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

In the format provided by the
authors and unedited

Supplementary Information

This PDF file includes:

Results for Non-Academic Journals	2
Instructions for Human Coding of Methodology	2
Training BERT-Based Models	2
Instructions for Human Coding of Causal Language	3
Examples of Causal and Descriptive Framing in Abstracts	4
Alternative Approaches to Assessing Causal Language	4
Comparison to Experimental Studies	5
Within-Year Impact-Factor Differences	5
Results by Causal Type	6
Full Text Analysis	6
GPT-4o Prompt to Label Journal Discipline	8
External Data from Elite Venues	8
Validating Elite Journals	9
Full Text Claim-Level Analysis: Additional Results	11
Experiment Stimuli	12
Exploratory Analysis for the Experiment	12
Sensitivity Analysis Without Flagged Participants	14
Residual Diagnostics	14
Power Analysis	16
First Experiment with GPT-4.1	16
Second Experiment with LLMs: Prompts	17
Second Experiment with LLMs: Causal Classifier	18
Second LLM Experiment: GPT-4.1 Robustness Checks	19
Second LLM Experiment: GPT-4.1 Claim-Level Analysis	20
Second Experiment with LLMs: Additional Tables	21
Causal Claims and Uncertainty	24

Results for non-academic journals

Our main descriptive analysis was conducted on the articles that ProQuest labeled as coming from peer-reviewed “Scholarly Journals” but our data also contain articles from other venues: dissertations, working papers, conference articles, and trade magazines. To what extent do we see causal language in cross-sectional articles from these venues?

Our data originally contained 318,357 non-duplicate articles from these venues of which our classifier labeled 191,152 as employing a cross-sectional methodology. In these articles, 26.0% contained some form of causal language, a figure substantially lower than what we saw for published work. When breaking the percentage down by venue type we see that trade publications show significantly lower rates (N = 64,888, 2% causal) than the others (Dissertations: N = 6,513, 56% causal; Conference papers: N = 2,337, 53% causal; Working Papers: N = 18,580, 46% causal, Other: N = 98,783, 35% causal), and this gap drives the discrepancy.

The database coverage of these articles is also not consistent. Rather than a continuous rise in the number of publications, as we saw for academic articles, these other venues have the highest rates of coverage around 2012 before dropping drastically for recent years. Such data sparsity makes over-time analysis fraught; however, when limiting to non-trade publications we see a similar rise in causal language to what we observed for academic articles. Specifically, in 1980, 19.9% of these articles contained causal language, and by 2020 the proportion rose to 51.0%.

Instructions for Human Coding of Methodology

Your task is to identify the methodology used in an academic study. Follow these steps:

1. Locate the Article: Use the provided citation to find the paper on Google Scholar. If available, access the full text.
2. Review the Methodology: Carefully examine the methodology to determine the study’s approach and analytical methods.
3. Classify the Study: Assign the paper to one of the following categories based on its methodology:
 - o Cross-sectional – Data collected at one point in time, usually using a survey or snapshot from a database.
 - o Mixed Methods – Quantitative survey + interviews or focus groups, or a combination of quantitative and qualitative data in parallel
 - o Qualitative – Reliance on interview transcripts, narrative data, thematic coding, or phenomenological analysis.
 - o Randomized Control Trial (RCT) – Participants are randomly assigned to conditions.
 - o Quasi-experimental – Involves intervention but lacks full randomization.
 - o Multi-Time Point, Longitudinal – Data collected across multiple time points.
 - o Review/Meta-analysis – Summarizes or statistically synthesizes findings from multiple studies.
 - o Multiple Studies – Includes multiple independent studies within the same paper.
 - o Other – For example, secondary coverage of an article, editorials, legal studies

Ensure that your classification accurately reflects the study’s primary research design. Many times this will be evident in the abstract alone. If uncertain, review the methods and results, sections for additional context.

