates reader impressions. To clarify scope and temper interpretation for readers of our own paper, we first made changes to our descriptions to clarify the nature of the data.

For example, in the *abstract*, we more clearly label that the experimental stimuli were abstracts.

- Previously: "...audiences often believe studies..."
- Now: "...readers frequently indicate abstracts..."

And in the *main text* we added:

- “Accordingly, we developed a three-step LLM-based workflow with human-in-the-loop validation to analyze social science articles across five ProQuest databases, primarily focusing on titles and abstracts, which are likely to receive disproportionate reader attention.”

We also added “abstracts” at several other places when describing the stimuli and methods.

Additionally, we added an extended description of this as a limitation.

- “Relatedly, our main results focused on titles and abstracts, which omit context that careful readers might glean from full texts. Although we found that full texts contain more causal language overall, they also include methodological caveats and associational claims that could temper interpretation. Our second GPT-4.1 experiment, which compared abstract-only versus full-text inputs, showed no difference in causal framing for the model summaries, but future human-subject experiments should test whether reading the full texts increases or decreases causal interpretations.”

As that limitation paragraph demonstrates, we ran an exploratory experiment with GPT output in which summaries were generated from title and abstract-only versus with full-text inputs. We found that summaries’ causal-claim classifications did not differ significantly across these conditions. See **Response 23** for details.

Finally, we conducted expanded full-text analyses and found results that cut both ways for how readers might infer causality with additional information from the main text (see details in **Response 21**). On the one hand, full texts were much more likely to contain a direct causal statement: all articles labeled as direct from the abstract contained an unhedged causal claim about their key findings somewhere in the main text, as did 72% of articles with only conditional or implied causal claims in the abstract and 44% of those with only descriptive statements. These unhedged causal claims were concentrated in the Introduction and Discussion and were also common in the Results, indicating that scholars are very likely to use causal language in the full text. On the other hand, most findings claims were associational (73.0%), which may temper any overall causal interpretation. These results are described alongside Figure 4 in the revised submission.

4. Use of “silicon participants” – The inclusion of AI participants is an innovative idea, though the added value could be explained in more detail. The results show higher compliance by silicon agents, which may reflect model parameters such as “Temperature” and “Presence Penalty.” Clarifying this point and discussing how such choices shape outcomes would increase transparency. One possibility would be to also assess silicon agents in settings with causal designs (e.g., randomized experiments or natural experiments). Contrasting outcomes in causal versus non-causal contexts could illustrate more clearly what these agents are capturing.

Response 14: We agree that the value of our experiments with GPT warrant clearer articulation and expansion. We have expanded the Methods to (i) explain what the AI agents add and (ii) document all model settings, and we added a limitation noting parameter/model dependence.

We made the following modifications:

- We added: “For our experiments with GPT-4.1, we use temperature = 1.0 and the default presence penalty (0). For multiple-choice items, max_tokens = 5; for summaries, max_tokens = 100. All other parameters were left at July 2025 batch-API defaults.”
- To the limitations we added “Moreover, our experiments with GPT-4.1 reflect a single model and a few prompt configurations. Other models or parameters may yield different outcomes, and we encourage continued evaluation of rhetorical overclaiming as new systems are deployed.”

Finally, as noted above and in response to Reviewer #3, we expanded our GPT experiments. While we considered adding studies with different methodological designs, we opted against this due to space constraints and to maintain a focused examination of causal overstatement within cross-sectional designs. Nonetheless, we partially address this point with our individual claim-level analysis (**Response 21 and 23**) where we now assess which claims GPT emphasizes in its summaries. We found that when an article contained multiple claims, only some of which were causal, GPT-4.1’s summaries routinely highlighted the causal claim over associational ones.

5. Figure 3 / time trends – It would be helpful to clarify whether the evidence in Figure 3 accounts for time effects, for example by ensuring that comparisons are made across journals within the same time period.

Response 15: We agree that Figure 3 could, in principle, reflect changes in journal composition over time. To address this, we estimated a multilevel logistic regression predicting whether an abstract contains causal overstatement, with impact-factor quintile as the focal predictor, year fixed effects to absorb time trends, and cluster-robust standard errors at the journal level. The association between impact factor and causal overstatement remains substantial and statistically significant after conditioning on year. We added a brief note to the main text and a full description of this analysis in the SI.

- We added, “A multilevel logistic regression of causal language on impact-factor quintile, with year fixed effects and journal-clustered standard errors, showed that this association persists within year (see SI, “Within-Year Impact-Factor Differences”)”
- In the SI, we added the section “Within-Year Impact-Factor Differences”: “In the main text we report differences in causal overstatement across time and by journal impact factor (IF). Because these patterns could be confounded if higher-IF journals contributed a growing share of articles over time, we fit a multilevel logistic model with IF quintile as the predictor, year fixed effects, and cluster-robust SEs at the journal level. Relative to the lowest IF quintile, the odds of causal overstatement were modestly higher in the second quintile (OR = 1.27, 95% CI [1.08, 1.48]), statistically indistinguishable in the third (OR = 0.97, 95% CI [0.74, 1.26]), and substantially higher in the fourth (OR = 1.59, 95% CI [1.28, 1.98]) and fifth (OR = 2.37, 95% CI [1.99, 2.82]) quintiles. Thus, higher-impact journals exhibit more causal overstatement even after conditioning on publication year, indicating that Figure 3 is not explained by time trends alone.”

6. SM: One-tailed tests (p.15) – Please provide a brief justification for the choice of one-tailed rather than two-tailed tests.

Response 16: We pre-registered directional hypotheses and specified one-tailed tests (with multiple-comparisons adjustments) for those contrasts. We believe a one-tailed test is appropriate when theory predicts a change in a single direction and an effect in the opposite direction would count against the hypothesis rather than for it. For transparency, we verified that our main inferences are unchanged under two-tailed tests; no result crosses the significance threshold when switching to two-tailed p-values.

- We added, “Given this power analysis, we used one-tailed tests with Holm-Bonferroni correction to increase power for detecting effects in the pre-specified direction; this decision, along with our directional hypotheses, was pre-registered, and all conclusions remain substantiated with a two-tailed test.”

7. SM: Transparency of results – The paper and supplementary material mention multiple-testing corrections, but standard errors and uncorrected p-values are not reported. Providing these would allow readers to better assess the robustness of the findings and verify the implementation of corrections.

Response 17: We agree that adding standard errors and uncorrected p-values can aid transparency; however, we note that our analyses utilize different modeling approaches (e.g., logistic and linear), hence SEs are not directly comparable across models. For consistency, we therefore report 95% confidence intervals for all model outputs as a substitute. We also judged that listing uncorrected p-values alongside corrected values would potentially confuse readers. To balance clarity with completeness, we added 95% CIs throughout the experimental results, and we keep the corrected p-values in the text.

Our code implementing the multiple-comparisons adjustments is publicly available for verification. On OSF (Experiment folder), see Analysis-Final.R: the Holm–Bonferroni corrections appear on lines 45, 63, and 88, and the corresponding uncorrected p-values are stored in the `p_one` variables immediately above those lines. For easy reference, we provide the uncorrected and corrected p-values here:

	Causal Inference		Subtle/Implied		Summaries	
	Raw P	Corrected	Raw P	Corrected	Raw P	Corrected
Rewrite	0.003	0.003	0.092	0.092	2.2×10^{-6}	6.5×10^{-6}
Note	3.1×10^{-5}	6.2×10^{-5}	0.045	0.089	0.078	0.078
AI	4.8×10^{-9}	1.4×10^{-8}	0.024	0.073	0.011	0.022

Under Holm–Bonferroni, the smallest p-value is multiplied by the number of tests, the next by $N-1$, and so on; this is the procedure applied here.

Here is an example of the revision made in the text, though we encourage reviewing the updated section for all of the edits.

- “In relative terms, participants who read a rewritten version with only associational language showed lower agreement ($b_{\text{rewrite}} = -0.36$, $p = 0.003$, 95% CI = [-0.63, -0.10]).”

Reviewer #3 (Remarks to the Author):

Thanks a lot for submitting this interesting article which clearly tackles an important problem.

Overall strength:

- (1) important and timely topic, and I think studying causal claims in cross-sectional social science articles is neat
- (2) using LLM + fine-tuned model to identify causal claims is reasonable and the model could be potentially used by other researchers
- (3) nice study on audience perception of causal claims

Response 18: We thank the reviewer for carefully reviewing our manuscript and offering useful suggestions for addressing remaining weaknesses. We hope we have addressed this through expanding our limitations and adding additional experiments with GPT-4.1.

Key weaknesses and suggestions

- (1) While studying cross-sectional social science articles is interesting, it is unclear how much these findings can be generalized to other types of articles. It would be meaningful to discuss the generalizability of the findings

Response 19: We agree that focusing on cross-sectional studies, while sharpening our ability to assess causal overstatement, has limited generalizability to other designs. We have expanded the discussion of this limitation and suggest exploring other forms of “narrative license” or rhetorical overreach in different study designs is an important direction for future research.

- “Second, our focus on cross-sectional studies, although helpful for creating clear criteria for identifying instances of causal overreach, does not imply that other designs are less susceptible to narrative license. For example, carefully executed cross-sectional evidence can, in some settings, be more informative for theory testing than, say, laboratory experiments with limited external validity. Our aim is not to criticize any single design or to single out any one manifestation of overclaiming, but rather to call attention to the broader phenomenon of narrative license and to introduce methods for identifying it. Future work should therefore examine other forms of narrative license within other research designs, including (quasi-)experimental designs, simulation studies, mathematical modeling, and ethnographic research. If systemic pressures such as funders’ emphasis on “impact” or publication incentives are driving these trends, we would expect parallel increases in other modes of overstatement cataloged in Table 2.”

(2) One key weakness of the study is that it focuses on abstracts instead of individual findings. A paper might have multiple conclusions and it's possible that they mostly used correlational language but use causal words in some. It would be nice to conduct finding-level analysis to understand whether specific scientific findings are presented in a causal way. Furthermore, I'm not sure if the current methodology could support the claim that "full texts are substantially more likely to contain causal language". It would be necessary to compare the same findings presented in the paper abstracts and the main text to better understand the dynamics behind this. There is a paper (<https://aclanthology.org/2022.emnlp-main.117.pdf>) with an associated package that allows you to easily identify pairs of scientific findings: <https://github.com/copenlu/scientific-information-change>. So basically, with clusters of findings presented in different sections, it should be fairly straightforward to study whether the authors convert a correlational claims to a causal claim in certain sections.

Response 20: We agree that a finding-level analysis would materially strengthen the paper. Our earlier full-text result speaks to document-level prevalence (e.g., using keywords and our classifier we observed that 81.8% of articles contain at least one causal claim), but it does not align specific findings across abstracts and main texts. The reviewer's proposed approach addresses this limitation.

