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

Calvin Isch[1,*], Timothy Dorr[1], Neil Fasching[1], Grace Jennings[1], and Duncan J. Watts[1]

[1]University of Pennsylvania, 3451 Walnut Street, Philadelphia, PA 19104, USA
[*]calvinis@upenn.edu

ABSTRACT

Across the social sciences, many studies use cross-sectional designs that reveal associations but are generally unable to support direct causal claims; yet authors of such papers may make or imply causal claims anyway. To examine the prevalence of such "overreaching" causal language, we analyzed 194,631 cross-sectional articles using large language models. 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%. 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 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.

Introduction

Reliable and trustworthy science communication requires measured claims that accurately reflect empirical evidence[1,2]. In academic publications, this ideal translates into reporting research questions, methods, results, and implications in a way that avoids distortion and misleading readers[3,4]. Yet scientific papers are not purely objective accounts of facts. Researchers retain considerable discretion in how they operationalize constructs and rhetorically frame their findings[5–7], providing flexibility that also opens the door to misinterpretation. An important case of what we call "Narrative License" occurs when scholars claim or imply that correlational findings are causal without conducting the appropriate tests[8]. In such cases, readers seeking evidence-based solutions[9] may find themselves persuaded of causal claims that are not based on causal evidence.

Conflating correlation with causation can yield serious real-world consequences when such claims inform policy, medicine, or communication. For example, in political science, influential work argued that civil wars were driven not by poverty or ethnic grievances but by structural opportunities for insurgency[10,11]. Although correlational, these findings were widely characterized as causal, including in *The New York Times*, and were credited with shaping World Bank policy. Yet, subsequent work found that the models in question performed poorly at predicting outcomes[12], implying that interventions based on these models could misallocate resources. Similar problems have arisen in public health: observational analyses from the Nurses' Health Study linked postmenopausal estrogen therapy to reduced cardiovascular risk[13], but randomized trials later showed the opposite[14], with benefits limited to younger women[15]. Uncritical application of the original causal claims could result in interventions that do more harm than good and erode public trust. Similar overclaiming also appeared in media-effects research linking violent media to aggression[16–18], with the relationship between the two remaining empirically contested[19–22]. Such cases illustrate how correlational findings are often interpreted as causal[23,24], sometimes amplified by scholarly language.

Concerns about conflating correlation with causation are longstanding in the social sciences; however, the prevalence of such conflation has been difficult to establish. In contrast with clearly spurious claims (e.g., featured in humorous compilations https://tylervigen.com/spurious-correlations), most scholarly articles test theoretically plausible hypotheses and employ empirical designs that can support causal inference under at least some conditions. The resulting ambiguity makes it difficult to systematically evaluate when causal claims overreach, especially at the scale of whole literatures.

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.

Even with these restrictions, studying the prevalence of causal claims in published articles at scale poses challenges. Authors are often aware of methodological limits and may use associational language when describing results, yet still imply causality by framing their work as testing causal hypotheses or by endorsing implications that only follow under causal interpretation. We also wish to examine differences in prevalence across several social science disciplines, which differ in topics, terminology, and article structure, thereby complicating the analysis in ways traditional text-analysis methods[31–35] often cannot handle reliably.

To address these difficulties, we develop methods that leverage the contextual understanding of large language models (LLMs). LLMs have been shown to perform well on diverse text-analysis tasks[36], often matching or exceeding state-of-the-art models and crowd workers[37,38] even with zero-shot prompts[39]. Still, expert validation remains essential[40,41], especially as LLMs are increasingly integrated into research systems[42–44].

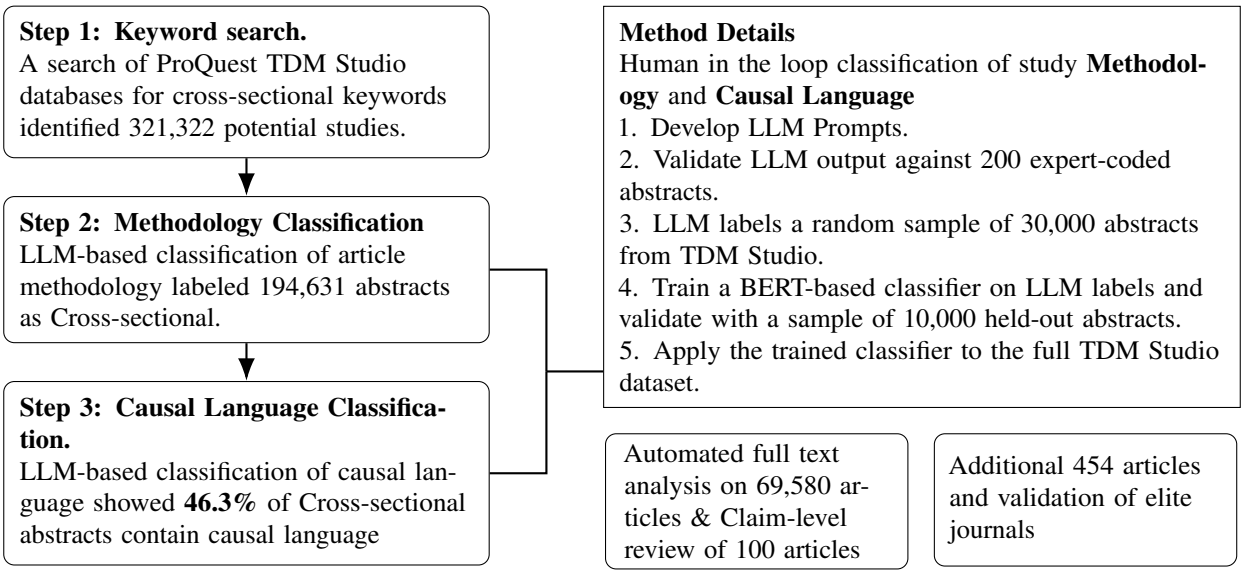

Figure 1. Overview of the three-step LLM-based workflow used to identify cross-sectional studies and detect causal language across five ProQuest databases. Step 1: We queried databases in the ProQuest walled-garden, data-mining platform TDM Studio using keywords associated with cross-sectional methods and exported 40,000 titles/abstracts to develop and evaluate models. Step 2: To verify that articles were truly cross-sectional, we designed a GPT-4o prompt and validated it against 200 expert-coded articles. The validated prompt was applied to all retrieved abstracts, and a fine-tuned BERT classifier (trained on 30k / evaluated on 10k) reproduced GPT and expert judgments. Applying this classifier to all keyword-identified article abstracts yielded 194,631 cross-sectional studies. Step 3: Using a parallel workflow, we coded causal language in titles and abstracts. A fine-tuned BERT classifier again matched GPT and expert labels, allowing classification of all cross-sectional articles and revealing that 46.3% used causal language, a notably high rate given the limited identification leverage of such designs. We additionally analyzed full texts when available and supplemented the dataset with articles from general science journals not included in ProQuest.

Accordingly, we developed a three-step LLM-based workflow with human-in-the-loop validation to analyze social science articles, primarily focusing on titles and abstracts, which are likely to receive disproportionate reader attention. As illustrated schematically in Figure 1, we first queried five ProQuest databases using keywords commonly associated with cross-sectional methodologies. Second, we exported a sample of titles and abstracts and constructed an LLM prompt to classify study methodology as cross-sectional or not. The resulting labels closely matched expert human judgments and were reproducible using a fine-tuned BERT-based classifier. Applying this classifier to the full dataset, we identified 194,631 studies that relied exclusively on a cross-sectional design (i.e. did not contain a mixture of cross-sectional and longitudinal or (quasi-)experimental designs). Third, we used a parallel procedure to code causal language in titles and abstracts (see Table 1 for example claims). A fine-tuned BERT classifier again replicated GPT- and expert-generated labels, enabling us to classify all cross-sectional articles for the presence of causal language and to examine temporal trends and variation across disciplines and journals. Further, we analyzed the full texts of a subset of articles to explore whether this increased contextual information presented higher or lower

rates of causal claims, supplementing the dataset with additional full texts and general science journal articles not represented in ProQuest. See Methods for more details of all stages of our pipeline and robustness checks.

ID	Quote
521097004	"...an unfavorable effect of 'time restriction' on well-being is expected...[we show that] the time restriction factor adds to the degree of exhaustion and the work-nonwork conflict, while time autonomy diminishes these outcome variables."
194739887	"Role of multihospital system membership in electronic medical record adoption...Multiple regression analysis was used to examine the impact of multihospital system affiliation on EMR level of adoption."
2847421536	"...we show that being in a dual-career household increases one's willingness and lowers the perceived risk of leaving their job and joining a startup venture."
1826809431	"[Our findings about the] moderation of self-regulatory efficacy suggests self-regulation minimises individual engagement in deviant acts. Thus, human resources managers in Nigerian universities should consider self-regulatory efficacy as a selection criterion when hiring academicians."
1966300327	"While prevalence studies have examined the incidence of problem gambling in other age groups, little attention has been paid to the impact of casino gambling on older adults. This study investigated the prevalence of casino gambling as a social activity for active senior citizens"
2135260445	"This study considers the effect of leadership and managerial constructs on lecturers' commitment to the newly implemented honours programmes in a Dutch University... visionary leadership and the perceived discussion culture on excellence are of large influence on lecturers' behaviour."
1647802301	"We investigate both the role of gender and feminism in [Friend's With Benefits] relationships at a United States college, and ask whether identification with feminist ideology impacts students' motivations and assessments of their relationships."
2627195225	"The disproportionate rates of police surveillance and encounters in many communities in the US may be contributing to inequities in health and violence."
1895058836	"It was found that only [stakeholder engagement/empowerment] has a direct positive impact on [project success]."
220966357	"The results show that policy ownership has a substantial, positive impact on service usage and costs, particularly for beneficiaries in fair or poor health. The greatest impact was found for policies that provide first-dollar coverage."