4. If you use more than the abstract to code the methodology, first come up with a probable label based on the article abstract alone. In the column marked “Full Text” note whether the methodology actually matched your probable label. If it did put “Matched” otherwise put “Different”. If you are able to label the study from the abstract alone, you can leave the “Notes” and “Full Text” column blank.

- a. Make a note in the last column about cases where it didn’t match, making sure that the correct methodological label is used in the “label” column.
- b. You can make other notes about the abstract in the last section, for example, “This one was very short”, “A review of another article”, etc.

Note: We will collapse these into “Single Study, Cross-sectional” vs everything else.

Training BERT-based Models

To develop accurate classifiers for identifying methodological design and causal language in academic abstracts, we used a dataset of 30,000 titles and abstracts, labeled for both methodology and causal language by GPT-4o. These labels served as the training data for a series of local models. We evaluated five different model architectures for each task—DistilBERT, SSCI-BERT, RoBERTa, LLaMA-3b, and MiniGPT—ultimately selecting SSCI-BERT due to its superior performance across both classification tasks.

SSCI-BERT is a domain-specific pre-trained language model designed to support natural language processing applications in the social sciences. While general-purpose language models have significantly improved the accuracy of text mining for broad textual domains, SSCI-BERT addresses the need for models tailored to scientific texts within the social sciences. Built on the BERT framework, SSCI-BERT was trained on abstracts from articles published in SSCI-indexed journals between 1986 and 2021. The training corpus contains over 503 million words and supports a range of downstream tasks, including text classification. With these advantages, it was a strong initial model for our classification task.

We fine-tuned SSCI-BERT on our labeled dataset and performed threshold optimization to maximize F1 score. Performance

was then evaluated on a held-out validation set of 10,000 academic abstracts for the methodology classifier, and a subset of 6,037 abstracts labeled as cross-sectional for the causal language classifier. SSCI-BERT achieved strong F1 scores in both tasks (Methodology F1: 0.92, Precision: 0.91, Recall: 0.93, Accuracy: 0.91; Causal language F1: 0.83, Precision: 0.84, Recall: 0.82, Accuracy: 0.84), leading us to adopt it as the final model for local classification. We uploaded the final model to TDM Studio to classify all 321,322 articles from scholarly journals, finding that 194,631 were labeled as cross-sectional and of those, 46.3% contained causal language.

The trained models were uploaded to hugging face and can be accessed at the following links:

Methodology: <https://huggingface.co/calvini/cross-sectional-ss-abstracts>

Causal Language: <https://huggingface.co/calvini/causal-language-ss-abstracts>

Instructions for Human Coding of Causal Language

Your task is to determine whether a Title and Abstract contains any of the three types of causal language listed below. Please read the entire abstract carefully, but do not worry about any causal language that refers exclusively to prior studies or general background (i.e., statements about what other researchers have found). We are only interested in how the authors describe their own study's results, conclusions, and aims.

Categories of Causal Language:

1. Direct Causal Statement

a. Definition: The authors explicitly claim that their results show one variable affects or causes another in their title, results, or conclusions.

b. Key Phrases: “influences,” “causes,” “impacts,” “role of”, “mechanisms,” “determines,” or similarly direct terms indicating cause-and-effect in their own study.

c. Note: Even causal words commonly used in path or mediation analysis should be marked with this category, though “interaction effect” should not.

2. Implied Causal

a. Definition: The abstract states (usually in the introduction) that the study investigates a causal relationship. This can occur with or without causal statements in their results.

b. Key Phrases: “This study examines the influence of. . .,” “we investigate the role of X on Y. . .,” or similar wording announcing a causal focus.

c. Additional definition: The authors imply a causal relationship in their conclusions or recommendations (e.g., suggesting a policy or intervention based on their results).

d. Key Phrases: “Based on our findings, we recommend. . .,” “Our results support implementing. . .,” or any statement that implies causality by advocating actions or next steps.

3. Conditional Causal Language

a. Definition: The authors hedge or qualify a causal statement regarding their results.

b. Key Phrases: “may cause,” “could explain,” “suggests a role of,” “potentially influences,” “possibly leads to,” or any other tentative or probabilistic causal wording referring to their own findings.