We implemented this new approach (see **Response 21**) and found that authors frequently shifted from correlational to causal language in the full text (matching claims using the provided package). All articles that included direct causal claims in their abstract also contained matching direct claims elsewhere in the manuscript, predominantly in the Introduction and Discussion sections and often in the Results. More importantly, 72% of articles with only conditional or implied causal claims in the abstract, and 44% of those with only descriptive statements, presented unhedged causal versions of these claims within the main text, again concentrated in the Introduction and Discussion and commonly appearing in Results. By contrast, the Methods section contained few causal claims, though this was due primarily to a lack of matched claims rather than deliberate avoidance of causal language. We discuss this in more detail below.

(3) the types of causal claims would also be an interesting thing to study. For example, would the authors hedge more when making causal claims? It would be nice to analyze the uncertainty of the claims and see whether the authors use hedges as a strategy to make causal claims for correlational findings. This also requires analysis at the finding level.

Response 21: We agree that examining hedging and uncertainty at the finding level is important. In our initial approach, we proxied certainty by classifying abstracts into four types: direct, implied (e.g., answering a causal question or drawing behavioral/policy implications that presume a causal relationship), conditional (hedged), and descriptive. When multiple claims appeared within an abstract, we coded the abstract according to the strongest type identified (Descriptive < Conditional < Implied < Direct). We ultimately deemphasized this strategy because coding at the abstract level yielded only moderate intercoder reliability, and modest agreement with our GPT-classifier. Nonetheless, we conducted additional analyses using these abstract-level labels; the results are presented in the new SI section "Results by Causal Type" and summarized in **Response 9** above. In brief, conditional (hedged) claims are substantially less common than direct or implied causal claims in abstracts.

Building on the reviewer's suggestion, we now implement finding-level coding, where these hedged claims and uncertainty can be more cleanly distinguished, as described in detail below.

In our methods, we added a new section “Full Text Claim-Level Analysis”:

“The primary abstract-level analysis was limited in that it relied on a single classification per abstract, even though abstracts often contain multiple claims, and full texts contain many more. This approach also conflated distinct forms of causal language (collapsing conditional with implied and direct causal claims). Although conditional causal claims represented only a small portion of the dataset, we acknowledge that such claims are more justifiable than direct or implied ones, and ideally should be analyzed separately. It is therefore important to assess how frequently full-text claims employ conditional (hedged) versus stronger statements that explicitly assert a causal relationship.

To address these issues, we conducted an additional analysis on a curated sample (N = 100) of articles representing different levels of causal strength, breaking them down into individual claims to allow multiple labels per article. To construct this sample, we randomly selected abstracts from the 30,000 originally exported from TDM Studio that had been classified by GPT into four categories (Descriptive > Conditional > Implied > Direct; see SI "Results by Causal Type"). Two authors reviewed these abstracts to verify the accuracy of GPT's classification, with the final sample consisting of 25 articles for each category where GPT and the two human coders showed complete agreement (Total N = 100).

To analyze individual claims, we examined sentences as a smaller unit of analysis. Abstracts contain sentences that serve multiple functions, including providing background, stating the research objective, describing methods, presenting results, and articulating implications. We split each abstract into sentences using the sentence tokenizer in the python spaCy package, classified all sentences into one of those categories, and restricted our analysis to those classified as about the results. This approach is conservative: it omits all implied claims made in the research objective as well as those that appear in the implications. Thus, any causal claims remaining in this analysis should explicitly concern the novel findings reported by the authors. Note: there were three abstracts that did not contain any sentences whose primary label was results; for these, we selected sentences that combined results with implications or results with background.

Using the *Scientific Information Change Package*⁸⁶, which was developed to measure the degree of similarity in the description of scientific findings, we compared claims across abstracts and main texts. Following the procedure described by Wright et al. ⁸⁶, we classified pairs of claims with an Information Matching Score (IMS) greater than 3 as representing the same finding. We retained only the first occurrence of each paired claim from the abstract, removing duplicates. This process yielded a total of 182 claims across 100 abstracts, with the number of claims per abstract varying (M = 1.82, SD = 1.1, Min = 1, Max = 6). We then applied our fine-tuned BERT-based classifier to determine whether each claim was causal or noncausal. In addition, we assessed the presence of hedging terms ("may", "would", "can", "might", "could", "appear to", "seem", "likely", "potential", "possible", "indicative", "probable", "suggest", "support", "model",

"relationship", derived from 87). Claims classified as causal without a hedge were classified as *unhedged causal*, causal claims with a hedge were classified as *conditional causal*, and noncausal statements were labeled *associational/descriptive*. Because we focused only on the results sentences, we removed the implied causal category.

This analysis revealed that many claims within each abstract-level causal category were associational or descriptive, although causal claims were more frequent in abstracts previously classified as causal across types. Instances of unhedged causal claims appearing in descriptive, conditional, or implied abstracts often arose when a sentence appeared associational in context but seemed causal when viewed in isolation. For example, the statement, "Our results show that negative returns to health outcomes set in at around 50 work hours per week, and that the negative effects of working long hours manifest earlier for women than men." immediately precedes a sentence clarifying the associational nature of the finding. While in context the statement is arguably noncausal, when analyzed in isolation, its phrasing ("set in," "manifest") was classified as an unhedged causal claim by our classifier. This example illustrates a central tradeoff in the sentence-based, claim-level classification: while it allows for greater granularity, it can obscure contextual nuance. The breakdown of claim classifications by abstract-level category is presented in SI Figure S8.

Next, we examined paired claims within the full text. We first gathered the full texts for all papers in the above analysis and manually divided each article into four sections: Introduction, Methods, Results, and Discussion. Not all journals adhere strictly to this structure—some intersperse discussion within the results, and others place descriptive results in the methods section—so we used our judgment to assign text to the section that best fit its function. For each section, we followed a procedure parallel to the abstract-level analysis. We tokenized the text into sentences, then used regex patterns to remove any sentence containing references to prior work, in order to focus on sentences describing the current study. We then matched all remaining claims to the unique abstract claims using the *Scientific Information Change* package, retaining all pairs with an information-matching score greater than 3.0. This generally yielded multiple matches for each abstract claim, with substantial variability (Total N = 3,134; Median = 11; M = 18.1; SD = 18.4; Min = 1; Max = 100). These matches were unevenly distributed across sections: the Introduction (M = 6.6) and Discussion (M = 6.3) contained more claims per abstract claim than the Results (M = 4.2) or Methods (M = 1.0). For our main full-text analysis, we retained all unique matched claims. We applied our causal language classifier and hedge detector to all matched full-text claims, categorizing each as unhedged causal, conditional causal, or associational/descriptive.

We note that this approach has some important limitations. First, by focusing only on explicit results claims, we overlook subtler ways in which authors imply causality in their framing of the research question or in their proposed implications, which may lead us to underestimate the prevalence of causal language relative to methods that incorporate these claim types. Second, because the claim-level analysis is conducted at the sentence level, it can sometimes miss contextual cues indicating that the results are associational or providing important limitations or caveats. These limitations are addressed with our abstract-level classification, indicating the importance of multiple approaches to analysis, and providing convergent validity for the

prevalence of overclaiming. Refer to SI “Full Text Claim Level Analysis - Additional Results” for additional details.

We also changed our full text figure (now figure 4) to illustrate these results:

“To test for this possibility, we conducted a second analysis in which we searched for specific claims in the full texts of a smaller subsample of articles ($N = 100$) that closely matched claims made within the abstract (see SI “Full Text Analysis” for details). For this set of matched claims, we observe both a lower relative rate and a higher overall prevalence of inappropriate causal claims compared to the title/abstracts. On the one hand, the majority of matched claims were associational or descriptive ($N = 2,287$; 73.0%), followed by strong, unhedged causal ($N = 487$; 15.5%) and more careful, conditional causal claims ($N = 360$; 11.5%), suggesting the main texts were on average more moderate than their titles and abstracts. On the other hand, consistent with our first analysis, articles frequently included at least one unhedged causal claim even when the title and abstract did not use causal language. As shown in Figure 4a, 100% of articles that contained an unhedged causal claim in the title or abstract also contained at least one matched causal claim in the full text. Moreover 72% of papers with conditional or implied claims and 44% with only descriptive claims in the title/abstract had at least one unhedged causal match in the full text. Figure 4b shows that these unhedged causal claims occurred mainly in introductions and discussions and were rare in the methods.”

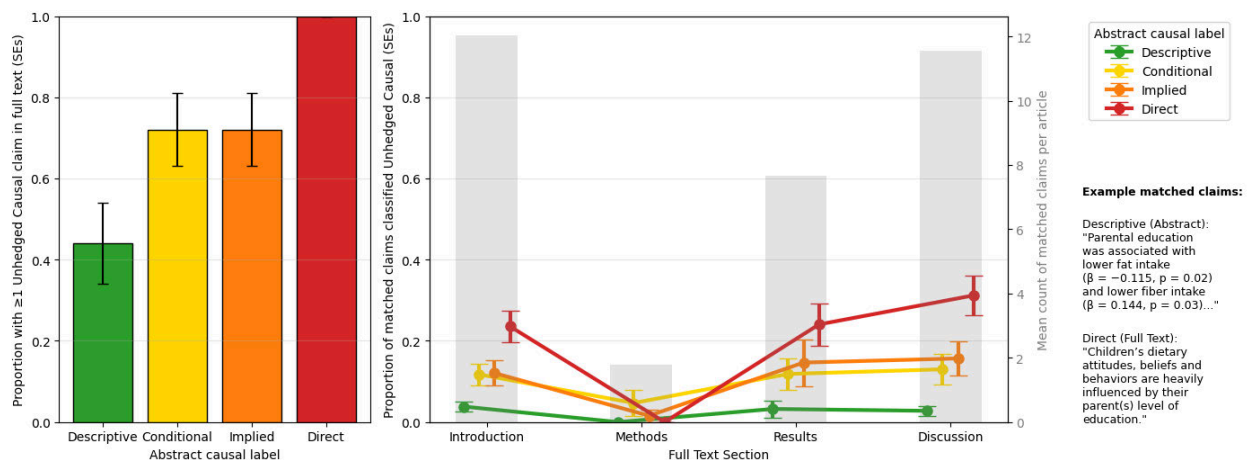


Figure 4. Causal claims present in the full text, grouped by article abstract-level classifications. (a) Proportion of articles with at least one unhedged causal claim in the full text that matched a claim made about the article's findings within the abstract. Even articles with abstracts that contain only conditional or descriptive claims frequently include unhedged causal claims in the full text that otherwise match the more carefully worded claim from the abstract. (b) Lines: Proportion of matched claims labeled unhedged causal by section and abstract-level causal classification. Bars: Mean number of matched claims in each section. Across abstract-level categories, most matched claims in the full text remain associational, and the unhedged causal claims that do appear are concentrated in the introduction and discussion, with far fewer in the methods. This lower proportion reflects both the smaller number of matched claims ($N_{Intro} = 1,150$; $N_{Method} = 168$; $N_{Results} = 720$; $N_{Discussion} = 1,099$) and their lower likelihood of

being causal (proportion unhedged causal: Intro = 17%, Methods = 4%, Results = 14%, Discussion = 17%; $\chi^2(6) = 81.0$, $p = 2.2 \times 10^{-15}$). An example of a matched abstract-full-text claim pair is shown to the right. Error bars represent standard errors.”