Table 1. Ten randomly selected examples of causal claims in cross-sectional titles and abstracts, as coded by our classifiers. IDs correspond to the unique GOID identifier from ProQuest, which can also be used to find article citation information in the supplemental data.

Beyond documenting the prevalence of causal language, we also examined its impact on readers. People often infer causality from associations[45,46], suggesting that overclaiming may stem as much from a cognitive "default toward causality" as from career or institutional pressures. To test how causal phrasing shapes interpretation, we ran experiments in which human participants read an abstract drawn from a stratified random sample of articles in elite journals within our dataset. Each abstract appeared either in its original form, containing causal language, or in a version rewritten to use purely associational phrasing. After reading, participants judged whether the study provided causal evidence, summarized findings (allowing us to code causal language), and evaluated hypothetical interventions that would only work if effects were causal. Following prior work[8,47], we also tested a brief methodological label and an "AI" warning that highlighted overclaiming as potential tools to improve reader inference.

Finally, given the increasing integration of LLMs into scientific workflows, and building on research that analyzes LLM outputs[48,49], we conducted a parallel experiment with GPT-4.1 to examine how its summaries shift under the same interventions. Using LLMs could exacerbate potential human misperception if, say, model summaries tend to present correlational evidence as causal, potentially generating summaries that are more misleading than the originals. 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.

Through this combination of large-scale descriptive analysis and controlled experimentation, we aim to contribute to metaknowledge[50] in three ways: first, by clarifying how often causal language appears in cross-sectional social science research; second, by assessing its practical consequences for audience interpretation and LLM summaries; and third, by testing potential interventions to improve communication.

Results

Prevalence of causal claims in academic articles

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. As we have just argued, cross-sectional designs are poorly suited to support such claims; we therefore describe these claims as "overreaching" or "inappropriate," and note that their near-half prevalence is strikingly high. Figure 2 also shows a sharp increase since 2000, rising from a relatively stable 20% to over 60% by 2024. 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"). Across five disciplines the trend is broadly similar, but business is highest (over 80% by 2024) and Sociology lowest, while Economics, Psychology, and Political Science sit near the overall average. Figure S1 shows article counts grew rapidly from 1980 onward, weighting the overall average toward recent years, and Figure S2 displays disciplinary counts.

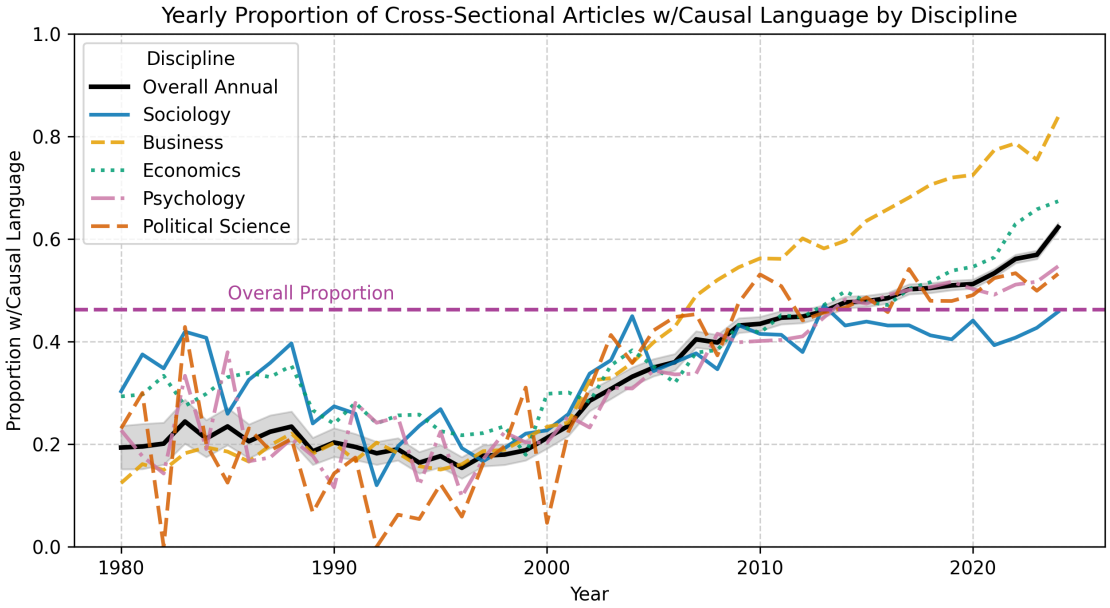

Figure 2. The yearly proportion of cross-sectional articles that contain causal language across all articles (black) and by discipline (colors). Overall, the proportion remained stable at roughly 20% from 1980 to 2000, then increased rapidly tripling by the 2020s. This increase appears across all disciplines, with notable recent variation: business shows the highest rates, while sociology shows the lowest. The gray band represents the annual 95% CI for all articles. The horizontal dashed line marks the pooled overall proportion across all articles (46.3%), which is disproportionately influenced by the higher volume of articles in later years.

We next examined variation in causal claims across journals with different prestige and impact. For this analysis, we focused on the 127,598 cross-sectional articles for which we could link SCImago Journal Rank (SJR) data. Interestingly, causal language was more prevalent among higher-impact journals. As Figure 3a shows, among journals above the median impact factor (IF > 3.03), 54.4% of abstracts contained causal language, compared to 43.3% in below-median journals. 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"). 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").

To further examine the relationship between journal prestige and causal language, we analyzed articles from five leading journals in each of four disciplines (economics, political science, sociology, and general social science; 20 journals total) and added a newer journal, the Journal of Quantitative Description: Digital Media, that explicitly disallows causal claims. For these journals, we manually validated both methodological classifications and causal-language labels (see SI, "Validating Elite Journals") and then plotted the proportion of abstracts containing causal language by journal group (Figure 3b). Across

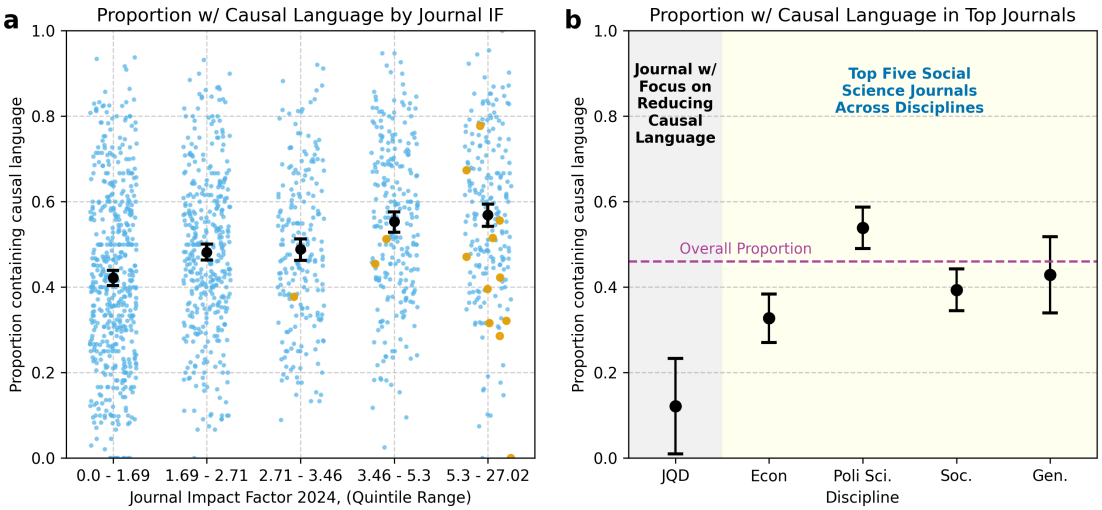

Figure 3. Proportion of cross-sectional articles containing causal claims, stratified by journal impact factor and discipline. (a) Proportion of cross-sectional abstracts with causal language by journal impact factor quintile. Blue dots represent proportions for individual journals with ten or more articles. Black dots represent the mean proportion across journals. Orange dots represent top journals from panel b. This plot shows that causal language is more common in higher impact journals. (b) The proportion of cross-sectional abstracts with causal language from five elite journals in four different disciplines (N = 20 total) and a new journal dedicated to using only descriptive language. The dashed line represents the proportion across all abstracts from all journals. Elite journals across disciplines exhibit substantial rates of causal language in cross-sectional articles, similar to the overall proportion. Error bars reflect 95% CIs

disciplines, elite journals showed rates of causal language (43.2%) slightly lower but comparable to those in the broader dataset. As expected, the Journal of Quantitative Description showed substantially lower use of causal language, although even here the rate was not zero.

Although claims made in the title and abstract are, by design, the most prominent claims made in a paper, the full text also offers authors many opportunities to assert claims that may reinforce, moderate, or even contradict their message. Using the same causal-language classifier, we analyzed full-text snippets from 69,580 cross-sectional articles available via TDM-Studio (see SI "Full Text Analysis"). As shown in Fig S7, 81.8% contained at least one causal snippet, well above title/abstract rates, suggesting authors often adopt causal framing in the main text even when abstracts are cautious. However, full texts contain many more total claims than abstracts, so the impact of these causal claims may be tempered by other associational statements.