4. None

a. Definition: The abstract uses purely associational or correlational language regarding the study's findings, with no indication (direct or implied) that one variable causes, affects, or influences another.

Step-by-Step Coding Procedure

1. Read the Title and Abstract in Full

2. Carefully review how the authors describe their study's objectives, methods, results, and conclusions to assess the presence of causal language.

3. Based on the strongest form of causal language present, classify the article overall, noting Direct > Implied > Conditional > None

Examples of Causal and Descriptive Framing in Abstracts

Supplementary Table 1. Examples of Causal and Descriptive Framing in Abstracts

True Label	Definition	Sentence	Classifier Label
Direct	The authors explicitly claim that their results show one variable affects or causes another in their title, results, or conclusions.	Social media causes an increase in depressive symptoms among adolescents.	Causal
		Higher income produces more happiness.	Causal
Implied	The abstract states that the study investigates a causal relationship. Alternatively, the authors imply a causal relationship in their conclusions by recommending interventions that require a causal relationship.	We explored the role of social media use in shaping depression outcomes.	Causal
		We found that students who used social media more demonstrated higher rates of depression. Parents and schools should limit children’s social media time to support mental health.	Causal
Conditional	The authors hedge or qualify a causal statement regarding their results.	Social media may contribute to depressive symptoms in teens.	Causal
		Fatty foods might raise the risk of cardiovascular disease.	Causal
Descriptive / Background	The abstract uses purely associational or correlational language regarding the study’s findings, with no indication (direct or implied) that one variable causes, affects, or influences another.	Social media usage was associated with depressive symptoms.	Descript.
		Individuals who consumed more fatty foods also had higher heart disease rates.	Descript.
		The COVID-19 pandemic has been linked to increased mental health issues.	Descript.

Alternative approaches to assessing causal language in titles and abstracts

In addition to leveraging general-purpose large language models, we employed two complementary techniques to assess the presence of causal language in titles and abstracts as robustness checks. First, we implemented a dictionary-based keyword approach. We developed a comprehensive list of terms and phrases commonly used to imply causality (e.g., “leads to,” “causes,” “role of”; see full list below). This lexicon was created through an iterative process that involved consulting with domain experts, querying GPT-4o for candidate phrases, and incorporating terms frequently observed during the manual coding of cross-sectional abstracts. While less capable of detecting nuance, this method is worth considering as results are directly interpretable, as opposed to the opaque nature of LLM labels. Second, we applied a pre-trained BERT-based classifier originally developed to detect causal claims in the concluding sections of biomedical research articles (Citation 35 main text). Because this model was optimized for sentence-level analysis, we segmented each abstract into individual sentences and applied the classifier to each one. We then calculated the number of sentences within each abstract flagged as containing causal language. Both approaches showed moderate agreement with GPT-4o classifications and produced similar, though slightly higher, overall estimates of the prevalence of causal language within cross-sectional articles. These complementary methods serve as robustness checks, supporting the validity of our primary classification strategy.

Causal Language Robustness Check 1: Keyword Analysis

General Large Language Models (LLMs) offer important advantages in classifying causal language, particularly their ability to account for context and nuance. For example, LLMs can distinguish between causal claims made in the current study and references to prior experimental findings—ignoring the latter when it is not relevant to the classification objective as is the case here. However, LLMs are inherently opaque, and their internal reasoning processes are generally uninterpretable. This lack of transparency motivates the present more interpretable method: using keywords to identify causal language. To support this complementary strategy, we developed a comprehensive list of causal keywords and phrases through an iterative process. First, we generated an initial list through discussions with fellow scholars. Second, we queried GPT-4o for additional phrases commonly associated with causality. Third, we incorporated terms frequently observed during the manual coding of 200 abstracts. The final set of terms is reflected in the regular expression string provided below.