Note: We moved our previous Figure 3 to the SI to accommodate this updated figure.

We also added more results to the SI:

“Full Text Claim Level Analysis - Additional Results

Here, we report additional results from our full-text, claim-level analysis. Please refer to that Methods subsection for details on how we identified and matched claims within the texts. First, Figure S8 shows the distribution of causal claim types among sentences classified as reporting results in the abstracts. As expected, most claims were associational/descriptive, and all causal abstract-level categories (Direct, Implied, Conditional) contained substantially more causal claims than abstracts labeled as purely associational.

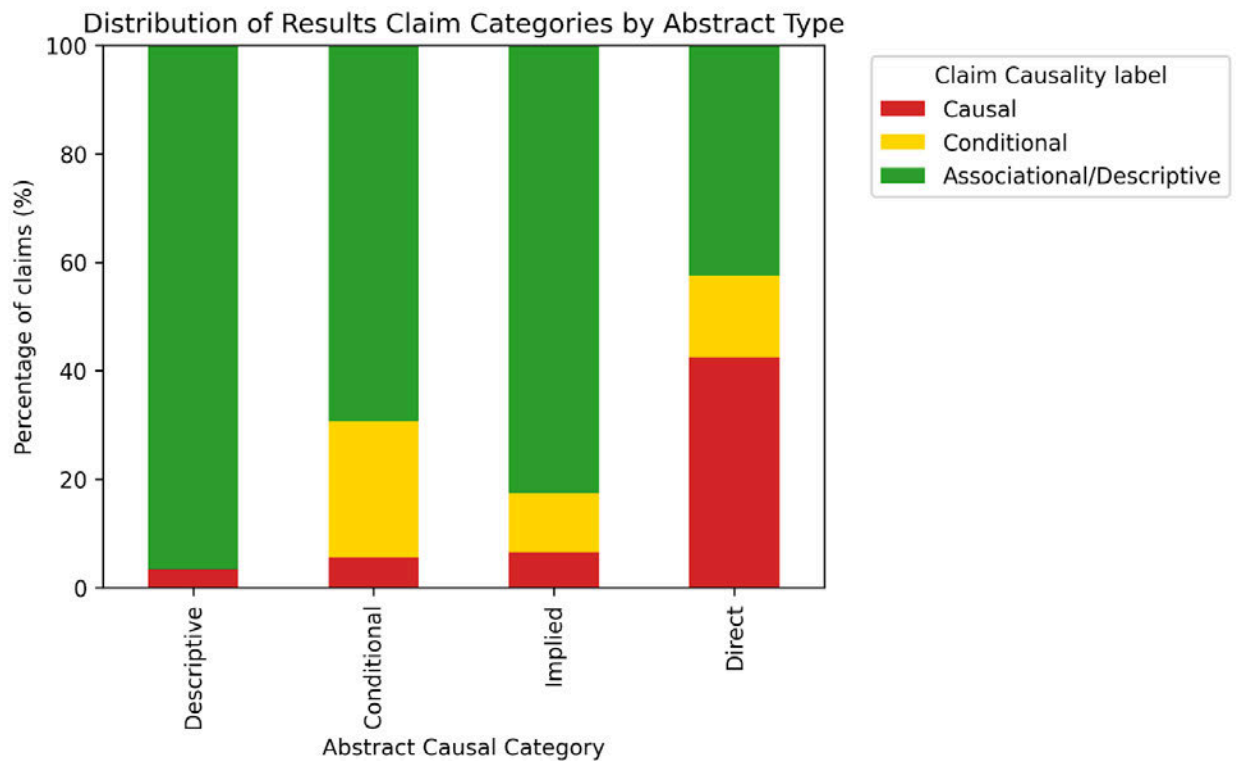


Figure S8. Distribution of Results-Claim Categories, within the abstract, by Abstract Type. All abstracts contained a large proportion of associational/descriptive claims about their results, though those with a causal abstract-level label included substantially more causal and conditional causal claims.

We next turn to matched claims within the full text. The majority of full-text claims were associational/descriptive ($N = 2,287$; 73.0%), followed by unhedged causal ($N = 487$; 15.5%) and conditional causal ($N = 360$; 11.5%) claims. When matched to abstract-level claim types, we observed that associational abstract claims ($N = 138$) had several associational matches in the full

text ($M = 11.2$), but also included conditional causal ($M = 1.3$) and unhedged causal ($M = 1.5$) counterparts. Conditional abstract claims ($N = 20$) showed a similar pattern but with more matches across all types (Associational/Descriptive = 17.3; Conditional = 5.1; Unhedged Causal = 4.7). Direct abstract claims ($N = 24$) produced even more unhedged causal matches (Associational/Descriptive = 18.6; Conditional = 4.2; Unhedged Causal = 8.0).

In all cases, associational claims were most common, yet papers frequently used causal language to describe results. When examining whether each abstract claim had at least one unhedged causal match in the full text, a binary indicator of causal overclaiming, we found that 38% of descriptive claims ($N = 50$), 75% of conditional claims ($N = 15$), and 92% of unhedged claims ($N = 22$) did so. Because abstracts often contained multiple result claims, aggregating to the paper level produced even higher proportions (as discussed in the main text). There we also analyzed the distribution of causal claims across sections of the full text, finding they were overrepresented in the Introduction and Discussion and comparatively rare in the Methods section. Note: Given the small sample sizes for each claim type, these proportions have wide margins of error.

(4) The language the authors used to describe GPT evaluation results is problematic and can be very misleading. For example, "GPT was already hesitant to say" has a serious anthropomorphism issue. I also STRONGLY discourage the authors from using "Silicon participants" in a serious research article, as it may create wrong perceptions that AI can actually be treated as research subjects like human participants. Using AI to conduct experiments is fine, but it must be described in an appropriate way.

Response 22: We thank the reviewer for pointing out this careless use of language on our part. We have removed the anthropomorphic phrasing and the term "silicon participants." Throughout, we now describe these as experiments with GPT rather than attributing intentions to the model. We outline many of these changes in **Response 4** above.

(5) I do like the GPT summary analysis as it provides additional evidence on how these inappropriate causal claims might become a bigger problem when everyone uses AI to access information. Please provide more details on how you use GPT to generate the summaries of the paper. To me, how AI models interpret the causal claims seems less exciting. It would be great if you could provide further empirical evidence on how the causality of the findings affects AI summaries. For example, if you provide 5 findings to AI, will AI be more likely to pick the causal one? This will further strengthen the implications of your findings.

Response 23: We agree that the downstream consequences for AI-generated summaries are an important part of the story, and we have both clarified our GPT procedures and extended the analysis to speak more directly to this concern.

First, we now provide fuller implementation details for the summarization experiments. In the revised manuscript, we specify the exact model (GPT-4.1), version, parameter settings (e.g., temperature, max tokens, number of responses), and prompt templates used to generate summaries. The full prompt text is

included in the SI, and all replication data and code are available on OSF to facilitate replication and extensions.

Second, to examine how the mix of findings corresponds with differences in GPT's summaries, we leverage the stratified sample of 100 articles for which we conducted sentence-level coding of claims in the full text (see **Response 21**). These data allow us to address the kind of question raised: when an article contains multiple findings (some associational, some causal) how likely is a GPT-4.1 summary to contain direct causal claim? Our analysis shows that, although associational claims are numerically dominant in these papers, GPT-4.1's summaries are very likely to include a direct causal statement whenever even a minority of the underlying claims are causal. For example, in a mixed-effects logistic model estimated on the new experimental data, if an article makes 10 claims and only 2 are coded as direct causal, the predicted probability that GPT-4.1's summary contains a direct causal claim is 58.4%. In other words, when several findings are available, GPT-4.1 disproportionately foregrounds causal ones, even when they are in the minority.

In summary, this analysis shows that GPT's summaries increase the prevalence and prominence of causal claims, across varied prompting strategies and regardless of whether the full text or only the abstract is provided.

We added the following to the methods:

“Second Experiment with GPT-4.1

Our second experiment with GPT-4.1 explored whether the probability of a direct causal claim appearing in GPT summaries varied based on the abstract-level causal classification, the prompt used for summarization, and whether the model was provided with the full text compared with only the abstract.

For our sample, we utilized the stratified sample of 100 articles used in the "Full Text Claim Level Analysis", which were evenly divided by abstract-level causal classification (Direct, Implied, Conditional, Associational/Descriptive). This design allowed us to test whether GPT added stronger causal language for articles whose abstracts were conditional, and whether it introduced causal claims when abstracts were purely associational from the start. We created four prompts representing diverse real-world uses of language models for interpreting academic texts: (1) a basic summarization prompt, (2) a careful prompt encouraging caution and skepticism, (3) a simplified prompt requesting a summary at an 8th-grade reading level, and (4) a practical prompt emphasizing real-world significance (See SI "Second Experiment with GPT-4.1 - Prompts").

For each article and prompt, we submitted two requests to GPT-4.1: one including only the title and abstract, and another including the full text, to test whether access to additional context affected causal framing within summaries. Each request used a temperature of 1, a maximum of 150 tokens, and generated five outputs, with all other parameters set to the default GPT-4.1 Batch API settings as of November 4, 2025. We developed a new classifier prompt tailored to these simplified summaries, which showed moderate alignment with the consensus human label on a held out validation sample of summaries (Macro F1 = 0.75, Precision = 0.75, Recall = 0.79, Accuracy = 0.75, N = 240; see SI "Second Experiment with GPT-4.1 – Causal Classifier"). Each

output summary received a single label corresponding to the strongest causal claim type present (Unhedged Causal > Conditional Causal > Associational/Descriptive). As with our full text analysis, we removed the implied category, focusing on causal claims about the findings. We also relabeled the direct category as "unhedged causal", with all unhedged causal claims receiving this label. For each article–prompt–text-type combination, we aggregated across the five outputs and assigned the majority label; in cases of ties, one label was randomly selected from the tied categories. Notably, in cases of mismatch between GPT and the consensus expert judgement, the human coders were substantially more likely to label summaries as containing an unhedged causal claim than GPT (Z test of proportions = 3.3, $p = 0.0008$). There were 44 cases where the human's labeled a sample as "Causal" when GPT's label was either conditional or associational, and only 8 with the reverse.

To examine whether our results depended on our method for identifying causal language, we conducted sentence-level analyses using our causal language classifier from the descriptive analysis, an independent Bert-based classifier trained on article conclusions³⁵, and a hedge detector, paralleling the procedures used in our full-text analysis (See SI "Second Experiment with GPT-4.1 - Robustness Checks"), finding similar results. We also tested whether abstract claim mix (descriptive vs unhedged causal) predicts causal language in LLM summaries, and found that more causal language in the article yields more causal claims in summaries (See SI "Second Experiment with GPT-4.1 - Claim Level Analysis").