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.

Impact on human interpretations and LLM summaries

Our descriptive analysis revealed the widespread prevalence of causal claims in cross-sectional work, but also raises the question of how this narrative overreach impacts reader inferences—and, if so, whether there are simple interventions to improve clarity. To answer these questions, we conducted an experiment to test how overreaching causal language in abstracts influences reader judgments. College-educated U.S. adults (N = 1,105) were randomly assigned to one of four conditions: (1) the original abstract containing a causal claim; (2) a rewritten version using strictly associational language; (3) the original abstract with a methodological note stating that the study was cross-sectional and therefore cannot on its own establish causality; or (4) the

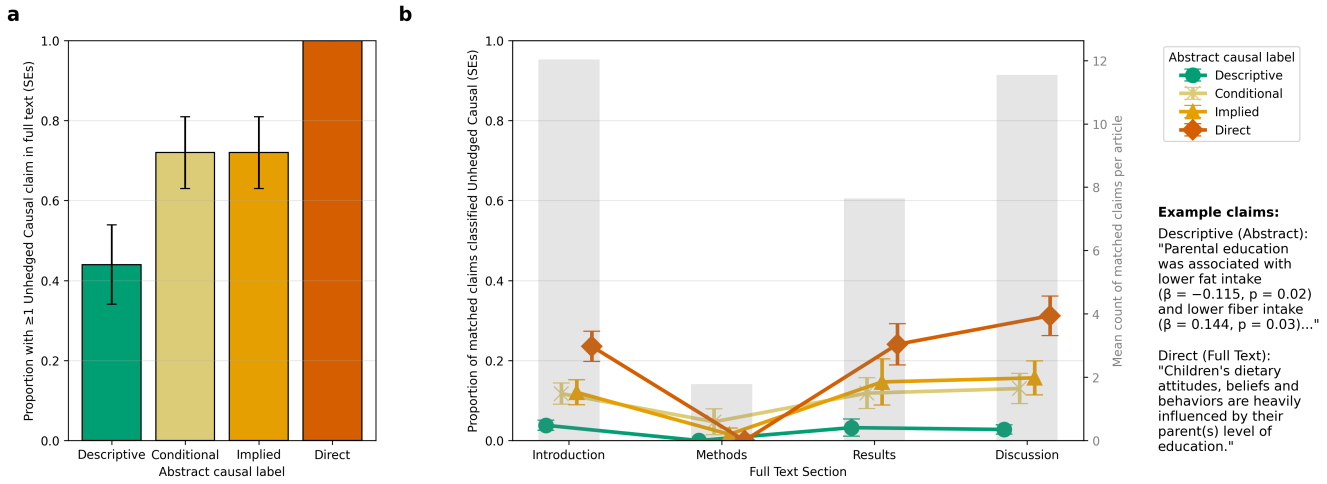

Figure 4. Causal claims present in the full text, grouped by article abstract-level classifications. (a) Proportion of articles with at least one unhedged causal claim in the full text that matched a claim made about the article's findings within the abstract. Even articles with abstracts that contain 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.

same note accompanied by "AI" feedback highlighting methodological and interpretive issues. Each participant read one of 28 abstracts sampled from elite journals, wrote a brief summary, and rated its causal implications (full preregistration, materials, and analytic models are available in the Methods and SI). The human-subjects experiment was preregistered, and we report one-sided p-values with Holm–Bonferroni corrections applied within each model family. A second experiment presented the same stimuli to GPT-4.1 to assess the influence of language on model output. We emphasize that these two experiments answer different questions: whereas the human experiment concerns how readers interpret causal claims, the experiment with GPT explores how AI-generated summaries might distort research findings.

Human participants

Overall, Fig. 5 (orange symbols) shows that human respondents have a strong latent tendency to make causal inferences regardless of the evidence, and that this tendency is resistant to corrective interventions but can be partially rectified for at least some outcomes (See Table S6 for all model estimates and confidence intervals).

First, Fig. 5a shows that in the original abstract condition, participants tended to agree that the studies provided causal evidence. In comparisons, participants who received a rewritten abstract with purely associational language, a methodological note, or "AI" feeback showed lower agreement, with medium effect sizes (standardized β's range from -0.25 to -0.55; p's $< p = 3.3 \times 10^{-3}$). Thus, removing or flagging causal language influenced reader interpretations, yet participants frequently inferred causal evidence regardless of condition, even when causal claims were absent or explicitly denied.

Second, Fig. 5b presents responses to a question about a hypothetical manipulation of the IV, implicitly invoking causality but not asking about the paper's evidence directly. Here, human respondents were even more likely to agree that the proposed relationship is causal and even less responsive to the interventions, with no significant difference observed across conditions (βs = -0.12 to -0.18; p's = 0.073 to 0.092). These results could arise either because humans are more likely to make causal inferences when asked about them indirectly or responses may reflect prior beliefs rather than paper-based evidence. To distinguish these, we included a "No Info" control condition in which participants responded without reading the paper. Responses did not differ significantly from those in the original abstract condition ($\beta = -0.07, p = 0.471$), suggesting that at least for these stimuli, latent beliefs played an important role.

Finally, Fig. 5c shows how likely participants were to use causal language when summarizing the paper's findings (see figure for example summaries). Participant summaries in the original-study condition were very likely to contain causal language.

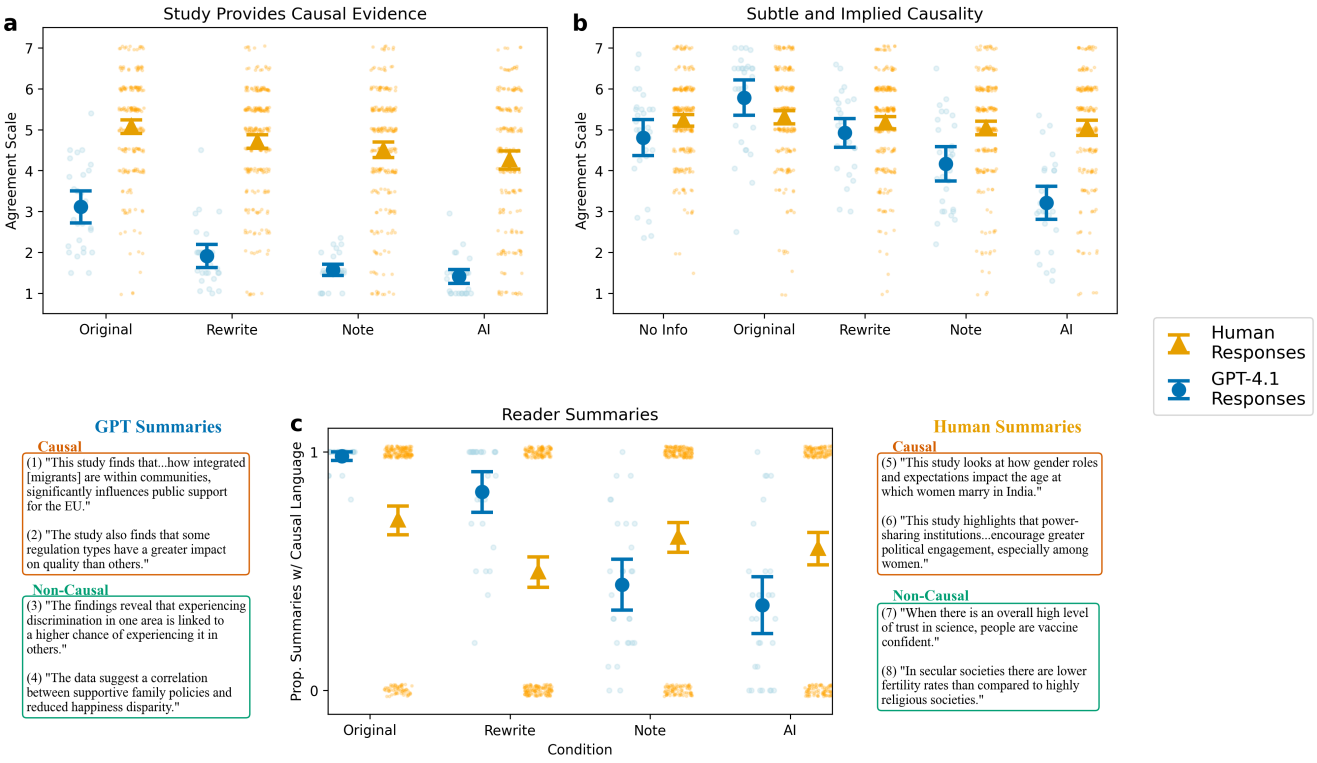

Figure 5. Causal beliefs indicated by human participants and in GPT-4.1 summmaries across experimental conditions. (a) Human participants were more likely to agree that the study established a causal relationship when reading the original abstracts compared to model outputs; both the rewrite and the two interventions significantly reduced agreement for human participants and in GPT-4.1 output. (b) GPT-4.1 output showed similar reductions on subtler items implying causality, and while human responses followed the same direction, they did not differ significantly across conditions. (c) Free-text research summaries from humans and GPT-4.1 frequently contained causal claims in the original condition; the interventions reduced this tendency, though the summaries continued to contain causal claims across conditions. GPT-4.1 produced 10 summaries per paper, so the model's causal-language measure ranges from 0 to 1, reflecting proportions rather than the binary judgments used for individual human responses. Example snippets from summaries with and without causal language are shown alongside the plot.