Regex Expression with Causal Phrases:

```
b(?:leads? to|results? in|because of|causes?|effects? of|
contributes? to|induced by|mechanisms? |influences|influence
of| impacts?|affects| affect of|effect of|affecting|roles?
of|affect on|benefits|profit from|enhances|due to|helps?
|promotes?|improves|improving|beneficial| prevention of)\b
```

We applied the finalized regular expression to detect the presence of causal language across all scholarly journal articles labeled as cross-sectional within the 30,000-article dataset exported from TDM Studio. This analysis revealed that 60.6% of titles and abstracts contained at least one causal phrase from our keyword list. On average, articles labeled by GPT-4o as containing causal language included 2.3 causal keywords, compared to just 0.8 in those labeled as descriptive—a substantial and statistically significant difference (Mann-Whitney U test, $p < 0.001$). When using the presence of at least one keyword as a binary classifier for causal language, we observed moderate agreement with GPT-4o’s labels: (F1: 0.69, Precision: 0.61, Recall: 0.80, Accuracy: 0.67). The relatively high recall compared to precision suggests that the keyword-based method tends to overestimate the presence of causal language relative to GPT, likely due to its inability to account for context or the nuanced use of causal terms.

Causal Language Robustness Check 2: BioBert Model

Previous work in the biomedical sciences has produced a BERT-based model trained to detect causal language within the conclusion sections of structured abstracts. While this model is well-suited for identifying causal claims in academic writing—having been trained on a large corpus of expert-validated examples—it was not originally designed for unstructured abstracts, which are more typical in our dataset. To adapt the model for our use, we segmented each abstract into individual sentences, applied the classifier to each sentence, and calculated the proportion of sentences within each abstract identified as containing causal language.

This approach revealed clear differences by GPT-4o classification: articles labeled as containing causal language by GPT had an average of 26.3% of their sentences flagged as causal by the BERT model, compared to just 9.1% for those labeled as descriptive, a significant difference (Mann Whitney $p < 0.001$). When treated as a binary outcome—whether at least one sentence in the abstract was classified as causal—the model identified 55% of articles as containing causal language. This method again demonstrated moderate alignment with GPT-4o’s classifications: (F1: 0.72, Precision: 0.63, Recall: 0.85, Accuracy: 0.69). Once more, recall was much higher than precision, suggesting that this method overcounts the proportion of articles with causal language. As with the keyword-based approach, the higher recall relative to precision suggests that this model overestimates the prevalence of causal language relative to GPT, likely due to limited sensitivity to context or conditional phrasing.

Comparison to Experimental Studies

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

Within-Year Impact-Factor Differences

In the main text we report differences in causal overstatement across time and by journal impact factor (IF). Because these patterns could be confounded if higher-IF journals contributed a growing share of articles over time, we fit a multilevel logistic model with IF quintile as the predictor, year fixed effects, and cluster-robust SEs at the journal level. Relative to the lowest IF quintile, the odds of causal overstatement were modestly higher in the second quintile (Wald $z = 2.99$, $p = 0.003$, OR = 1.27, 95% CI = [1.08, 1.48]), statistically indistinguishable in the third (Wald $z = -0.21$, $p = 0.831$, OR = 0.97, 95% CI = [0.74, 1.26]), and substantially higher in the fourth (Wald $z = 4.20$, $p < 0.001$, OR = 1.59, 95% CI = [1.28, 1.98]) and fifth (Wald $z = 9.72$, $p < 0.001$, OR = 2.37, 95% CI = [1.99, 2.82]) quintiles. Thus, higher-impact journals exhibit more causal overstatement even after conditioning on publication year, indicating that Figure 3 is not explained by time trends alone.

Results by Causal Type

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

Extended Data Figure 3 shows how the proportion of causal claim types changed over time. In 1980, 83% of cross-sectional articles were classified as descriptive; among those with causal language, 6.8% were direct, 7.1% implied, and 3.1% conditional. Beginning in the 1990s, causal phrasing increased as descriptive language declined. By 2020, descriptive classifications fell to 44%; direct (26.4%) and implied (23.7%) claims became much more common, while conditional claims remained comparatively rare (5.5%).