Analytical approach:

We report descriptive statistics showing the proportion of summaries containing each causal claim type, disaggregated by prompting strategy and by the article's abstract-level causal classification. To test whether the presence of causal claims in research summaries differed across prompts and abstract categories, we conducted Chi-square tests of independence. To compare summaries generated from abstracts versus full texts, we used the Stuart–Maxwell test to assess whether the probability distribution of output categories differed across these paired input types. Because each article includes a complete set of summaries across input types (abstract vs. full text) and all four prompting strategies, matched on abstract-level causal labels, these dimensions were analyzed independently without additional statistical controls. Finally, we compare the proportion of causal language across key dimensions using two proportion z-tests. As this experiment was not pre-registered, we report unadjusted, two-sided p-values for significance testing.”

And this to the results:

“The [first experiment with GPT-4.1] had three important limitations: it relied on a single prompt; it did not account for the type of causal claim in the abstract; and it used only abstracts rather than full texts. To address these limitations, we conducted a second GPT experiment that employed four summarization prompts: a basic summarization, similar to the previous experiment; a simplifying summary, in which GPT was asked to “use plain, everyday words and short sentences” at an 8th grade reading level; a summary with a specific focus on practical impact; and a “careful” summary that focused on methodological limitations. In addition, we sampled articles stratified by the three types of causal claims—Direct, Implied, and Conditional—as well as

Descriptive (i.e. no causality claimed). Finally, we ran a separate condition in which GPT was given the full text of the article rather than the title and abstract alone. Stuart-Maxwell tests revealed no significant differences between full-text and abstract-based summaries for any prompt (all p s > 0.17); thus, we present results for abstracts only.

Fig. 6 shows that across this more diverse sample of articles and prompting strategies, GPT-4.1 summaries continued to contain unhedged causal language at a substantially higher rate (54.4%, 95% CI = [50.9%, 57.8%]) than in the abstracts themselves (34.0% 95% CI = [24.7%, 43.3%], $Z = 3.8$, $p = 1.0 \times 10^{-4}$, using a summary-tailored, causal language classifier). However, Figure 6 also shows two important sources of variation. First, although all types of abstracts yielded some amount of causal claims in the summaries, the prevalence varied sharply by the abstract's causal classification ($\chi^2(3) = 208.7$, $p = 5.7 \times 10^{-45}$): abstracts with direct claims were the most likely to generate causal summaries (90.5% (95% CI=[86.4%,94.6%]) followed by abstracts containing implied (59.5% ([52.7%,66.3%])), conditional (48.0%, ([41.1%,54.9%])), and descriptive (19.5% ([14.0%,25.0%])) claims respectively. Second, the prompt also had a strong and systematic effect ($\chi^2(3) = 95.7$, $p = 1.3 \times 10^{-20}$). Figure 6a, shows that for the basic summarization prompt, 56.0% of summaries contained unhedged causal claims (95% CI=[49.1%,62.9%]). Next, Figs. 6b and c show that this rate increased for the simplified summary condition (64.5%, 95% CI=[57.9%,71.1%]) and increased even more for the summary emphasizing practical applications (71.0%, 95% CI=[64.7%,77.3%]). Finally, the rate dropped to 26.0% (95% CI=[19.9%,32.1%]) in the "careful" summary condition, which prompted GPT to focus on methodological limitations, although we note that even the careful prompt did not significantly reduce the prevalence of unhedged causal language ($p = 0.14$, Z -test); summaries for this prompt largely preserved causal claims when the abstracts stated them directly and occasionally introduced them even when abstracts were only conditional or descriptive. In other words, Fig. 6 largely replicates and extends the main insights of Fig. 5: GPT-4.1 is prone to produce summaries of empirical work that state causality even when the evidence is purely associational; but it is also responsive to corrective interventions, which reduce but do not eliminate this tendency

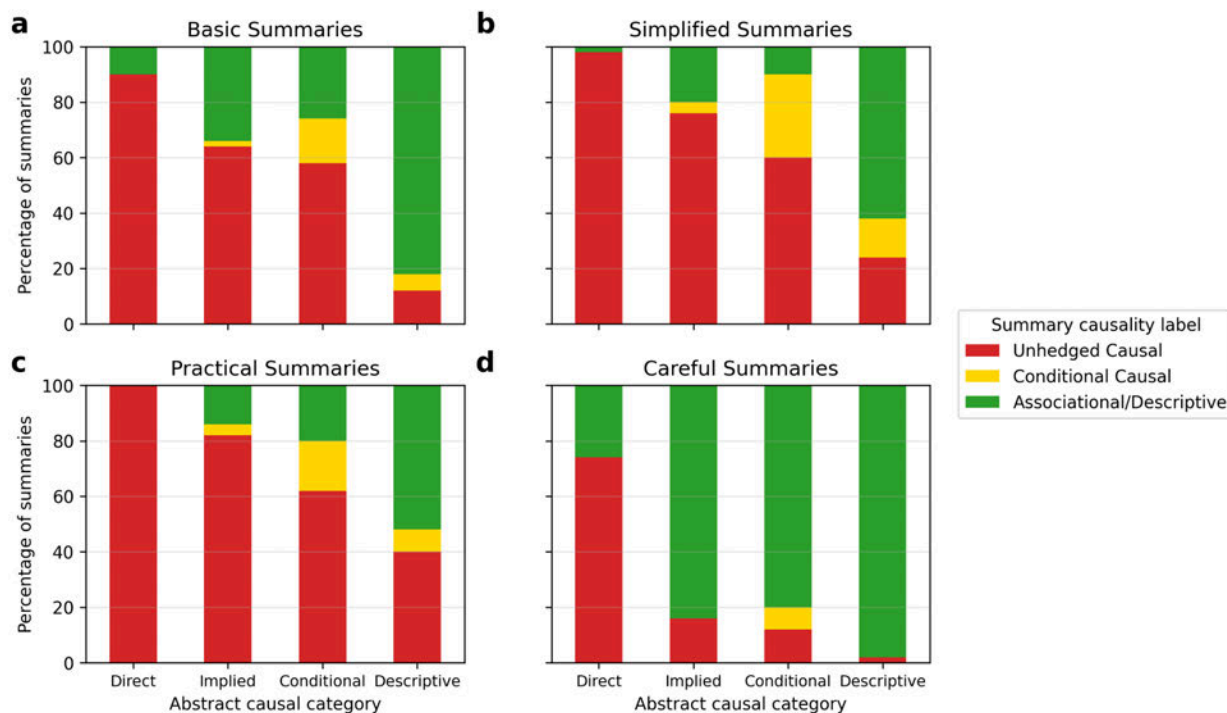


Figure 6. Proportion of GPT-4.1 summaries with causal claims by prompting strategy and abstract-level causal category. (a) With the basic summarization prompt, GPT frequently introduced unhedged causal claims across all abstract causal categories, including in more than half of the summaries for articles whose abstracts contained only conditional language. (b) A prompt requesting a simplified summary produced even higher rates of unhedged causal claims across abstract causal categories, even when the original abstract contained only associational language. (c) A prompt emphasizing practical implications increased the prevalence of these unhedged causal claims in GPT summaries even more. (d) A cautious prompt, asking GPT to write from the perspective of a skeptical reviewer, substantially reduced this tendency, though it did not meaningfully lower causal overstatement for abstracts that already overclaimed.

To the SI, we added our prompts:

SI Second Experiment with GPT-4.1- Prompts

simplified_summary = ""Summarize the text for a general reader (~8th-grade level). Use plain, everyday words and short sentences. Focus on the most important idea(s), avoiding jargon and numbers unless essential. Output one paragraph ~50 words (45–60). No preamble or bullets.

TEXT: ""

practical_summary = ""Summarize the text with emphasis on practical implications; describe the methods, and translate the findings into concrete takeaways or actions. Output one paragraph ~50 words (45–60). No preamble or bullets.

TEXT: ""

basic_summary = ""Give a concise summary capturing the central question, key finding/argument, and the minimum context needed to understand it. Keep wording neutral and factual. Output one paragraph ~50 words (45–60). No preamble or bullets.

TEXT: ""

careful_summary = ""Summarize this text from the perspective of a critical peer-reviewer. State the key findings and any important limitations (e.g., overclaiming, bias, confounding, generalizability). Output one paragraph ~50 words (45–60). No preamble or bullets.

TEXT: ""

We also added details about how we created the classifier for these summaries in the SI (Section “Second Experiment with GPT-4.1 – Causal Classifier”). Further, we added additional robustness tests that confirm the key findings do not depend on our specific causal language classifier (SI Section “Second Experiment with GPT-4.1 - Robustness Checks”).

Finally, we conducted a claim-level analysis in SI section “Second Experiment with GPT-4.1 - Claim Level Analysis.” Briefly, we found that GPT summaries tend to emphasize causal framing even when most underlying claims are descriptive. This tendency increases substantially as the proportion of unhedged causal claims rises, with even a modest share of direct claims sharply increasing the likelihood that GPT produces a unhedged causal statement. Using dummy data and the model predictions, we find the following:

“For an abstract containing three descriptive-only claims, the model predicts a 41.4% probability that GPT’s summary includes an unhedged causal statement. If one of the three claims is direct (33%), the probability rises to 61.6%. For full texts, when ten claims are all descriptive, the predicted probability is 46.1%; if one-fifth are direct, it rises to 58.4%, and if half are direct, to 74.7%.”

Overall I think this is an interesting paper and congratulations on finding this neat angle to study the overuse of causal claims!

Thank you! We hope you find the updated draft of substantially higher quality with the updated analyses, and look forward to your additional feedback.

Nature Human Behaviour

Reviewer expertise:

Reviewer #1: Science of science

Reviewer #2: Econometrics, experiments

Reviewer #3: NLP, Computational Social Science

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The paper is much improved, and I only have a few additional minor suggestions, mostly about narrative and presentation:

1. The figures are not colorblind accessible; the bars should have fill patterns (especially when it is just red and green), and the lines should have line styles.

Response: Our apologies, we updated the figures with accessibility in mind, adding fill patterns, varied markers, and line styles.