The Rewrite ($OR = 0.38$, $p = 6.5 \times 10^{-6}$) and AI ($OR = 0.60$, $p = 0.022$) conditions showed significant reductions in this tendency, and although the Note condition was in the expected direction, it did not differ significantly ($OR = 0.73$, $p = 0.078$) . In all conditions, participants were more likely than not to use causal language in their summaries, again suggesting that the tendency to impute causality to observed associations is strong by default and challenging to correct.

First Experiment on GPT-4.1

In addition to our human-subjects experiment, we conducted a second experiment in which we provided the same experimental stimuli to GPT-4.1. Overall, Fig. 5 (blue symbols) shows that the model tended to describe correlational findings as causal. This tendency was sometimes weaker and sometimes stronger than humans, but in both cases was much more responsive to our interventions. Specifically, Fig. 5a shows that in the original condition GPT output was substantially less likely than humans to report that the study provided causal evidence and also showed larger reductions in the Rewrite, Methodological Note, and AI Feedback conditions (β range = -1.23 to -1.75, all $p < 5.0 \times 10^{-16}$). In the AI condition, for which the effect was largest, model output almost always indicated that the study did not provide causal evidence. For the implicit causal inference outcome, Fig. 5b shows that GPT output was about as likely as humans to agree with the causal interpretation in the No Info condition and slightly more likely in the original condition. As with the explicit question in Fig. 5a, however, the Rewrite, Note, and AI Feedback interventions all significantly reduced this tendency (β range = -0.60 to -1.79, all $p < 5.0 \times 10^{-8}$). Finally, Fig. 5c shows that model-generated summaries based on the original abstracts almost always included causal claims (99.3%), but that once again all three interventions showed significantly lower likelihoods (Rewrite 89.8%, Methodological Note 41.7%, AI Feedback (30.3%), all $p < 3.0 \times 10^{-8}$). Notably, the Note and AI treatments had quite large effects, but no intervention fully eliminated causal framing (see Table S7 for all model estimates).

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 $p > 0.09$, see Table S9); thus, we present results for abstracts only.

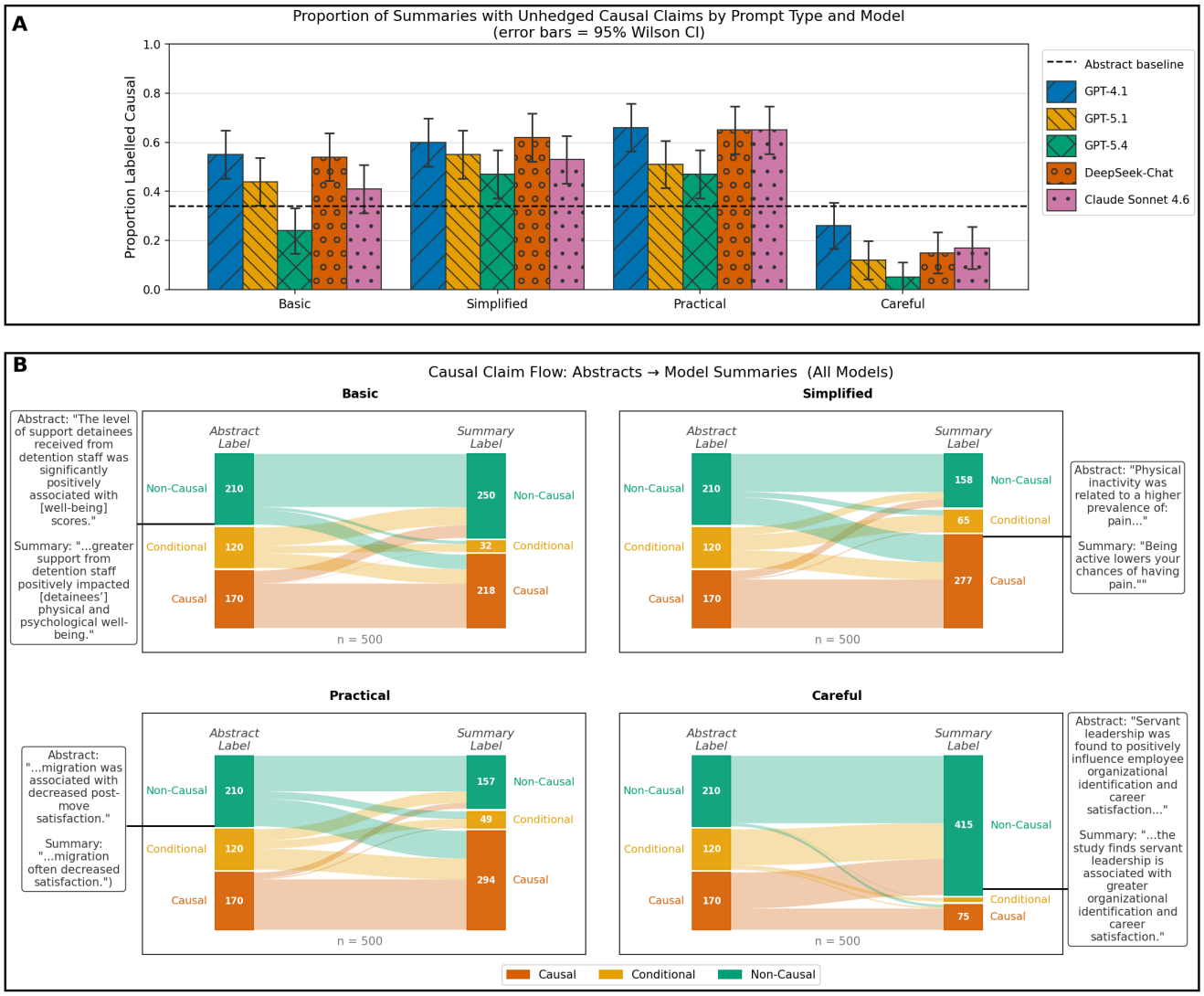

Figure 6. Panel A shows the proportion of summaries from each of the five models tested that contain unhedged causal statements, with the horizontal line representing the baseline rate in abstracts. Simplifying and practical prompts produced significantly higher causal claim rates across all models; 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 S10 for full 2×2 tables and Table S11 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 S12). 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.

Discussion

In this paper we have demonstrated that overreaching causal language is pervasive across the social sciences. Restricting our analysis to nonexperimental studies that rely exclusively on cross-sectional data, we find that nearly half (46.3%) of such papers published since 1980 contained causal claims in either their titles or abstracts. Even this large number understates the prevalence in recent years, which is over 60%, reflecting a threefold increase since 2000—a trend that occurred across all disciplines, albeit to different extents. Interestingly, these results applied equally to the most prestigious and selective journals and somewhat more so to high-impact journals in general. Moreover, full-text analyses showed that even when abstracts used cautious, associational language, corresponding papers often included direct causal statements elsewhere. Together, these results, which align with other evidence of causal overclaiming in psychology[51] and biomedical science[35], indicate that rhetorical overreach is not limited to isolated cases or to low-status outlets but reflects a systemic feature of how correlational findings are communicated.

In addition to studying the prevalence of inappropriate causal claims, we also studied their effects on humans and language model summaries in a series of three experiments in which we also tested various interventions to improve reader inferences. For humans, we found that our interventions somewhat reduced the share of respondents who agreed that the paper in question provided causal evidence, but that respondents continued to perceive causality at surprisingly high rates in all conditions. Even when abstracts were rewritten in strictly associational terms or corrective notes were added to the text, participants often reintroduced causal phrasing in their summaries. These findings suggest that causal interpretation is deeply ingrained in humans and is difficult to eliminate even with more careful phrasing. 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. These results raise concerns about the integration of LLMs into research and communication workflows, but also suggest that simple prompting modifications can ameliorate potential misinterpretations within model summaries.

What might explain the widespread use of causal language observed in this study? Although we do not test mechanisms, we speculate that several interdependent factors likely contribute. First, universities' hiring and promotion practices heavily prioritize the number and status of candidates' publications[53,54]. Second, journals, especially the highly selective journals that are most prized by hiring and promotion committees, often explicitly select for "novel" and "important" findings–criteria that encourage rhetorical overclaiming[55–58]. Our finding that higher-impact journals contain more causal language is consistent with this interpretation. Third, funding agencies' increasing emphasis on "impact"[59,60] may also incentivize causal framing as a way to signal relevance. Fourth, questionable research practices remain widespread in the social sciences[61,62], potentially normalizing overstatement and reducing individual incentives to resist. Finally, our results are consistent with prior findings that humans routinely infer causality from correlation[45–47], a tendency that may stem from a fundamental cognitive preference for causal, narrative explanations[63]. Scholars are presumably not immune to these biases[64], and this inclination may intensify by professional incentives or ideological commitments[7,65]. 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.

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. Our results also suggest that LLMs, while potentially a part of the problem, can also be part of the solution. For example, by flagging instances of causal overreaching, LLM-powered tools can assist authors, reviewers, editors, and readers alike to more clearly communicate and more accurately interpret scientific findings.

Before concluding, we emphasize that our study has several important methodological limitations, which also serve as motivation for future research directions.

First, our focus on inappropriate causal language is just one manifestation of verbal misrepresentation. As illustrated in Table 2, researchers have many rhetorical tools that can subtly distort findings—a broader phenomenon that we call "Narrative License," which captures the tendency to shape results into compelling stories at the expense of accuracy. Narrative License closely resembles prior work on "spin,"[4], but foregrounds narrative as the primary underlying mechanism. Much of the spin literature has focused on presenting null results as positive[67,68], but other forms have been examined[35,51,69,70] and found to alter reader judgments[71,72]. 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[73–75] and validated detection methods already exist for scientific claims[76]. 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, reinforcing the value of examining different types of Narrative License.