Extended Data Figure 4 shows the breakdown of causal claim type by journal impact factor, split into quintiles for this sample. It reveals that higher impact factor journals have substantially more causal claims overall, as reported in the main text. In the lowest impact-factor quintile (0–1.69), descriptive language remains a majority (58.9%), with direct (17.9%), implied (19.0%), and conditional (4.2%) claims less common. In the highest quintile (5.34–19.2), descriptive falls to 41.6%, while direct (28.7%) and implied (24.2%) are over-represented; conditional claims remain a small share (5.6%).

Full Text Analysis

ProQuest databases store full-text content within a dedicated `<Text>` field in the XML data. To examine causal language in full-text articles, we followed a multi-step process. First, we assessed the presence of full text across the dataset of 672,944 articles. Of these, 176,785 included a populated full-text field. Among articles from scholarly journals ($N = 327,598$), the availability of full text was higher, with 109,718 containing this field. We limited our analysis to the 69,580 articles classified as cross-sectional from these (70,472 before dropping duplicates).

For each article with full text, we preprocessed the content by removing sections following the references. We used regular expressions to detect common headings that indicate the start of a reference section, such as “References”, “Bibliography”, “Citations”, or “Works Cited”.

```
r'(?i)^\s*(?:\d+\.?\s*)?(references|bibliography|  
citations|works cited)\s*$'
```

To identify instances of potential causal language, we applied the same set of keywords used in our earlier keyword analysis. For every keyword match, we extracted a 200-character snippet—100 characters before and after the keyword—to provide contextual information.

```
big_regex = re.compile(r"\b(?:leads? to|results? in|because of|causes?|effects? of|contributes? to|induced by|mechanisms?)|"  
r"influence of|impacts?|affects|affect of|affecting|roles? of|affect on|benefits?|profit from|enhances|helps?|"  
r"promotes?|improves?|improving|beneficial|prevention of|interventions|implications?|programs?influences?)\b",  
re.IGNORECASE)  
  
def detect_causal_language(t, context_length = 100):  
    matches_list = []  
    all_matches = list(big_regex.finditer(t))  
    for m in all_matches:  
        matches_list.append(t[max(0,m.start()-context_length):min(m.end()+context_length, len(t))])  
    return matches_list
```

Supplementary Figure 1. Regex code utilized to gather text surrounding causal keywords within the full text of articles.

Next, we applied a custom regular expression to detect and exclude segments that included in-text citations (e.g., parenthetical references or bracketed numbers). While this filtering step may exclude valid causal claims that also reference prior work, we prioritized precision in our analysis, focusing on unambiguous instances of causal language that appear to describe the authors' own findings in cross-sectional studies.

Finally, we ran the remaining text segments through the BERT-based causal language classifier used in our primary analysis. Thus, for a text chunk to be labeled as containing causal language, it had to meet three criteria: (1) contain a causal keyword, (2) be free of in-text citations, and (3) be classified as causal by the BERT model.