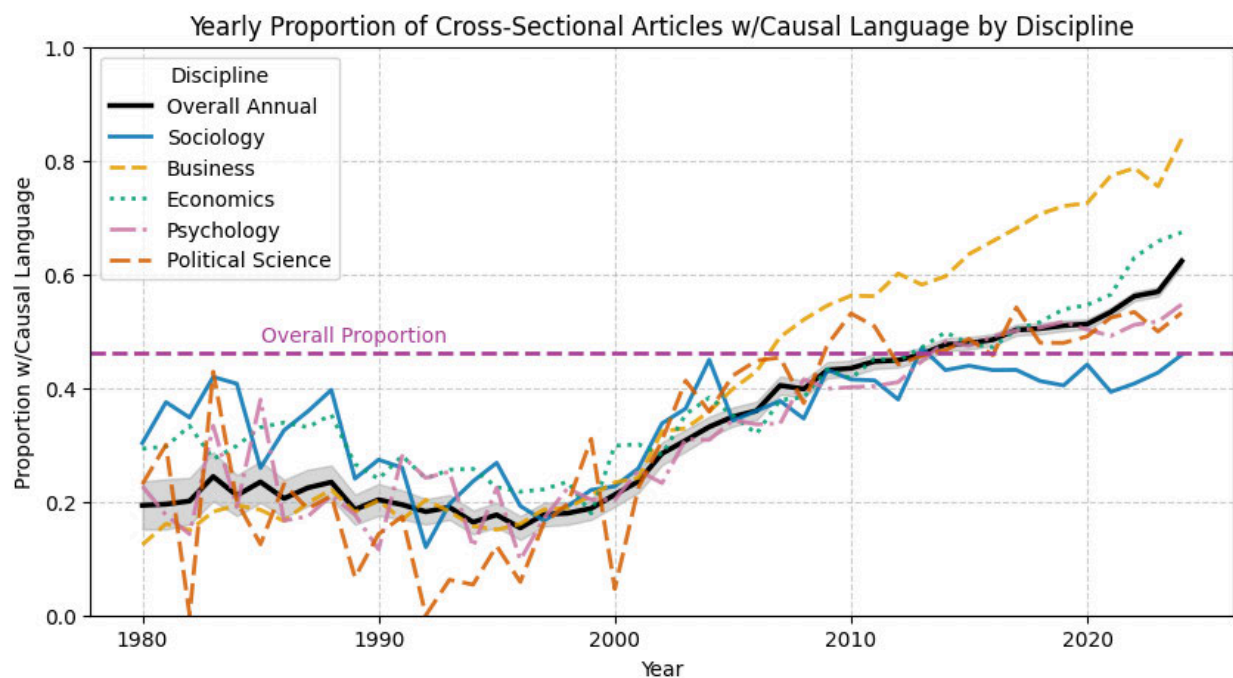


Figure 2

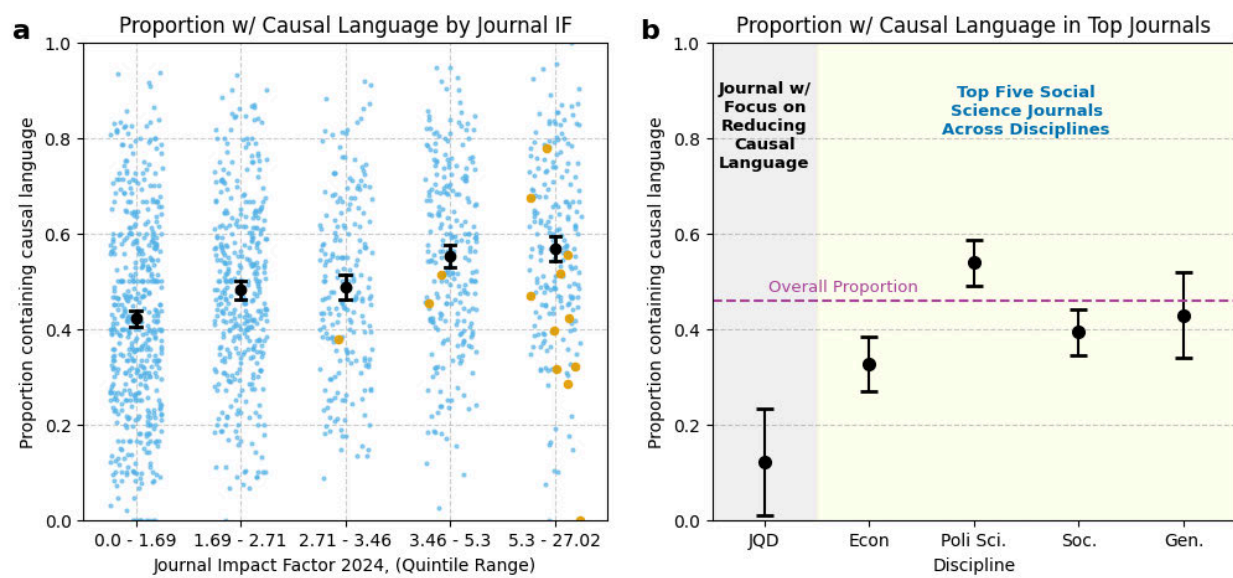


Figure 3

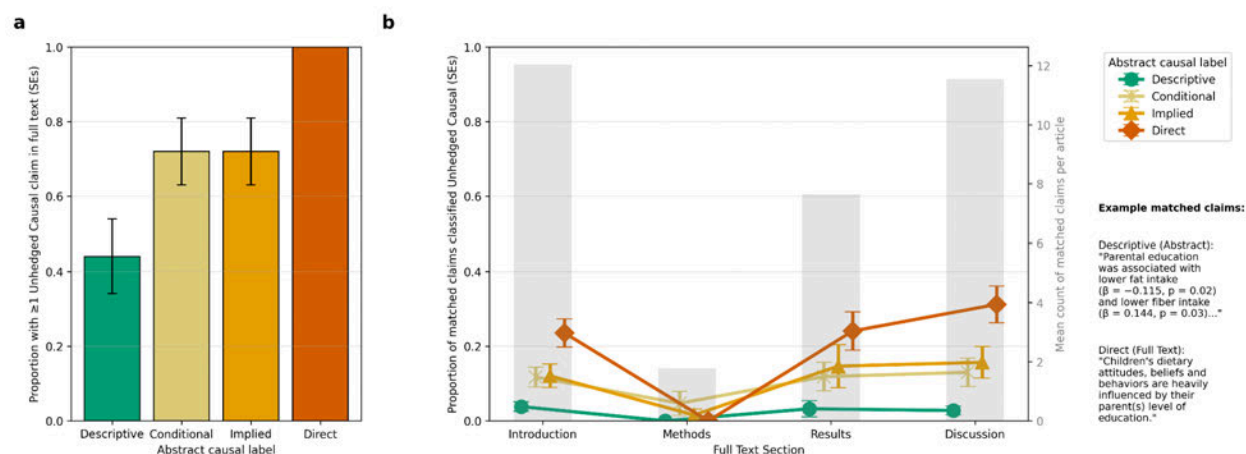


Figure 4.

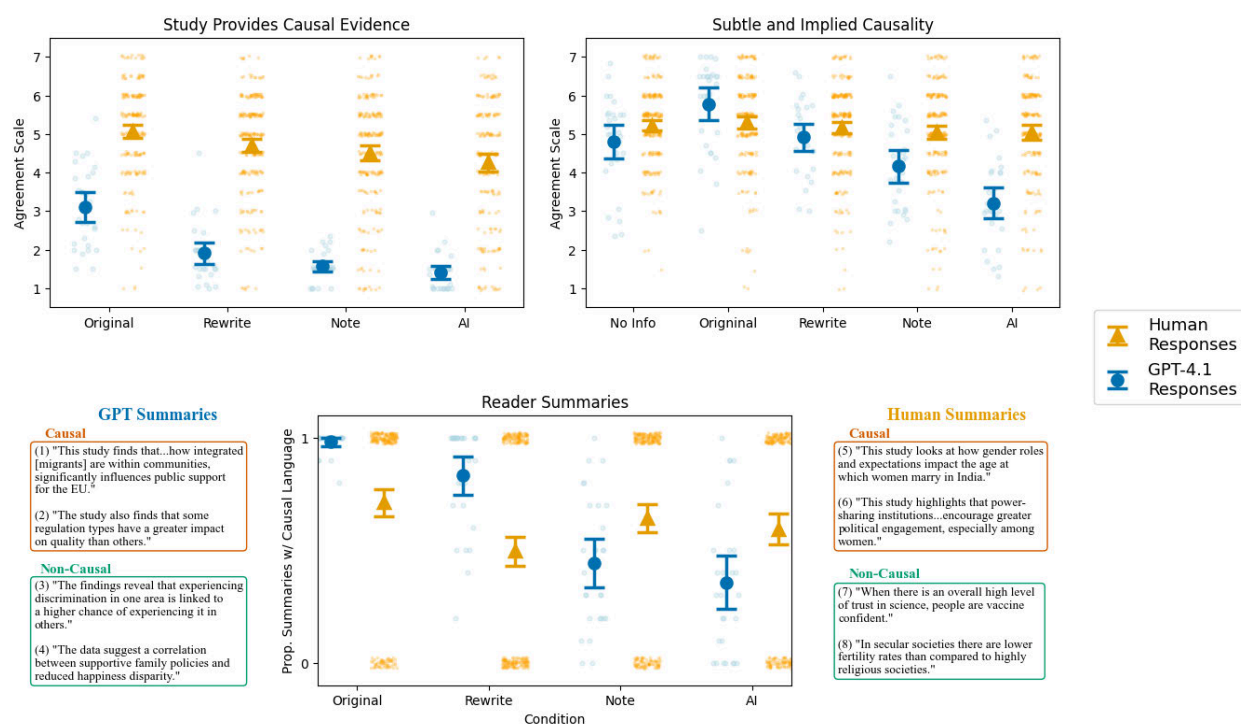


Figure 5

2. Have the authors run their own detection method on this manuscript? The authors should report whether the tool flagged any causal claims that were based on their cross-sectional analysis. This could make for an interesting discussion caveat, if it turns out to be hard to avoid slipping into a causal-implying mode, even for the authors.

Response: We appreciate this idea, and so we applied our full-text analysis classification strategy to the portions of the paper that were exclusively on the cross-sectional analysis.

There was one finding-claim made in the abstract about this analysis: “Over the period 1980-2024, an average of forty-six percent of articles contained causal language in their titles or abstracts, where the annual rate has risen almost three-fold since 2000 from 20% to 60%”. This was classified as non-causal by our classifier. It matched to 5 claims in the full text, of which all but one were also labeled non-causal.

The claim that was classified as causal was: “Figure 2 shows that from 1980-2024 our classifiers coded 46.3% of the 194,631 abstracts as making causal claims about the paper’s empirical results.” This appears to be a false positive, driven by our inclusion of the word “causal” in the sentence: when we change the descriptor from “causal” to “strong”, “universal”, or omit it entirely, this sentence’s classification label becomes non-causal. This illustrates a unique limitation of our classifier, which was designed on more general social science articles: on papers such as ours that are about the use of causal language or causality, it may be less accurate. Fortunately, we do not expect such articles to appear frequently in our data.

It is important to note that, given the topic of our analysis, we were very careful to search for and remove overreaching causal language. In earlier drafts, we had included causal claims that we needed to hone down or remove, and when we discuss the project with others, we sometimes find it difficult to avoid causal overclaiming. We attempted to illustrate this in the intro (“default toward causality”) and discussion (“...this finding is correlational, and likely reflects some degree of self-selection among submissions; yet it invites causal interpretation anyway.” along with the previous Stroop analogy discussed next.) We believe causal overclaiming is indeed very challenging to avoid, even for us, and hope that is clearly communicated as one of the key takeaways of the paper.

3. I don't like the Stroop task analogy. It implies a particular mechanism (cognitive load) that isn't actually born out by any evidence here. It could be that authors introduce causal language deliberately and strategically, not because they're tired. (Proving once again that narrative license is hard to avoid!)

Response: Upon further reflection, we take your point that the Stroop analogy was non-ideal. And although we acknowledge that there are alternative reasons why scholars may make overreaching causal claims, the point of this section was meant to align with your previous comment: that even for those trying very deliberately to avoid overclaiming, it can be a challenge. We modified our manuscript to remove the Stroop comparison and highlight this point.