Table 2. Practices of Narrative License within social science articles

Practice	Definition	Examples
Selective Reporting	Authors emphasize findings that support their narrative while minimizing or omitting conflicting evidence.	- Citing only literature that supports the hypothesis - Highlighting aligned results, downplaying others - Entirely omitting complicating results
Unreported Deviation	Changes to methods or analysis plans are made without disclosure.	- Switching hypotheses after seeing results - Polishing or simplifying methods without transparency - Deviating from a pre-registered plan without explanation
Rhetorical Flourishes	Language is used to enhance appeal or obscure limitations.	- "Hype" or "boosting" language - Minimizing limitations through vague acknowledgments - Adding excitement or certainty through editorializing
Overgeneralizing	Claims extend beyond what the data directly support.	- Applying findings to broader populations than studied - Generalizing results to different settings or timeframes - Treating specific measures as if they reflect broad concepts
Overclaiming	Authors draw stronger conclusions than the evidence justifies.	- Interpreting statistical significance as practical impact - Describing small effects as robust or definitive - Inferring causal relationships from correlational designs*

* In this analysis, we focus on causal language within correlational designs, but this is just one form of rhetorical overreach, or Narrative License, which may appear in academic publications.

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. 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.

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[78,79]. 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[80–82]. We also note that humans are, at their core, storytelling creatures[63,83], 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.

Fourth, our human subjects experiment targeted college-educated participants but did not, to our knowledge, include practicing social scientists. Prior work on statistical misperceptions[84] shows that even expert judgments vary with how data are presented, but experts might respond differently in this specific case. Relatedly, our experiment focused on titles and abstracts, which omit context that careful readers might glean from full texts. Although our second LLM experiment found no difference between abstract-only and full-text inputs, future human-subject studies should examine whether full-text exposure affects causal interpretations. 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. Future studies should also 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.

Fifth, our study deliberately focused on academic publications as that is where practicing scientists have the most control, and hence the most accountability, over messaging. However, science communication unfolds across multiple interacting stages[85]; For instance, spin in journal articles can shape press releases[86], which in turn affect the accuracy of media reports[87]. With recent advances in LLMs, there is substantial opportunity to scale fidelity detection from prior work[88]. Future research could deploy LLMs across the full science communication pipeline to flag misaligned claims and better align interpretations with underlying evidence.

In conclusion, constructing coherent narratives around empirical findings can make it difficult to avoid causal overreach and related forms of overclaiming. Resisting these pressures requires sustained attention to what Feynman described as the first principle of scientific integrity: not to fool ourselves, despite being the easiest ones to fool[89]. This intellectual discipline is an essential task required to ensure the integrity of science and the gradual accumulation of knowledge.

Methods

We present our methodology in two parts: first, the procedures for the descriptive analysis of academic articles; second, the methods used in the preregistered human-subjects experiment and experiments with LLMs.

Ethical considerations

For the preregistered, human-subjects experiment, the study protocol was reviewed and approved by the University of Pennsylvania's Institutional Review Board (IRB #858559). Participants provided informed consent electronically and could withdraw at any time. Data were collected anonymously and stored in compliance with IRB guidelines. Participants were compensated $2.25, which slightly exceeded Prolific's fair compensation guidelines for the estimated study duration. The experiments with LLMs and the descriptive analysis were not pre-registered.

Methods for descriptive analysis

Open Science statement

This observational study was not preregistered. We share all analysis code, prompts, and derived article-level datasets (methodology and causal-language labels, metadata, and identifiers) on OSF (https://osf.io/zpua6/?view_only=e290bc25f1634205b3c9c809fbd4faeb). The underlying article texts are licensed through ProQuest TDM Studio and cannot be redistributed; however, our repository includes the exact search strategy and processing pipeline so that researchers with institutional access can reconstruct the corpus and replicate our analyses.

Data & sample

This study draws primarily on data obtained through ProQuest's TDM Studio, a text and data mining platform that provides access to the full XML content of articles across several ProQuest databases. We supplemented this dataset with articles from elite journals not available within the five databases analyzed (described below). As our main analyses focus on the TDM Studio articles, we describe that dataset first.

TDM Studio operates within a secure, browser-based workbench (often referred to as a "walled garden") which enables large-scale text mining of academic and journalistic content. The XML files include the title and abstract for all academic

articles, and occasionally full-text content as well. Within TDM Studio, we limited our search to five ProQuest databases that collectively represent a range of disciplines within the social sciences:

- *ABI/INFORM Global* provides comprehensive coverage of business, management, and economics research, including scholarly journals, trade publications, and market reports.
- *Worldwide Political Science Abstracts* indexes scholarly literature in political science, international relations, public policy, and related disciplines, offering insights into global governance and political theory.
- *International Bibliography of the Social Sciences* is a multidisciplinary database covering economics, sociology, anthropology, and political science, with an emphasis on international and interdisciplinary research.
- *PAIS Index* focuses on public policy, international relations, and social issues, indexing government documents, think tank reports, and academic literature on political and social affairs.
- *Sociological Abstracts* provides access to sociological research, including studies on social behavior, culture, demography, and social institutions.

To identify potential cross-sectional studies, we queried for relevant keywords ("cross-sectional" OR "crosssectional" OR "survey" OR "observational study" OR "prevalence study" OR "questionnaire") within the abstracts of all articles in these databases. This returned 672,944 total articles. ProQuest labels whether articles came from scholarly journals. Roughly half of the articles contained that classification (48.7%, N = 327,598). Our primary analysis is conducted on these scholarly articles that passed peer review; however, we include additional analysis on the full set of articles, which includes working papers and preprints, dissertations and theses, trade magazines, and conference submissions (See SI "Results for non-academic journals"). While each article had a unique identifier within these databases, a small portion (1.9%, N = 6,276) contained duplicate titles. These duplicates were dropped from our analysis, leaving a final sample size of 321,322 unique article abstracts. Figure 1 presents a visualization of how papers were collected and analyzed.

Automatic detection of study methodology and causal language

To identify cross-sectional studies and determine whether they used causal language, we developed a computational approach designed to operate within the constraints of the TDM Studio platform. A key limitation of this "walled garden" environment is that all processing must be conducted locally; it does not allow for easy sharing of content among coders or large-scale use of language models within its infrastructure (Note: After our analyses were conducted, TDM Studio opened access to OpenAI's API.). However, TDM Studio does support the export of small batches of metadata, including titles and abstracts, which enabled limited external processing. Since all articles contained this metadata, but only a subset contained full texts, we needed to build classifiers that relied solely on titles and abstracts.

To construct these classifiers, we followed the steps outlined in Figure 1. First, we developed coding procedures for humans and corresponding prompts for LLMs to identify two features: (1) whether the study used a cross-sectional design and (2) whether it employed causal language in describing its results. Human coders then rated a sample of 200 titles and abstracts to validate these classifications, alongside ratings from GPT-4o. We used this validation set to assess the performance of the language model. After validation, we applied the LLM to a larger dataset (N = 30,000) and used its outputs to train a local BERT-based classifier that could approximate the LLM's judgments. This classifier was then validated on a held-out sample (N = 10,000) and then deployed within TDM Studio to classify the full dataset. This approach allowed us to harness the contextual understanding of LLMs while operating within the platform's technical constraints. Both the language model and the local classifier achieved strong accuracy for both classification tasks, as described in detail below.

Identifying cross-sectional articles

Cross-sectional studies are observational research designs that examine empirical data collected from different entities. A key feature is that they do not track repeated measurements and do not involve any (quasi-)experimental intervention. To identify such studies, human coders independently reviewed the methodological design of 50 articles, consulting full-text content when abstracts were ambiguous (see SI "Instructions for Human Coding of Methodology"). Ultimately, each article was assigned a binary label: cross-sectional or non-cross-sectional. The coders achieved acceptable agreement (Krippendorff's $\alpha = 0.84$) and resolved discrepancies through discussion to reach a consensus on final labels. They then coded an additional 150 articles to create the validation set of 200 randomly-selected articles.

Using our coding instructions as a foundation, we developed a corresponding prompt for GPT-4o to classify study methodology (see SI "GPT-4o Methodology Prompt"). On our held-out validation set of 200 articles, the model showed strong alignment with human judgments (F1: 0.87, Precision: 0.84, Recall: 0.91, Accuracy: 0.85). We applied this prompt to the 30,000 titles and abstracts exported from TDM Studio and used these labels to train a BERT-based classifier (see SI "Training

BERT-based Models"). The classifier demonstrated strong predictive performance on a held-out validation set of 10,000 scholarly journal articles (F1 = 0.92, Precision = 0.91, Recall = 0.93, Accuracy = 0.91), indicating that it closely approximates GPT-4o's judgments and is suitable for large-scale automated coding. The trained model was then uploaded to TDM Studio and applied to the full dataset. Of the 321,322 unique, scholarly articles retrieved through our query, the classifier identified 194,631 as using cross-sectional designs. All subsequent analyses were conducted on this subset.