We display the proportion of causal claims within the full texts of articles by abstract type in Supplementary Information Figure 3. Both article types were very likely to contain causal snippets of text.

```

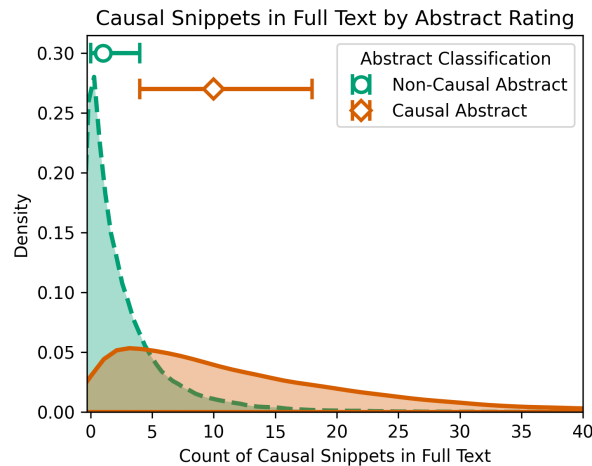
numeric_pattern = r'[\s*\d+(?:\s*,\s*\d+)*\s*\d+]'
author_year_pattern = (
    r'(\s*'                                     # opening parenthesis with optional whitespace
    r'(?:(?:e.g.\s+)?'                           # optional "e.g." prefix
    r'[A-Z][A-Za-z]+'                             # first author's last name
    r'(?:(?:\s+(?:et\s+al\.|and|&)\s+[A-Z][A-Za-z]+)?' # optionally "et al." or "and" + second author
    r'\s+\d{4}'                                     # year (four digits)
    r'(?:(?:\s*;\s*'                               # additional citations separated by semicolons
    r'[A-Z][A-Za-z]+'
    r'(?:(?:\s+(?:et\s+al\.|and|&)\s+[A-Z][A-Za-z]+)?'
    r'\s+\d{4})*)'
    r'\s*\s*'
)
preceding_pattern = (
    r'[A-Z][A-Za-z]+(?:\s+(?:and|&)\s+[A-Z][A-Za-z]+)+' # one or more authors separated by "and" or "&"
    r'\s+(\s*\d{4}\s*)'                                   # year in parentheses with optional whitespace
)
full_reference_pattern = r'.*\b\d{4}\b.*'

numeric_re = re.compile(numeric_pattern)
author_year_re = re.compile(author_year_pattern, flags=re.DOTALL) # DOTALL so newlines are handled
preceding_re = re.compile(preceding_pattern)
full_reference_re = re.compile(full_reference_pattern)

# Contains citation?
def contains_citation(text):
    return bool(numeric_re.search(text)) or bool(author_year_re.search(text)) or bool(preceding_re.search(text)) or
    bool(full_reference_re.search(text))

```

Supplementary Figure 2. Regex code utilized to remove causal snippets that were likely to contain citations



Supplementary Figure 3. Density plots with articles based on the number of causal snippets contained in the full text, colored by the abstract categorization of causal or not. Points and error bars show medians and IQRs for each group. Across all articles, 81.8% of full text contained a causal snippet, such snippets were more common in articles with abstracts that also contained causal claims.

GPT-4o prompt to label journal discipline

I will provide you with a list of names for **twenty academic journal**.
Your task is to respond **only** with an array of JSON objects following these
→ guidelines:

1. **journal**: Echo back the exact name of each journal provided.
2. **discipline**: Identify each journal's primary discipline. Use a broad category
→ such as:
 - Psychology
 - Communication
 - Economics
 - Anthropology
 - Political Science
 - Sociology
 - Education
 - Public Health
 - Business
 - Biology
 - Chemistry
 - Physics
 - Computer Science
 - Medicine
 - Engineering
 - Interdisciplinary
 - Other (if unclear)
3. **quality**: Rate each journal's overall impact or prestige, choosing exactly
→ one of the following labels:
 - "High Impact" - "Moderate Impact" - "Low Impact" - "Uncertain"
4. **reputation**: Indicate how widely recognized each journal is, choosing exactly
→ one of:
 - "Highly Reputable"
 - "Well-known"
 - "Less Known"
 - "Obscure"

If you cannot determine the discipline, quality, or reputation, use "Other" for
→ discipline and "Uncertain" for quality or "Obscure" for reputation.

