

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	DESCRIPTIVE ANALYSIS: Article labels were generated using OpenAI GPT models accessed through the OpenAI API. BERT-based classifiers were trained and applied in Python. Article information was accessed on ProQuest's TDM Studio platform. EXPERIMENT: The human-subjects experiment was administered in Qualtrics. OpenAI GPT models were used to classify selected participant responses and generate AI-based summaries. Platform API's were also used for Claude and Deepseek Models. Statistical analyses were conducted in R and Python.
Data analysis	All analysis was conducted in Python and R. Complete reproduction materials are publicly available online: https://osf.io/zpua6/. This folder contains scripts, datasets, and package information for reproduction of the results, though the complete text for all articles in our descriptive analysis must be accessed on TDM Studio. Package versions for python and R are listed below, and the OSF repository provides instructions for how to install these (along with all their appropriate dependencies, and additional details for supplemental analysis) to enable full reproduction. Python: Python 3.9.6, pandas 2.3.3, NumPy 2.0.2, SciPy 1.13.1, statsmodels 0.14.6, scikit-learn 1.6.1, matplotlib 3.9.4, openpyxl 3.1.5 R: R 4.5.1, lme4 1.1-37, lmerTest 3.1-3, dplyr 1.1.4, broom 1.0.9, emmeans 1.11.2, ggplot2 3.5.2, clubSandwich 0.6.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All article-level (including citation information) and experimental results datasets required to reproduce the findings of this study are available on the Open Science Framework (OSF) at: <https://osf.io/zpua6/>. The underlying article texts and metadata were obtained through ProQuest TDM Studio and are subject to licensing restrictions. These data are therefore not redistributed by the authors but may be obtained directly through ProQuest's TDM Studio platform (<https://tdmstudio.proquest.com/home>) subject to institutional access and licensing agreements.

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	We do not report on sex or gender differences in this study.
Reporting on race, ethnicity, or other socially relevant groupings	We do not report on race, ethnicity, or other socially relevant differences in this study.
Population characteristics	See above
Recruitment	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. Participants received fair monetary compensation consistent with Prolific's recommended hourly rate.
Ethics oversight	The study protocol has been 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.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	This paper contains two quantitative analyses: A large scale descriptive study on the presence of causal claims in cross sectional studies and an experiment that assesses how these claims impact reader (and LLM) interpretations of a studies implications.
Research sample	DESCRIPTIVE ANALYSIS: The final dataset contained 194,631 cross-sectional articles published between 1980 and 2024. Articles were drawn from five ProQuest TDM Studio databases and included article titles, abstracts, and associated metadata. Duplicate records with different identifiers but identical titles were removed. EXPERIMENT: The experiment included 1,105 participants recruited through Prolific by the Wharton Behavioral Lab. Eligibility was restricted to adults residing in the United States who held at least a bachelor's degree. The target sample size was 1,100 participants; Prolific slightly over-recruited, resulting in a final sample of 1,105 participants. No demographic information beyond these eligibility criteria was collected. Because participation on Prolific is voluntary, the sample reflects a self-selected group of online research participants and is not intended to be representative of the broader U.S. population. LLM EXPERIMENTS: The language-model experiments examined models from three leading AI providers at the time of data collection: OpenAI (GPT-4.1, GPT-5.1, and GPT-5.4), Anthropic (Claude Sonnet 4.6), and DeepSeek (DeepSeek Chat-V3.2).
Sampling strategy	DESCRIPTIVE ANALYSIS: Articles were identified using a keyword-based search of five ProQuest TDM Studio databases and therefore constitute a purposive rather than random sample of likely cross-sectional studies. The resulting corpus was intended to provide broad coverage of cross-sectional social-science research rather than a representative sample of all academic publications. No formal

sample size calculation was performed. Instead, all eligible articles returned by the search and classification procedures were retained, resulting in a final sample of 194,631 cross-sectional articles.

EXPERIMENT: Participants were recruited through Prolific using a convenience sample restricted to U.S. adults with at least a bachelor's degree. A priori power analyses were conducted (full details in main text) for establishing our target sample. The final sample size of 1,105 participants slightly exceeded this target. Although not nationally representative, the sample was selected to provide a sufficiently powered test of the experimental manipulations.