Identifying causal language in titles and abstracts

To identify causal language in the abstracts of the 194,631 cross-sectional studies, we followed the same procedure used to classify study design. We began by randomly sampling 50 articles, which were independently assessed by three expert coders to determine whether the title or abstract contained causal language about the article's novel empirical findings, excluding claims regarding context or background literature (see SI "Instructions for Human Coding of Causal Language" and Table S1 for examples). Causal language was defined broadly to include direct, implied, and conditional forms. For direct causal language, the authors explicitly claim that their results show one variable affects or causes another in their title or abstract (e.g., "We show that social media causes an increase in depressive symptoms among adolescents"). For implied statements, the abstract suggests that their study investigates a causal relationship (e.g., "We explored the role of social media use in shaping depression outcomes.") or proposes an intervention that entails causality without listing their assumptions or hedging their implication (e.g., "Based on our findings, individuals should limit their social media consumption to improve their mental health outcomes."). Conditional statements are hedged or qualified causal claims (e.g., "Social media may contribute to depressive symptoms in teens."). 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 $\alpha = 0.86$). Coders then classified 150 more articles to create the validation set of 200 randomly selected titles and abstracts, though ten articles were removed because they were in a language other than English.

Using our coding instructions, we developed a prompt for GPT-4o to classify the presence of causal language in an article's title and abstract (see SI "GPT-4o Causal Language Prompt"). GPT-4o showed strong alignment with the human-coded labels on our held-out validation set (F1 = 0.81, Precision = 0.91, Recall = 0.74, Accuracy = 0.78). Notably, the model's higher precision than recall suggests it tends to miss some causal claims, leading to a slight underestimation of their prevalence compared to human judgments. We then applied the GPT-4o prompt to the set of 30,000 titles and abstracts exported from TDM Studio and used the resulting labels to train a BERT-based classifier. This model demonstrated strong predictive performance on the held-out sample of 6,027 cross-sectional articles drawn from the 10,000-article validation set (F1: 0.83, Precision: 0.84, Recall: 0.82, Accuracy: 0.84), suggesting it effectively approximates GPT's judgments and is suitable for large-scale automated coding. The trained model was subsequently uploaded and deployed within the TDM Studio environment, allowing us to apply it to the full dataset of cross-sectional articles on the platform. Table 1 presents examples of causal language found in randomly sampled titles and abstracts from our full dataset (See also SI "Table with Examples of Causal and Descriptive Framing in Abstracts").

Robustness checks

To evaluate the sensitivity of our analysis to measurement strategy, we conducted two alternative approaches to identifying causal language, including keyword searches and a causal language model developed in prior research[35] and found even greater prevalence rates (See SI "Alternative approaches to assessing causal language in titles and abstracts").

Full text analysis

A subset of our cross-sectional articles from scholarly journals (N = 69,580, 35.7%) included full text in the XML. Although these articles are not randomly distributed, we aimed to determine whether their full texts were more or less likely to contain causal language compared to titles and abstracts. We required a solution that could be implemented in TDM Studio.

To achieve this, we first searched the full-text documents for a set of causal keywords (See SI "Full Text Analysis"). When a keyword was identified, we extracted a snippet of text spanning 100 characters on either side of the keyword to capture the relevant context describing the outcome. We opted for snippets rather than complete sentences because the XML data is highly variable and often too messy for sentence classifiers to work properly. Moreover, additional context is sometimes necessary to establish whether the causal language is relevant to new results or instead comes from previous work. Next, we used regex to

detect citation patterns within these snippets. Any snippet containing a citation was excluded from further analysis. Finally, we applied our BERT-based causal language identifier to these snippets.

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 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[90], 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[91], 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.[91], we classified pairs of claims with an Information Matching Score (IMS) greater than 3.0 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[92]). 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.

Supplemental data

We gathered information about every journal in our dataset using two complementary approaches. First, we used GPT-4o to classify the disciplinary focus of each journal based on the publication title (See SI "GPT-4o prompt to label journal discipline"). Second, we matched journal titles to the SCImago Journal Rank (SJR) database, a widely used platform that provides detailed bibliometric data for academic journals. The SJR database covered a substantial portion of the journals in our dataset (N = 127,598 cross-sectional articles), allowing us to incorporate standardized impact metrics in our analysis.

In addition to the TDM Studio dataset, we gathered articles from several high-impact general science journals that were not included in the ProQuest databases. Specifically, we collected content from *PNAS*, *Nature*, *Nature Human Behaviour*, *Science*, and *Science Advances* (See SI "External Data from Elite Venues"). For each journal, we filtered articles to ensure they focused on the social and behavioral sciences and included one of the keywords used to identify articles on TDM Studio (Final N = 454). Note: Our scraping of PNAS also includes articles published in PNAS Nexus following its launch in 2022. These sources helped ensure broader disciplinary coverage and representation of methodological variation within the behavioral sciences.

Given the prominence of these journals and their rigorous review processes, we conducted additional validation, both for them and for 15 other prestigious disciplinary venues included in TDM Studio, to ensure accurate identification of cross-sectional designs and causal language (See SI "Validating Elite Journals"). Because the sample from these venues was relatively small (N = 3,713 abstracts), we used GPT-4o for classification. Then we drew a stratified random sample of up to four cross-sectional articles per venue and had two authors manually review the full text, confirming that the majority (68%) met strictly cross-sectional criteria. These coders also labeled the presence of causal language in abstracts (blind to GPT's labels), and their judgments showed strong agreement with our automatic classifier (F1 = 0.82; Precision = 0.91; Recall = 0.75; Accuracy = 0.82), with notably higher precision than recall, indicating that GPT's labeling underestimates the rate of causal language relative to human judgments. Finally, we incorporated articles (N = 101) from a specialized journal with editorial policies that explicitly target limiting causal language in descriptive scientific work (*Journal of Quantitative Description*).

Methods for experiments

The human-subjects experiment examined the extent to which readers inferred causality from cross-sectional research abstracts that contained overreaching causal language, as identified in Part 1. Additionally, we tested three interventions: (1) a rewrite condition in which the abstracts were revised to retain the same content but use only associational language (2) a simple methodological label clarifying that the study was cross-sectional and does not establish causality, and (3) an "AI" feedback condition that combined this label with brief commentary and in-text highlights that drew attention to key methodological and interpretive features of the abstract. We were interested in the following research question and hypotheses.

RQ1 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.

Open Science statement

All procedures, materials, hypotheses, and analytic plans for this experiment were preregistered on AsPredicted (registration #238174; https://aspredicted.org/rhz3-jtg5.pdf) on the morning of 2025-07-14, prior to data collection, which began later that day. We followed this plan exactly, with no deviations from the protocol except for minor overrecruitment by the platform. De-identified data, analysis scripts, and stimuli files are deposited on the Open Science Framework for full reproduction of our results: https://osf.io/zpua6/?view_only=e290bc25f1634205b3c9c809fbd4faeb. The experiments with LLMs were not pre-registered.

Design

We recruited 1,100 adult participants through Prolific using the Wharton Behavioral Lab. Eligibility was restricted to individuals who hold at least a bachelor's degree and reside in the United States. A priori power analysis indicated that this sample size achieved over 80% power of detecting a modest effect size conservatively estimated based on pilot data results (See SI "Power Analysis"). The platform overrecruited an additional 5 participants, resulting in a final $N = 1,105$.

The experiment employed a between-subjects design with four main conditions. The factors were: (a) Original: the title and abstract containing causal phrasing. (b) Rewrite: rewritten title and abstract using strictly associational phrasing. (c) Methodological Note: a brief note highlighting the cross-sectional methodology with the original title and abstract. (d) "AI" Feedback: the methodological label above plus brief "AI" feedback and inline highlights that flag key areas which may impact reader inferences. While participants were informed that this was AI generated, we manually created this label and the highlights to ensure consistency across articles. Participants were randomly assigned to one of the four cells. See supplemental Figure S9 for an illustration of the experiment design. We also included an exploratory "No Info" condition for certain outcomes that did not require reading the study abstract. We randomized participants to conditions with 216 in the Original, 240 in the Rewrite, 229 in the Note, 202 in the AI, and 218 in the No Info condition.

We created a pool of 28 cross-sectional research abstracts drawn from our manual validation of articles in top-tier journals (See SI "Validating Elite Journals"). Articles were selected through stratified random sampling to ensure disciplinary diversity, and all original articles with both a cross-sectional methodology and any causal claim in their title or abstract were utilized as stimuli. For each article we prepared two matched versions, the original title and abstract and a rewritten version that used purely associational language, while keeping the content as similar as possible otherwise (See SI "Experiment Stimuli").

The primary dependent variable was participants' averaged responses to two questions about whether the study provided causal evidence, the first tailored to the abstract they read, and the second stated more abstractly: "Based solely on the evidence presented in this paper (not your prior beliefs or intuition), do you agree that this paper shows that {changes in IV} directly causes {changes in the DV}? (1 Strongly disagree - 7 Strongly agree)" "Do you agree that the study design used in this paper can establish a causal relationship between the variables of interest? (1 Strongly disagree - 7 Strongly agree)" The specific IV and DV terms were populated from the abstract assigned to each participant, ensuring the language was consistent with how the variables were described in the text.

In addition to our primary outcome, we collected two secondary outcome variables: (1) Free response item: "In a few sentences (at least 25 words) of your own words, summarize what this study shows about {Independent Variable} and {Dependent Variable}". Responses to this question were processed using a slightly updated causal language classification prompt with minor updates for these summaries, yielding a binary indicator of whether the participant used causal phrasing (See prompt in SI "First Experiment with GPT-4.1"). (2) Subtle and Implied Causal Inference: Averaged responses to two items, one about a hypothetical behavior/policy-intervention and another about their personal beliefs. "Given this information and using your intuition, do you agree that if {Group changes X} we can expect {Y to change} as a result? (1 Strongly disagree - 7 Strongly agree)". "Based on your intuition and the information provided, do you agree with the claim that {changes in IV} directly cause {changes in DV}? (1 Strongly disagree - 7 Strongly agree)" (See SI "Experiment Stimuli").