\\#\\#\\# Output Format:

```
[{\n  "journal": "The journal name you received as input",\n  "discipline": "One of the listed \\disciplines or 'Other'",\n  "quality": "One of ['High Impact', 'Moderate Impact', 'Low Impact',\n    → 'Uncertain']",\n  "reputation": "One of ['Highly Reputable', 'Well-known', 'Less Known',\n    → 'Obscure']"\\}, ...]
```

\\#\\#\\# Input:

Journal Names:

External Data from Elite Venues

We collected article data from several high-impact journals, *PNAS*, *Nature*, *Science*, *Nature Human Behaviour*, and *Science Advances*, that were not included in the ProQuest databases studied. These journals publicly provide data including article titles, abstracts, author information, and publication date on their online platforms. We implemented a modular web scraping

pipeline in Python using Selenium, a popular suite of web scraping tools, to systematically gather this data from each journal. Each journal was handled by a separate module, tailored to the source’s specific site layout. The resulting dataset contained the fields: *link*, *title*, *abstract*, *authors*, *publish date*, and *topic*. The topic field was left blank when not provided by the source and additional filtering was applied using a GPT-based classifier as specified below.

Because each journal follows a different system for organizing and categorizing content online, we applied tailored collection criteria and methods for each source. We excluded content pre-1980 for relevance.

- **PNAS** specifies the social sciences to include the following topics: “Demography”, “Economic Sciences”, “Environmental Sciences”, “Political Sciences”, “Psychological and Cognitive Sciences”, “Social Sciences”. We filtered by these topics in the journal’s online search function and included all articles containing abstracts.
- **Nature** specifies the social sciences to include the following topics: “Environmental Social Sciences”, “Health Care”, “Health Occupations”, “Scientific Community”, “Social Sciences”, “Business and Industry”. We filtered by these topics in the journal’s online search function and applied a filter by article type to include only Research Articles.
- **Science** does not provide topical categorizations for the social sciences. We collected data from all articles containing abstracts from journal issues published between 1980 and the present from the journal’s online “Archive” of issues.
- **Science Advances** provides numerous, specific topical categorizations that are contained within a broader Social Science category. Consequently, we did not limit our initial collection of articles by topic.
- **Nature Human Behaviour** is a journal that specializes in social science topics; therefore, we included all articles published in this venue.

As we collected articles from these journals, we restricted our dataset to those that mentioned the same keywords used in identifying articles on TDM Studio (i.e., crosssectional, cross-sectional, survey, observational study, prevalence study, questionnaire). For articles from *Science*, we further filtered the sample by running each title and abstract through a GPT-based classifier to determine whether the article primarily pertained to behavioral or social science. Below is this prompt:

You will be shown the title and abstract of a study. Determine whether the study is
→ primarily relevant to the behavioral and social sciences. Respond with a single
→ word, either Yes or No"

After reviewing the articles from *Nature*, we found that a substantial portion did not pertain to behavioral science, so we applied the same GPT-based classification prompt to these articles. Finally, *Science Advances* contains a large quantity of topics (N = 395), we manually reviewed these and labeled a subset of topics as relevant to the behavioral and social sciences (N = 68). We kept all articles in these topics within our final sample. We provide the number of articles contained at each stage of filtering in Supplementary Information Table 2. We applied both our methodology classifier and causal language classifier to all of the articles that were behavioral and containing relevant keywords (the last column of the table).

Supplementary Table 2. Counts of articles collected from external sources at each stage of filtering

Journal	Total Collected	Behavioral Articles	Matching Keywords
<i>Science</i>	39,569	1,327	34
<i>Nature</i>	488	167	19
<i>PNAS</i>	4,491	4,491	294
<i>Nature Human Behaviour</i>	756	756	51
<i>Science Advances</i>	14,889	1,039	56

We collected data from the (*Journal of Quantitative Description* using the same process as for the other journals. However, because this journal is focused on the behavioral sciences and contains a small number of articles at the time of collection (N = 101), we did not apply either the behavioral or keyword filter to the articles from this venue, relying only on GPT’s labeling of whether the article was cross-sectional or not.

Validating Elite Journals

Our dataset included several journals widely recognized as flagship outlets within their respective disciplines. Articles published in these venues often undergo more rigorous review processes and feature more sophisticated analyses, potentially making them categorically different from the broader sample, which includes many articles from less prestigious sources. As a result, classifier performance may vary on this subset.