- Previously: “As a result, when researchers hold strong theoretical priors, resisting causal interpretation can resemble a Stroop task \cite{stroop1935studies}: demanding cognitive effort and sustained, deliberate attention.”
- Now [Line 241]: “As a result, when researchers hold strong theoretical priors, resisting causal interpretation can be extremely challenging, demanding cognitive effort and sustained, deliberate attention to avoid.”

4. The Journal of Quantitative Description example is really interesting. To me this brings up the key counterexample to the whole paper, which is that I find it completely plausible for there to be a causal effect between the journal's policies and the extent of causal language used in its pages, and for this to be absolutely admissible evidence in favor of a policy intervention (that journals should restrict causal language). Don't you? I think this issue deserves a more prominent discussion.

Response: Although we agree that the finding for JQD may be indicative of such a policy's efficacy, it does not, alone, illustrate that these policies are effective. With this correlational analysis, all we can tell is that articles published in JQD tend to have less causal overclaiming than most other outlets. The source of this difference may come from these policies, and they may also emerge from other places. For example, two plausible other mechanisms include (1) self-selection from authors: the journal's focus may attract researchers more interested in descriptive work with carefully hedged claims whose articles are less likely to overclaim (2) self-selection from editors: editors in this venue may be particularly attuned for detecting over-reaching language and require edits for or reject papers with such claims; they may behave similarly with or without such a policy.

Our analysis offers little leverage to understand the counterfactual: how the same papers, reviewed by the same editors in a comparable venue, would read without a strict language policy. While we agree that editorial policies are promising targets for reducing causal overclaiming, this view rests primarily on prior reasoning rather than our data; the correlational finding that JQD exhibits less overclaiming only marginally changes those beliefs. We expanded our discussion of this point to highlight these important nuances.

- [Line 243] “Combating these mutually reinforcing mechanisms poses a formidable challenge, though publishing reforms offer perhaps the most obvious potential remedies. For example, The Journal of Quantitative Description, which explicitly discourages causal language, exhibited substantially lower rates of causal claims than most outlets. Ironically, this very observation illustrates the challenge we aim to highlight with this paper: this finding is correlational, and may reflect other mechanisms entirely (e.g., self-selection among authors focused on more careful claims, or editors already predisposed to methodological precision) yet it invites causal interpretation anyway. Nonetheless, we believe that clearer guidance from journals regarding permissible language could help realign scientific communication with methodological reality^{7, 66}, provided such standards were consistently enforced rather than merely stated.”

5. Relatedly, there is no magic bullet technique that grants someone narrative license, not even RCTs. I see Yarkoni cited in the introduction, but it would be useful to revisit that point in the discussion along with something like Deaton & Cartwright (2018):

<https://www.sciencedirect.com/science/article/pii/S0277953617307359>.

Response: We agree that experiments are not magic bullets. We edited our discussion on this point, adding your suggested references:

- [Line 265] “Second, our focus on cross-sectional studies, although helpful for creating clear criteria for identifying instances of causal overreach, does not imply that other designs are less susceptible to narrative license. For example, carefully executed cross-sectional evidence can, in some settings, be more informative for theory testing than, say, laboratory experiments with limited external validity. Indeed, RCTs carry a host of methodological limitations that constrain the inferential scope of their findings^{6, 77}. Our aim is not to criticize any single design or to single out any one manifestation of overclaiming, but rather to call attention to the broader phenomenon of narrative license and to introduce methods for identifying it. Future work should therefore examine other forms of narrative license within other research designs.”

6. "Prior work on statistical misperceptions [80] has found that experts are no less susceptible than lay participants" overstates that particular finding. Experts are also susceptible, but they may be (and one certainly hopes they are) less susceptible than lay participants.

Response: Thank you for catching this. We agree it was overstated and changed the language to match your description (and the title) from that citation.

- [Line 285] “Prior work on statistical misperceptions⁸⁴ shows that expert judgments vary with how data are presented...”

Conditioned on addressing these points, I am happy to recommend this manuscript for publication.

Reviewer #2 (Remarks to the Author):

The revised version of the paper addresses all my comments from the previous review round.

Response: We are glad to see that the reviewers' comments have been addressed and we thank them for their constructive feedback in the previous round of review.

Reviewer #3 (Remarks to the Author):

Thank you so much for revising the manuscript and the updated manuscript looks great!

Below are some suggestions and comments

1. For AI experiments, while I do believe the results on GPT 4.1 could be generalized to other models, it would be nice to add more LLMs since GPT 4.1 is only one (and outdated) model. The authors should use the latest LLMs when preparing the final version. This will help strengthen the paper, especially when it is being discussed by the public.

Response: We agree that assessing the generalizability of LLM overclaiming across models is important. In response, we replicated the experiment using several more recent models: GPT-5.1, GPT-5.4, DeepSeek-Chat-V3.2, and Claude Sonnet 4.6.

Overall, the main pattern of results largely replicates, with two notable exceptions. First, the basic summarization prompt only produces significantly higher rates of causal claims for two of the five models (after adjusting for multiple comparisons); GPT-5.4 summaries actually contain slightly fewer causal claims than abstracts, though not significantly so. Whether this reflects a broader improvement in OpenAI's models or a quirk specific to GPT-5.4 and these papers remains unclear. Second, the "Careful" prompt consistently reduces causal language across all new models tested. This pattern suggests that newer models may be increasingly sensitive to the limitations of the underlying research design and can appropriately modulate their claims when prompted to do so.

We have revised the manuscript to reflect these updated findings and Figure 6, which presents cross-model and prompt comparisons.

Updates to manuscript:

- [Abstract, last sentence] "Experiments with LLMs revealed that model summaries of these articles can amplify causal overstatement, removing hedges and introducing causal claims where articles used strictly associational phrasing; however, prompting caution diminishes this pattern."
- [Line 77] "To test this possibility, we further examined model summaries of a larger set of articles with varying types of causal language and evaluated how prompting strategies influenced different models' output, expanding the analysis across five leading LLMs."
- [Line 224] "Our experiments revealed that LLMs frequently introduced causal language when summarizing study findings, under basic, simplifying, and practical-oriented prompting strategies. This mirrors recent evidence that LLMs overclaim when summarizing scientific findings, for instance, by overgeneralizing.⁵² However, LLMs proved substantially responsive to intervention. Simple textual augmentations, study contribution labels, and a "careful" prompt requesting caution in summarization all substantially reduced the presence of causal language."
- [Line 289] "Moreover, our experiments with language models reflect only a few current models under four prompt configurations. Other models, prompts, or parameters may yield different outcomes, and we encourage continued evaluation of rhetorical overclaiming as new systems are deployed. Indeed, GPT-5.4 no longer showed the pattern observed in GPT-4.1; further testing will be necessary to determine whether this trend will persist in future models."
- [METHODS UPDATES, Line 619] "**Second Experiment with Large Language Models**

Our second experiment with LLMs explored whether the probability of a direct causal claim appearing in model summaries varied by the abstract-level causal classification, the summarization prompt, and whether the model was provided with the full text rather than only the abstract.

For our sample, we used the stratified sample of 100 articles from the “Full Text Claim Level Analysis”, which were evenly divided by abstract-level causal classification (Direct, Implied, Conditional, Associational/Descriptive). This design allowed us to test whether models added stronger causal language for articles whose abstracts were conditional, and whether they introduced causal claims when abstracts were purely associational from the start. We created four prompts representing diverse real-world uses of language models for interpreting academic texts: (1) a basic summarization prompt, (2) a careful prompt encouraging caution and skepticism, (3) a simplified prompt requesting a summary at an 8th-grade reading level, and (4) a practical prompt emphasizing real-world significance (See SI “Second Experiment with Large Language Models - Prompts”).

For each article and prompt, we submitted two requests to each model: one containing only the title and abstract, and another containing the full text, to test whether access to additional context affected causal framing in summaries.

We began with only GPT-4.1. Each request used a temperature of 1, a maximum of 150 tokens, and generated five outputs, with all other parameters set to the default GPT-4.1 Batch API settings as of November 4, 2025. Because new leading models were deployed throughout the peer review process, we replicated these results with subsequent models to test generalizability: specifically, OpenAI's GPT-5.1 and GPT-5.4, DeepSeek Chat-V3.2, and Anthropic's Claude Sonnet 4.6. For these models, we returned a single summary per query via the chat API and collected all responses from March 25 to 28, 2026.

For our outcome measure, we developed a new classifier tailored to these simplified summaries. As with our full text analysis, this classifier did not target the implied category, and we also relabeled the direct category as "unhedged causal", resulting in three categories (Unhedged Causal > Conditional Causal > Associational/Descriptive). Each output summary received a single label corresponding to the strongest causal claim type present, and the model classifier showed moderate alignment with the consensus human label on a held-out validation sample of summaries (Macro F1 = 0.75, Precision = 0.75, Recall = 0.79, Accuracy = 0.75, N = 240; see SI “Second Experiment with LLMs – Causal Classifier”). Notably, in cases where models and the consensus expert judgement mismatched, human coders were substantially more likely to label summaries as containing an unhedged causal claim than the model (Z test of proportions = 3.3, $p = 0.0008$). There were 44 cases where the humans labeled a summary as "Causal" when the model's label was either conditional or associational, and only 8 with the reverse. For GPT-4.1 responses, we aggregated across the five outputs and assigned the majority label to each article-prompt-content type; in cases of ties, we randomly selected one label from the tied categories. For all other models where we collected only one response, we assigned that label only.

Analytical approach: We first assessed whether summaries generated from full texts versus abstracts differed systematically in their causal classifications. The summary coding

scheme included three categories of causal language (i.e., Descriptive, Conditional, Unhedged). To preserve this structure, we report overall agreement, Cohen's κ , and p-values from Stuart–Maxwell tests to evaluate whether the marginal distributions of causal classifications differed across content types. We report both raw and false discovery rate (FDR)–adjusted p-values, analyzing every prompt-model configuration separately. These statistics are presented in Table S9. Agreement statistics should be interpreted alongside these distributional comparisons, as κ can be sensitive to differences in category prevalence.

For the main analysis, we collapsed the coding scheme into a binary distinction of whether the summary contained an unhedged (i.e., direct) causal language or not. This focuses the analysis on the most substantively consequential form of overreach. We then assessed whether model-generated summaries were more likely to contain such overstatements relative to the original abstracts. Because summaries generated from full texts and abstracts did not differ meaningfully, we restrict this section to abstract-based summaries to avoid double-counting.

We employed complementary analytical approaches. First, we present a complete set of 2×2 contingency tables comparing causal classifications in abstracts and summaries for every model/prompt pair (Table S10). For each configuration, we report overall agreement, differences in the proportion of summaries classified as containing unhedged causal language, and results from McNemar's test, including both raw and FDR-adjusted p-values (Table S11). These analyses provide a condition-specific assessment of whether certain models or prompts tend to inflate or attenuate causal overstatement relative to abstracts. Given the nested structure of the data, with multiple observations derived from the same article, assessing the statistical significance of these differences across models and prompts is complex. Accordingly, these results should be interpreted in conjunction with the multilevel models reported below, which explicitly account for this clustering.