Procedure

After consenting, participants read brief background information about one randomly selected study, which provided definitions and other relevant details for comprehension. They were then provided with the associated title and abstract. Formatting (font, size) was identical across conditions, and all abstracts were PNGs to prevent copy and pasting. They then answered the outcome questions, always completing the summary first to ensure greater engagement, and then the other items were presented following a Latin square to account for order effects. They next provided demographic information about their level of (statistical) education. Finally, they had the option to leave comments before exiting the survey. Participants in the note and AI condition also answered a manipulation check question about the message, and participants in the exploratory "No Info" control condition also answered the Subtle and Implied causality questions immediately after the background but before reading the study abstract.

Participants who failed the attention check or submitted responses in under one-third the median pilot time were flagged. Consistent with the preregistration, no participants with complete data were excluded from the main analysis; flagged cases were retained and examined for robustness in sensitivity analyses (SI "Sensitivity analysis without flagged participants"). Median completion time was 8 minutes and 32 seconds; 11 participants were below the cutoff time, and an additional 25 participants failed the attention check, resulting in a total of 36 participant responses that were flagged (3.3%).

Statistical analysis

Analyses were conducted in R. Our primary model was a multilevel linear regression with random intercepts for article (equation 1). We were principally interested in main effects, though we report additional moderation analysis in the supplement (SI "Exploratory analysis for the experiment"). Our primary dependent variable was the average response to the two questions

about whether the study provided causal evidence, and we ran a parallel model with the subtle and implied causality questions as well as a similar logistic regression with the binary classification of whether the summaries contained causal claims.

$$Y_{ij} = \beta_0 + \beta_1 W_j + \beta_2 L_{1j} + \beta_3 L_{2j} + u_j + \varepsilon_{ij} \quad (1)$$

Y_{ij}: study provides causal evidence for participant i reading article j
W_j: Rewrite condition (0 = Original, 1 = Rewritten)
L_{1j}: Methodological Note condition (0 = no Note, 1 = Note present)
L_{2j}: AI condition (0 = not AI, 1 = AI condition)
u_j: random intercept for article j
ε_{ij}: residual error

Model assumptions were verified via residual plots, and the main results were found to be similar across specifications (See SI "Residual Diagnostics"). All preregistered hypotheses were evaluated with one-tailed tests at $\alpha = 0.05$ with a Holm-Bonferroni correction for multiple comparisons within each model family. In R, we employed the lme4 package for mixed-effects modeling[93]. In Python, we used statsmodels for statistical modeling and matplotlib for visualization[94,95].

Exploratory analysis

We were also interested in how both features of the stimuli and characteristics of the participants moderated causal inference. Specifically, we explored the plausibility of the causal relationship from each abstract (as rated by language models), the strength of causal language used in the original version of the abstract (as rated by statistical experts), and the complexity of the abstract's language, including both Flesch-Kincaid reading level and subjective expert judgments. We also considered participants' statistical education, based on a self-report question. We examined how these variables related to our three outcome measures using multilevel models with interaction terms (See SI "Exploratory analysis for the experiment").

First Experiment with GPT-4.1

Given the growing role of AI in both consuming and generating social science content, we conducted a parallel experiment on GPT-4.1, in which this large language model was presented with the same set of abstracts and questions as human participants. We gathered 10 responses from the AI agent for every article and condition and report their responses alongside those of human readers for comparison (See SI "First Experiment with GPT-4.1"). When querying the model, we used 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. Importantly, we do not assess AI judgments based on how closely they resemble human responses; rather, we examine them in their own right, given the increasing role of language models as mediators and communicators of scientific information.

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 $F_1 = 0.75$, 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 due to 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[35], 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").

Data and Code Availability

All analysis code, derived article-level data (including citation information), and experimental datasets required to reproduce the findings of this study are available on the Open Science Framework (OSF) at: https://osf.io/zpua6/?view_only=e290bc25f1634205b3c9c809fbd4faeb. Access to the full-text databases used for text and metadata extraction can be requested through ProQuest's TDM Studio platform (https://tdmstudio.proquest.com/home).

Acknowledgements

The authors thank members of the PennMAP group for their extensive feedback on this work, with special thanks to Amir Tohidi. We are also grateful to the customer support team at ProQuest's TDM Studio for their prompt and helpful assistance throughout the research process. We thank the Wharton Behavioral Lab for their support in running the experiment. This research was developed with funding from the Defense Advanced Research Projects Agency's (DARPA) SciFy program (Agreement No. HR00112520300). The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. Finally, CI gratefully acknowledges funding from the Institute for Humane Studies (Grant No. IHS019218). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author Contributions Statement

C.I., T.D., N.F., and D.J.W. conceived the study. C.I. and G.J. collected the data and conducted the analyses. C.I., T.D., N.F., G.J., and D.J.W. designed the experiment, which was conducted and analyzed by C.I. C.I. and D.J.W. wrote the manuscript, and all authors reviewed and approved the final version.

Competing Interests Statement

The authors declare no competing interests.

References

1. Kueffer, C. & Larson, B. M. Responsible use of language in scientific writing and science communication. *BioScience* **64**, 719–724 (2014).
2. Jamieson, K. H. 1 the need for communication: Communicating science's values and norms. *The Oxf. handbook science science communication* 15 (2017).
3. Wager, E. & Kleinert, S. Responsible research publication: international standards for authors. *Promot. Res. Integr. a Glob. Environ. Singapore* 309–16 (2010).
4. Boutron, I. & Ravaud, P. Misrepresentation and distortion of research in biomedical literature. *Proc. Natl. Acad. Sci.* **115**, 2613–2619 (2018).
5. Fletcher, R. H. & Black, B. Spin in scientific writing: scientific mischief and legal jeopardy. *Med. & L.* **26**, 511 (2007).
6. Yarkoni, T. The generalizability crisis. *Behav. Brain Sci.* **45**, e1 (2022).
7. Clark, C. J., Costello, T., Mitchell, G. & Tetlock, P. E. Keep your enemies close: Adversarial collaborations will improve behavioral science. *J. Appl. Res. Mem. Cogn.* **11**, 1 (2022).
8. Hofman, J. M. *et al.* Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
9. Head, B. W. Reconsidering evidence-based policy: Key issues and challenges (2010).
10. Fearon, J. D. & Laitin, D. D. Ethnicity, insurgency, and civil war. *Am. political science review* **97**, 75–90 (2003).
11. Collier, P. & Hoeffler, A. Greed and grievance in civil war. *Oxf. economic papers* **56**, 563–595 (2004).
12. Ward, M. D., Greenhill, B. D. & Bakke, K. M. The perils of policy by p-value: Predicting civil conflicts. *J. peace research* **47**, 363–375 (2010).
13. Stampfer, M. J. *et al.* Postmenopausal estrogen therapy and cardiovascular disease: ten-year follow-up from the nurses' health study. *New Engl. journal medicine* **325**, 756–762 (1991).
14. Rossouw, J. E. *et al.* Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *Jama* **288**, 321–333 (2002).
15. Lobo, R. A. Hormone-replacement therapy: current thinking. *Nat. Rev. Endocrinol.* **13**, 220–231 (2017).
16. Rubin, A. M., West, D. V. & Mitchell, W. S. Differences in aggression, attitudes toward women, and distrust as reflected in popular music preferences. *Media Psychol.* **3**, 25–42 (2001).
17. Willoughby, T., Adachi, P. J. & Good, M. A longitudinal study of the association between violent video game play and aggression among adolescents. *Dev. psychology* **48**, 1044 (2012).
18. Huesmann, L. R., Lagerspetz, K. & Eron, L. D. Intervening variables in the tv violence–aggression relation: Evidence from two countries. *Dev. psychology* **20**, 746 (1984).
19. Dou, Y. & Zhang, M. Longitudinal reciprocal relationship between media violence exposure and aggression among junior high school students in china: a cross-lagged analysis. *Front. Psychol.* **15**, 1441738 (2025).
20. Hilgard, J., Engelhardt, C. R., Rouder, J. N., Segert, I. L. & Bartholow, B. D. Null effects of game violence, game difficulty, and 2d: 4d digit ratio on aggressive behavior. *Psychol. science* **30**, 606–616 (2019).
21. Han, L. *et al.* The long-term effect of media violence exposure on aggression of youngsters. *Comput. human behavior* **106**, 106257 (2020).
22. Ferguson, C. J. Does media violence predict societal violence? it depends on what you look at and when. *J. Commun.* **65**, E1–E22 (2015).
23. Nunes, K. L., Hatton, C. E., Pham, A. T., Blank, C. & Maimone, S. Causal interpretations of correlational evidence regarding violence. *J. Interpers. Violence* 08862605241285996 (2024).
24. Nyhan, B. *et al.* Like-minded sources on facebook are prevalent but not polarizing. *Nature* **620**, 137–144 (2023).
25. Hill, A. B. The environment and disease: association or causation? (1965).
26. Schünemann, H., Hill, S., Guyatt, G., Akl, E. A. & Ahmed, F. The grade approach and bradford hill's criteria for causation. *J. Epidemiol. & Community Heal.* **65**, 392–395 (2011).
27. Rosenbaum, P. R., Rosenbaum, P. & Briskman. *Design of observational studies*, vol. 10 (Springer, 2010).
28. Angrist, J. D. & Pischke, J.-S. *Mostly harmless econometrics: An empiricist's companion* (Princeton university press, 2009).