LLM EXPERIMENTS: Models were purposively selected from three leading AI providers and were not intended to represent all available language models. No formal power analyses were conducted. The first GPT-4.1 experiment was exploratory. For the second LLM experiment, 100 articles were selected to provide a diverse set of stimuli spanning multiple causal-language categories while remaining computationally feasible. Findings should therefore be interpreted as evidence regarding the sampled models and articles rather than all possible language models or research papers.

Data collection

DESCRIPTIVE ANALYSIS: Article titles, abstracts, and metadata were collected through ProQuest TDM Studio. A random sample of 40,000 articles was exported and labeled using GPT-4o via the OpenAI API. These labels were used to train local BERT-based classifiers, which were subsequently applied to the full corpus within the secure TDM Studio environment. The resulting article-level labels, together with citation information and unique identifiers, were exported for statistical analysis and are publicly available. No human participants were involved, and researcher blinding was therefore not applicable.

EXPERIMENT: Data were collected using Qualtrics and administered online through Prolific. Participants completed the study remotely without the presence of researchers or other participants. Assignment to experimental condition was performed automatically through Qualtrics using embedded JavaScript randomization procedures. Researchers were aware of the study hypotheses but did not interact directly with participants during data collection. Because assignment was randomized and all procedures, exclusion criteria, and analyses were preregistered prior to data collection, researcher awareness of the hypotheses could not influence condition assignment or participant responses. Raw Qualtrics data were exported following data collection and all cleaning and analysis procedures are documented in the public OSF repository.

LLM EXPERIMENTS: Data were collected through the application programming interfaces (APIs) of OpenAI, Anthropic, and DeepSeek. Models were queried programmatically using standardized prompts and parameter settings described in the manuscript and Supplementary Information. Responses were recorded automatically and stored for analysis. No human participants were involved, and no researchers were present during data generation. Researcher blinding was not applicable because the outputs were generated through automated API calls using predefined procedures.

Timing

DESCRIPTIVE ANALYSIS: The ProQuest TDM Studio databases were queried on February 7, 2025. A training dataset of 30,000 articles was labeled using GPT-4o through OpenAI's Batch API on March 2, 2025. A separate validation dataset of 10,000 articles was labeled using the same procedure on April 29, 2025. The resulting classifiers were deployed within TDM Studio and applied to the full corpus during March and April 2025.

EXPERIMENT: The GPT-4.1 experiment was conducted using OpenAI's Batch API on July 3, 2025. The human-subjects experiment was conducted on Prolific on July 14, 2025.

SECOND LLM EXPERIMENT: The second LLM experiment, which evaluated multiple language models, prompting strategies, and article samples, was conducted between March 25 and March 28, 2026.

Data exclusions

DESCRIPTIVE ANALYSIS: Duplicate articles with identical titles but multiple identifiers were excluded prior to analysis (N = 6,276, 1.9%). Analyses reported in the main text were additionally restricted to articles published in peer-reviewed academic journals; non-journal publications (e.g., dissertations, conference proceedings, reports) were excluded from the primary analyses (N = 318,357). Both exclusion criteria were established prior to analysis. Supplementary analyses including all publication types yielded substantively similar results.

EXPERIMENT: No participants were excluded from the primary analyses. A preregistered attention check was included, and supplementary sensitivity analyses excluding participants who either failed this check, or completed the task in less than 1/3 the median pilot time (N = 36) are reported in the Supplementary Information.

LLM EXPERIMENTS: No model-generated responses or summaries were excluded from the analyses. All successfully generated outputs were retained and analyzed according to the preregistered procedures.

Non-participation

No participants dropped out or declined participation.

Randomization

DESCRIPTIVE ANALYSIS: No randomization was performed because this component of the study analyzed an existing corpus of published articles.

EXPERIMENT: Participants were randomly assigned by Qualtrics to one of five conditions: Original (N = 216), Rewrite (N = 240), Methodological Note (N = 229), AI Feedback (N = 202), and No Information (N = 218). Randomization was performed automatically by the survey software prior to participants viewing study materials. Additional details regarding the experimental conditions are provided in the manuscript.

LLM EXPERIMENTS: No randomization was performed. For each article and prompt condition, responses were generated separately from each language model through independent API calls. All models received the same stimuli and prompts, allowing direct comparisons across models and prompting conditions.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involvement in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involvement in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	NA
Novel plant genotypes	NA
Authentication	NA