To assess systematic differences across models and prompts, and to examine how causal language in abstracts relates to causal language in summaries, we estimated a multilevel logistic regression model using the lme4 package in R. In this model, the presence of unhedged causal language in the summary was predicted by the model, the prompt, their interaction, and the abstract's causal classification (Descriptive, Conditional, Implied, Direct). All predictors were treated as factors. We included a random intercept for article to account for repeated observations. Model estimates are reported in Table S12. Importantly, although this model enables clear comparisons across models, prompts, and abstract causal categories, it does not permit direct comparisons of causal language rates relative to the original abstracts, which is why we also reported the 2×2 contingency analyses described above.

We summarize these results at a high level in the main text, providing point estimates with confidence bands only in the SI out of space constraints.

We also conducted several robustness checks. To examine whether our results depended on our method for identifying causal language, we conducted sentence-level analyses on the responses produced by GPT-4.1, using our causal language classifier from the descriptive analysis, an independent Bert-based classifier trained on article conclusions \cite{yu2019detecting}, and a hedge detector, paralleling the procedures used in our full-text

analysis (See SI "Second LLM Experiment - GPT-4.1 Robustness Checks"), finding similar results. We also tested whether abstract claim mix (descriptive vs unhedged causal) predicts causal language in LLM summaries, and found that more causal language in the article yields more causal claims in summaries (See SI "Second LLM Experiment - GPT-4.1 Claim Level Analysis").”

- [RESULTS, Line 181] **“Second Experiment on LLMs**

The experiment just discussed had four important limitations: it relied on a single prompt; used a single model; did not account for the type of causal claim in the abstract; and relied only on abstracts rather than full texts. To address these limitations, we conducted a second experiment that employed four summarization prompts: a basic summarization, similar to the previous experiment; a simplifying summary, in which the models were asked to "use plain, everyday words and short sentences" at an 8th grade reading level; a summary with a specific focus on practical impact; and a "careful" summary that focused on methodological limitations. For our outcome, we used a new dedicated classifier tailored to these summaries. In addition, we sampled articles stratified by the three types of causal claims---Direct, Implied, and Conditional---as well as Descriptive (i.e. no causality claimed), and we tested 5 different models (GPT-4.1, GPT-5.1, GPT-5.4, DeepSeek Chat-V3.2, and Anthropic's Claude Sonnet 4.6). Finally, we ran a separate condition in which each model was given the full text of the article rather than the title and abstract alone. Stuart-Maxwell tests revealed no significant differences in causal language between full-text and abstract-based summaries for any prompt or model (all p 's > 0.09 , see Table S9); thus, we present results for abstracts only.

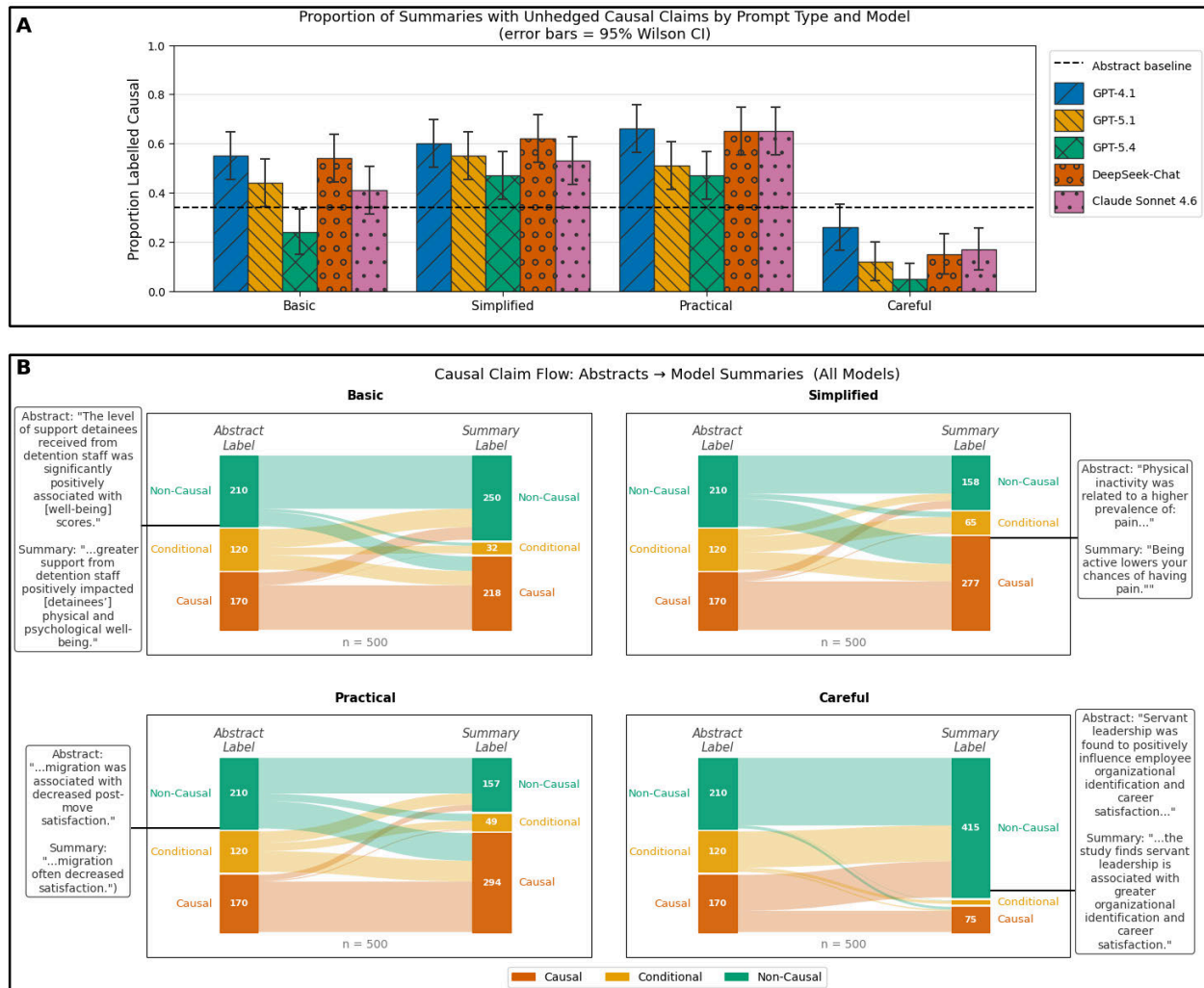


Figure 6. Panel A shows the proportion of summaries from each of the five models tested that contain unhedged causal statements, with the horizontal line representing the baseline rate in abstracts. Simplifying and practical prompts produced significantly higher causal claim rates across all models; basic prompts also led to higher causal claim rates, except for GPT-5.4. A careful prompt reduced causal language across models. Panel B tracks how claim types shift from abstracts to summaries. Conditional causal claims with more hedges and less certainty drop sharply in summaries. Example claim changes are presented which move from associational in the abstracts to causal in the summaries for each prompt except “Careful,” where the reverse is displayed.

Fig. 6a shows that, across this more diverse sample of articles, LLM-generated summaries contained unhedged causal language at substantially higher rates than the abstracts themselves overall, with notable variation across prompts (see Table 10 for full 2×2 tables and Table 11 for corresponding statistics). Under the basic summarization prompt, GPT-4.1 and DeepSeek-Chat-V3.2 produced significantly more causal claims than the abstracts, whereas the remaining models did not differ from the abstracts after correcting for multiple comparisons. The simplifying and practical prompts also yielded higher rates of causal

claims, this time across all models (all $p < .031$). In contrast, the careful prompt ($M = 15.0\%$) reduced causal language across all models except GPT-4.1.

A multilevel logistic regression confirmed that these differences across prompts and models were statistically meaningful when accounting for repeated measures by article (see Table 12). It further showed that abstracts containing any form of causal language (conditional, implied, or direct) were more likely to yield summaries with unhedged causal claims relative to purely descriptive abstracts, with predicted probabilities increasing with the strength of the original causal language. Fig. 6b illustrates how causal claims shift from abstracts to summaries across models for each prompt condition.

In short, Fig. 6 largely replicates and extends the main LLM insights of Fig. 5: language models appear prone to produce summaries of empirical work that overstate causality even when the evidence is purely associational; but they also are responsive to corrective interventions, including a “careful” prompt.”

[SI Additional Tables]

Table S8. Agreement of Causal Classifications for Summaries Generated when provided with the Full Texts Compared to Only the Abstract

Model	Prompt	κ	Agreement	Stuart–Maxwell p	p_{adj}
gpt-5.1	Basic	0.69	0.83	0.433	0.610
gpt-5.1	Simplified	0.56	0.74	0.005	0.094
gpt-5.1	Practical	0.49	0.7	0.758	0.846
gpt-5.1	Careful	0.59	0.92	0.485	0.610
gpt-5.4	Basic	0.49	0.78	0.799	0.846
gpt-5.4	Simplified	0.46	0.66	0.183	0.565
gpt-5.4	Practical	0.45	0.69	0.192	0.565
gpt-5.4	Careful	0.27	0.91	0.488	0.610
claude-sonnet-4-6	Basic	0.54	0.74	0.372	0.610
claude-sonnet-4-6	Simplified	0.48	0.71	0.220	0.565
claude-sonnet-4-6	Practical	0.39	0.7	0.049	0.373
claude-sonnet-4-6	Careful	0.23	0.81	0.086	0.432
deepseek-chat-v3.2	Basic	0.45	0.69	0.307	0.610
deepseek-chat-v3.2	Simplified	0.32	0.62	0.267	0.593
deepseek-chat-v3.2	Practical	0.54	0.77	0.873	0.873
deepseek-chat-v3.2	Careful	0.4	0.81	0.419	0.610
gpt-4.1	Basic	0.61	0.79	0.343	0.610
gpt-4.1	Careful	0.73	0.89	0.804	0.846
gpt-4.1	Practical	0.48	0.77	0.056	0.373
gpt-4.1	Simplified	0.64	0.81	0.226	0.565

Note. Overall, the distribution of causal language classification (descriptive, conditional, unhedged) did not differ meaningfully between summaries generated from full texts and those generated from abstracts alone. Consistent with this, all Stuart–Maxwell tests were nonsignificant after FDR correction (all $p_{\text{adj}} > 0.09$), and Cohen’s κ indicated moderate agreement across conditions. Note that in cases where κ is low, overall agreement remains high, suggesting that these values are driven by class imbalance rather than substantial disagreement.