29. Pearl, J. *Causality* (Cambridge university press, 2009).

30. Wooldridge, J. M. *Econometric analysis of cross section and panel data* (MIT press, 2010).

31. Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *J. language social psychology* **29**, 24–54 (2010).

32. Hutto, C. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, vol. 8, 216–225 (2014).

33. Gentzkow, M., Kelly, B. & Taddy, M. Text as data. *J. Econ. Lit.* **57**, 535–574 (2019).

34. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 4171–4186 (2019).

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

36. Bubeck, S. *et al.* Sparks of artificial general intelligence: Early experiments with gpt-4 (2023).

37. Rathje, S. *et al.* Gpt is an effective tool for multilingual psychological text analysis. *Proc. Natl. Acad. Sci.* **121**, e2308950121 (2024).

38. Gilardi, F., Alizadeh, M. & Kubli, M. Chatgpt outperforms crowd workers for text-annotation tasks. *Proc. Natl. Acad. Sci.* **120**, e2305016120 (2023).

39. Hsu, T.-Y. *et al.* Gpt-4 as an effective zero-shot evaluator for scientific figure captions. *arXiv preprint arXiv:2310.15405* (2023).

40. Pangakis, N., Wolken, S. & Fasching, N. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176* (2023).

41. Amirizaniani, M. *et al.* Developing a framework for auditing large language models using human-in-the-loop. *arXiv preprint arXiv:2402.09346* (2024).

42. Thirunavukarasu, A. J. *et al.* Large language models in medicine. *Nat. medicine* **29**, 1930–1940 (2023).

43. Demszky, D. *et al.* Using large language models in psychology. *Nat. Rev. Psychol.* **2**, 688–701 (2023).

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

45. Gershman, S. J. & Ullman, T. D. Causal implicatures from correlational statements. *PloS one* **18**, e0286067 (2023).

46. Xiong, C., Shapiro, J., Hullman, J. & Franconeri, S. Illusion of causality in visualized data. *IEEE transactions on visualization computer graphics* **26**, 853–862 (2019).

47. Law, P.-M., Lo, L. Y.-H., Endert, A., Stasko, J. & Qu, H. Causal perception in question-answering systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15 (2021).

48. Hewitt, L., Ashokkumar, A., Ghezae, I. & Willer, R. Predicting results of social science experiments using large language models. *Preprint* (2024).

49. Cui, Z., Li, N. & Zhou, H. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nat. Comput. Sci.* 1–8 (2025).

50. Evans, J. A. & Foster, J. G. Metaknowledge. *Science* **331**, 721–725 (2011).

51. Bleske-Rechek, A., Gunseor, M. M. & Maly, J. R. Does the language fit the evidence? unwarranted causal language in psychological scientists' scholarly work. *The Behav. Ther.* (2018).

52. Peters, U. & Chin-Yee, B. Generalization bias in large language model summarization of scientific research. *Royal Soc. Open Sci.* **12**, 241776 (2025).

53. of Sciences, N. A., Medicine, Affairs, G. & on Responsible Science, C. *Fostering integrity in research* (National Academies Press, 2017).

54. Rawat, S. & Meena, S. Publish or perish: Where are we heading? *J. research medical sciences: official journal Isfahan Univ. Med. Sci.* **19**, 87 (2014).

55. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. science* **22**, 1359–1366 (2011).

56. Ioannidis, J. P. Why most published research findings are false. *PLoS medicine* **2**, e124 (2005).

57. Debrouwere, S. & Rosseel, Y. The conceptual, cunning, and conclusive experiment in psychology. *Perspectives on Psychol. Sci.* **17**, 852–862 (2022).

58. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *Royal Soc. open science* **3**, 160384 (2016).

59. National Science Foundation. Important notice no. 121: New criteria for nsf proposals (1997). Announces adoption of the two merit review criteria: Intellectual Merit and Broader Impacts.

60. Zerhouni, E. Medicine. the nih roadmap. *Science* **302**, 63–72, DOI: 10.1126/science.1091867 (2003).

61. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. science* **23**, 524–532 (2012).

62. Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N. & Lelkes, Y. Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *J. Commun.* **71**, 715–738 (2021).

63. Bruner, J. The narrative construction of reality. *Critical inquiry* **18**, 1–21 (1991).

64. Watts, D. J. Common sense and sociological explanations. *Am. J. Sociol.* **120**, 313–351 (2014).

65. Duarte, J. L. *et al.* Political diversity will improve social psychological science1. *Behav. brain sciences* **38**, e130 (2015).

66. Mitchell, G. & Tetlock, P. E. The internal validity obsession. *Behav. & Brain Sci.* (2022).

67. Jellison, S. *et al.* Evaluation of spin in abstracts of papers in psychiatry and psychology journals. *BMJ evidence-based medicine* **25**, 178–181 (2020).

68. Ito, C., Hashimoto, A., Uemura, K. & Oba, K. Misleading reporting (spin) in noninferiority randomized clinical trials in oncology with statistically not significant results: a systematic review. *JAMA network open* **4**, e2135765–e2135765 (2021).

69. Navarro, C. L. A. *et al.* Systematic review finds "spin" practices and poor reporting standards in studies on machine learning-based prediction models. *J. clinical epidemiology* **158**, 99–110 (2023).

70. Chiu, K., Grundy, Q. & Bero, L. 'spin'in published biomedical literature: a methodological systematic review. *PLoS Biol.* **15**, e2002173 (2017).

71. Boutron, I. *et al.* Impact of spin in the abstracts of articles reporting results of randomized controlled trials in the field of cancer: the spiin randomized controlled trial. *J. Clin. Oncol.* **32**, 4120–4126 (2014).

72. Boutron, I. *et al.* Three randomized controlled trials evaluating the impact of "spin" in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit. *BMC medicine* **17**, 1–10 (2019).

73. Hyland, K. Writing without conviction? hedging in science research articles. *Appl. linguistics* **17**, 433–454 (1996).

74. Friedman, S. M., Dunwoody, S. & Rogers, C. L. *Communicating uncertainty: Media coverage of new and controversial science* (Routledge, 2012).

75. Gustafson, A. & Rice, R. E. The effects of uncertainty frames in three science communication topics. *Sci. Commun.* **41**, 679–706 (2019).

76. Pei, J. & Jurgens, D. Measuring sentence-level and aspect-level (un) certainty in science communications. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9959–10011 (2021).

77. Deaton, A. & Cartwright, N. Understanding and misunderstanding randomized controlled trials. *Soc. science & medicine* **210**, 2–21 (2018).

78. Stolley, P. D. When genius errs: Ra fisher and the lung cancer controversy. *Am. J. Epidemiol.* **133**, 416–425 (1991).

79. Parascandola, M. Two approaches to etiology: the debate over smoking and lung cancer in the 1950s. *Endeavour* **28**, 81–86 (2004).

80. Lee, G. & Zhai, X. Reconceptualizing epistemic dependence for future scientific literacy: A lesson from the lk-99 case. *Res. Sci. Educ.* 1–23 (2025).

81. Bem, D. J. Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *J. personality social psychology* **100**, 407 (2011).

82. Berlinguette, C. P. *et al.* Revisiting the cold case of cold fusion. *Nature* **570**, 45–51 (2019).

83. Dahlstrom, M. F. & Scheufele, D. A. (escaping) the paradox of scientific storytelling. *PLoS Biol.* **16**, e2006720 (2018).

84. Zhang, S. *et al.* An illusion of predictability in scientific results: Even experts confuse inferential uncertainty and outcome variability. *Proc. Natl. Acad. Sci.* **120**, e2302491120 (2023).

85. De Semir, V., Ribas, C. & Revuelta, G. Press releases of science journal articles and subsequent newspaper stories on the same topic. *Jama* **280**, 294–295 (1998).

86. Yavchitz, A. *et al.* Misrepresentation of randomized controlled trials in press releases and news coverage: a cohort study. *PLOS medicine* (2012).

87. Schwartz, L. M., Woloshin, S., Andrews, A. & Stukel, T. A. Influence of medical journal press releases on the quality of associated newspaper coverage: retrospective cohort study. *Bmj* **344** (2012).

88. Yu, J., Yeung, C. & Soman, D. Are media reports of published research an accurate representation of the research? *Behav. Public Policy* 1–21 (2025).

89. Feynman, R. P. Cargo cult science. In *The art and science of analog circuit design*, 55–61 (Elsevier, 1998).

90. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. spaCy: Industrial-strength Natural Language Processing in Python. *Python* DOI: 10.5281/zenodo.1212303 (2020).

91. Wright, D., Pei, J., Jurgens, D. & Augenstein, I. Modeling information change in science communication with semantically matched paraphrases. *arXiv preprint arXiv:2210.13001* (2022).

92. Yao, M., Wei, Y. & Wang, H. Promoting research by reducing uncertainty in academic writing: a large-scale diachronic case study on hedging in science research articles across 25 years. *Scientometrics* **128**, 4541–4558 (2023).

93. Bates, D. *et al.* Package 'lme4'. *convergence* **12**, 2 (2015).

94. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with python. *SciPy* **7**, 92–96 (2010).

95. Barrett, P., Hunter, J., Miller, J. T., Hsu, J.-C. & Greenfield, P. matplotlib–a portable python plotting package. In *Astronomical data analysis software and systems XIV*, vol. 347, 91 (2005).