Table S9. Contingency tables comparing abstract and LLM summary causal classifications

	Basic			Careful			Practical			Simplified			Total		
	Abs.	Sum.		Abs.	Sum.		Abs.	Sum.		Abs.	Sum.		Abs.	Sum.	
		0	1		0	1		0	1		0	1		0	1
GPT-4.1	0	43	23	0	62	4	0	32	34	0	38	28	0	175	89
	1	2	32	1	12	22	1	2	32	1	2	32	1	18	118
		Sum.			Sum.			Sum.			Sum.			Sum.	
claude-sonnet-4-6	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1
	0	52	14	0	62	4	0	30	36	0	41	25	0	185	79
	1	7	27	1	21	13	1	5	29	1	6	28	1	39	97
deepseek-chat-v3.2		Sum.			Sum.			Sum.			Sum.			Sum.	
	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1
	0	42	24	0	65	1	0	33	33	0	34	32	0	174	90
gpt-5.1	1	4	30	1	20	14	1	2	32	1	4	30	1	30	106
		Sum.			Sum.			Sum.			Sum.			Sum.	
	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1
gpt-5.4	0	48	18	0	64	2	0	43	23	0	39	27	0	194	70
	1	8	26	1	24	10	1	6	28	1	6	28	1	44	92
		Sum.			Sum.			Sum.			Sum.			Sum.	
Total	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1	Abs.	0	1
	0	244	86	0	318	12	0	186	144	0	197	133	0	945	375
	1	38	132	1	107	63	1	20	150	1	26	144	1	191	489

Note. Each cell displays a 2x2 contingency table comparing abstract (rows) and summary (columns) causal classifications. Values represent counts of articles.

Table S10. Comparison of Causal Language in Abstracts and LLM Summaries

Model	Prompt	Agreement	Δ Proportion	McNemar p	McNemar p_{adj}
gpt-5.1	Basic	0.74	0.1	0.0755	0.0809
gpt-5.1	Simplified	0.67	0.21	0.0003	0.0006
gpt-5.1	Practical	0.71	0.17	0.0023	0.0033
gpt-5.1	Careful	0.74	-0.22	1.05×10^{-5}	3.00×10^{-5}
gpt-5.4	Basic	0.76	-0.1	0.0639	0.0752
gpt-5.4	Simplified	0.71	0.13	0.0241	0.0301
gpt-5.4	Practical	0.77	0.13	0.0106	0.0142
gpt-5.4	Careful	0.69	-0.29	2.98×10^{-8}	2.45×10^{-7}
claude-sonnet-4-6	Basic	0.79	0.07	0.1892	0.1892
claude-sonnet-4-6	Simplified	0.69	0.19	0.0009	0.0014
claude-sonnet-4-6	Practical	0.59	0.31	7.84×10^{-7}	3.47×10^{-6}
claude-sonnet-4-6	Careful	0.75	-0.17	0.0009	0.0014
deepseek-chat-v3.2	Basic	0.72	0.2	0.0002	0.0004
deepseek-chat-v3.2	Simplified	0.64	0.28	1.94×10^{-6}	6.47×10^{-6}
deepseek-chat-v3.2	Practical	0.65	0.31	3.67×10^{-8}	2.45×10^{-7}
deepseek-chat-v3.2	Careful	0.79	-0.19	2.10×10^{-5}	4.66×10^{-5}
GPT-4.1	Basic	0.75	0.21	1.94×10^{-5}	4.66×10^{-5}
GPT-4.1	Careful	0.84	-0.08	0.0768	0.0809
GPT-4.1	Practical	0.64	0.32	1.94×10^{-8}	2.45×10^{-7}
GPT-4.1	Simplified	0.7	0.26	8.68×10^{-7}	3.47×10^{-6}

Note. These values correspond to the 2x2 contingency tables presented in Table S9. McNemar tests were used to assess whether the rate of causal language in model-generated summaries differed from that in the paired abstracts for each model–prompt condition. We also report differences in proportions (Δ Proportion), with positive values indicating that summaries contained more causal language and negative values indicating that abstracts did. p_{adj} denotes p-values after FDR adjustment for multiple comparisons.

Table S11. Multilevel Logistic Regression Predicting Unhedged Causal Claims in Summaries by Model, Prompt, Their Interaction, and Abstract Causal Category

	OR (95% CIs)
Intercept Odds (Reference: gpt-5.1, Basic prompt, Descriptive abstract; not OR)	0.06 [0.03, 0.16]
Model: Claude Sonnet 4.6	0.78 [0.36, 1.71]
Model: DeepSeek-Chat-V3.2	2.18 [1.01, 4.71]
Model: GPT-4.1	2.36 [1.09, 5.09]
Model: GPT-5.4	0.17 [0.07, 0.39]
Prompt: Careful	0.04 [0.01, 0.10]
Prompt: Practical	1.73 [0.80, 3.74]
Prompt: Simplified	2.36 [1.09, 5.09]
Abstract causal category: Conditional	6.04 [2.24, 16.31]
Abstract causal category: Implied	11.88 [4.38, 32.22]
Abstract causal category: Direct	178.92 [62.16, 514.96]
Claude Sonnet 4.6 × Careful	2.57 [0.70, 9.49]
DeepSeek-Chat-V3.2 × Careful	0.71 [0.19, 2.63]
GPT-4.1 × Careful	2.34 [0.66, 8.29]
GPT-5.4 × Careful	1.42 [0.29, 6.89]
Claude Sonnet 4.6 × Practical	3.79 [1.26, 11.36]
DeepSeek-Chat-V3.2 × Practical	1.36 [0.46, 4.03]
GPT-4.1 × Practical	1.36 [0.46, 4.04]
GPT-5.4 × Practical	4.41 [1.41, 13.79]
Claude Sonnet 4.6 × Simplified	1.09 [0.37, 3.24]
DeepSeek-Chat-V3.2 × Simplified	0.79 [0.27, 2.32]
GPT-4.1 × Simplified	0.62 [0.21, 1.84]
GPT-5.4 × Simplified	3.24 [1.04, 10.11]

Note. The intercept represents the baseline odds of unhedged causal language for the reference condition (gpt-5.1, basic prompt, descriptive abstract). All other coefficients are reported as odds ratios relative to this baseline (along with 95% CIs). To obtain predicted odds for a given condition, multiply the intercept odds by the relevant odds ratios (including interaction terms where applicable), and convert to probabilities as $p = \frac{\text{odds}}{1 + \text{odds}}$. A random intercept was included for article. For each of the 100 tested articles, 5 LLMs generated summaries using 4 separate prompts yielding a total $N = 2,000$. Fixed effects explain a substantial portion of variation in causal language ($R^2 = 0.54$), with additional variance attributable to article-level clustering ($ICC = 0.40$), yielding strong overall model fit ($R^2 = 0.74$).

2. Thanks a lot for the effort on studying the connections between uncertainty and causality expressions. While I did suggest this in earlier reviews, I find the current results a bit confusing and does not add too much to the overall narrative. One way to improve it is to directly look at claim-level uncertainty and causal expressions. Please consider using <https://pypi.org/project/certainty-estimator/> when assessing the certainty of scientific claims, as it is a fine-tuned model designed specifically for measuring the uncertainty of findings. It could also help to scale up the experiment on uncertainty and causality, and it could actually be a fairly straightforward experiment, given that you already have the model for causal expressions.

Please also consider adding more references regarding hedges and uncertainty expressions in science communication:

Hyland, Ken. "Writing without conviction? Hedging in science research articles." *Applied linguistics* 17.4 (1996): 433-454.

Friedman, Sharon M., Sharon Dunwoody, and Carol L. Rogers. *Communicating uncertainty: Media coverage of new and controversial science*. Routledge, 2012.

Gustafson, Abel, and Ronald E. Rice. "The effects of uncertainty frames in three science communication topics." *Science Communication* 41.6 (2019): 679-706.

While I do think a refined analysis on this could help, I also feel it is okay to leave this out of the main text to make a coherent narrative.

Response: (Un-)Certainty within claims is an important dimension of narrative license, and we appreciate the recommended references and package, which we have incorporated into the manuscript. We agree that more detailed certainty analyses would be valuable; however, due to space constraints, we have included them primarily in the SI with a brief reference in the discussion where we flag them as a productive direction for future work.

- *We added:* [Line 260] “Future research might leverage LLMs to systematically detect other forms of narrative license at scale. Hedging and certainty within claims represent a particularly promising target: both have been extensively studied \cite{hyland1996convict, friedman2012uncertain, gustafson2019frames} and validated detection methods already exist for scientific claims \cite{pei2021}. Preliminary analyses suggest that certainty metrics are only moderately correlated with causal overclaiming (see SI, “Causal Claims and Uncertainty”), indicating that these rhetorical elements operate somewhat independently and reinforcing the value of examining different types of narrative license.”
- *New Analysis (In the SI):* “Do causal claims carry more or less certainty than non-causal claims? Applying the certainty-estimator package \citep{pei2021} to all findings from our full-text analysis, we found that causal claims were expressed with reliably lower certainty ($M = 4.47$, $SD = 0.73$, $N = 847$) than non-causal claims ($M = 4.70$, $SD = 0.54$, $N = 2,287$), with a modest but significant effect size (Cohen's $d = -0.36$, $p < .001$). This difference was driven largely by conditionally causal claims ($M = 3.95$, $N = 360$), which showed markedly lower certainty than direct causal claims ($M = 4.86$, $N = 487$). Direct causal claims, in turn, exhibited higher certainty than non-causal claims. The limited overlap in certainty scores across these three categories suggests they represent meaningfully distinct phenomena warranting separate analysis. Figure S12 illustrates the level of certainty across causal claim types.”



Figures S12. Uncertainty scores for individual claims (light blue) and aggregated by their corresponding causal classification (black).

3. I'm not sure if Figure 6 does the best job of presenting the findings on LLMs' summarization. It might be helpful to have a river plot highlighting how many correlational claims of a paper got translated into causal claims through LLM summarization. It would also help to add a more illustrative example highlighting how LLM summarization pick/present causal claims. Another idea would be ask LLM to do deep research on some controlled topics and see how it picks and summarizes the results. I really think the LLM summarization experiment presents a potentially very impactful finding when so many people are using AI to do deep research (basically a deep summarization of literature). And I would really hope the authors could further strengthen this result.

Response: We agree that our previous Figure 6 could be improved. In response, we substantially revised the visualization. Specifically, per your suggestion, we now present a river plot that directly illustrates how claims shift from correlational to causal language through the summarization process (and vice versa), aggregated across all tested models. We incorporated concrete examples into this figure.

We appreciate the suggestion to examine this phenomena within “deep research”. We agree that this is an important extension. At the same time, we view this suggestion as beyond the scope of the present study, which also includes several findings outside the LLM results discussed here (i.e., findings about the prevalence of causal overclaiming in the literature and the impacts of those claims

on human interpretation). Accordingly, we have added the following to the discussion to highlight this direction for future work:

- [Line 292] “Future studies should examine whether LLMs similarly modify claims in more naturalistic usage contexts, such as multi-turn interactions or ‘deep research’ prompts on targeted topics, where models must select, integrate, and summarize evidence across multiple sources.”

Again, I would like to highlight that I really like this paper and I would love to see it being published. I hope the authors could consider my suggestions and further strengthen it to generate a larger impact in the AI world.

Response: Thanks for the helpful suggestions! We hope you find the paper appropriately strengthened and now suitable for publication.