Supplementary Table 3. Cross-Sectional Articles from Elite Journals by Journal and Discipline

Publication Title	Count	Cross-Sectional	Prop Causal	Discipline
Econometrica	56	5	0.20	Economics
The American Economic Review	497	172	0.30	Economics
The Journal of Political Economy	135	50	0.36	Economics
The Quarterly Journal of Economics	75	23	0.39	Economics
The Review of Economic Studies	36	10	0.50	Economics
Nature Human Behaviour	51	10	0.50	General
Nature	19	1	0.00	General
Proceedings of the National Academy of Science	294	80	0.40	General
Science	34	10	0.30	General
Science Advances	56	18	0.61	General
Journal of Quantitative Description: Digital Media	101	33	0.12	JQD: Other
American Journal of Political Science	287	85	0.55	Political Science
British Journal of Political Science	224	71	0.52	Political Science
Journal of European Public Policy	101	52	0.63	Political Science
The American Political Science Review	231	69	0.48	Political Science
The Journal of Politics	405	128	0.53	Political Science
American Sociological Review	322	110	0.44	Sociology
Annual Review of Sociology	27	1	0.00	Sociology
Demography	380	104	0.38	Sociology
Socius: Sociological Research for a Dynamic World	272	129	0.36	Sociology
The American Journal of Sociology	110	40	0.42	Sociology

To assess whether our automated classification system effectively identified causal language within cross-sectional articles from these top-tier journals, we selected five prestigious publications from four broader disciplines economics, sociology, political science, and general social science for extensive data quality checks.

Our selection of prestigious publications was based on features other than journal impact factor. For economics, we used the “Top Five” as widely recognized in the field: *The American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*. For Political Science, we gathered the top five journals as listed on Google Scholar as of March 20, 2025: *American Political Science Review*, *British Journal of Political Science*, *American Journal of Political Science*, *Journal of European Public Policy*, and *The Journal of Politics*. For Sociology we used four of the five top journals listed on Google Scholar; *American Sociological Review*, *Annual Review of Sociology*, *Demography*, and *Socius*; as well the *American Journal of Sociology*. All of these journals were available within the data on TDM Studio. We supplemented these with other top general science journals that were not contained in these data: the *Proceedings of the National Academy of Science* (PNAS), *Nature*, *Nature Human Behaviour*, *Science*, and *Science Advances*. Note that *PNAS* also included articles published in *PNAS Nexus*. See main text “Supplemental Data” for details about how articles were incorporated from these outside sources. See counts of the number of articles classified as cross-sectional and the proportion of those that were classified as causal in Supplementary Information Table 3.

We drew a stratified (across journals) random sample ($N = 4$, or less if fewer articles were available) of articles initially labeled as cross-sectional by GPT and two authors manually reviewed the full text to verify that they met standard criteria for cross-sectional designs. We also assessed each article’s title and abstract for the presence of causal language. This review confirmed that 50 of the 74 articles (68%) were correctly identified as cross-sectional. Part of the reason why this performance was lower than in our full dataset may be because many of these articles contained several analyses within their main text. Often this included a small time-series analysis or instrumental variable component. All articles that had any analyses like these were labeled as not cross-sectional in this validation check, to ensure our data only included clearly cross-sectional articles.

Supplementary Table 4. Convenience sample of cross-sectional articles with causal language from Elite Journals

Title	Publication Title	Brief Description Methods	Why Causal?
Private protection in Russia and Poland	American Journal of Political Science	Regression analysis on responses to a survey of 240 small private business in three Russian cities	"...disorganized regulation increases informal activity and promotes private protection."
Atheists As "Other": Moral Boundaries and Cultural Membership in American Society	American Sociological Review	Regression analysis on survey responses from a nationally representative random-digit dial (RDD) telephone survey (N = 2081)	"This distrust of atheists is driven by religious predictors..."
Economic effects of quality regulations in the day-care industry	The American Economic Review	SEM on survey data from the 1990 National Child Care Survey (NCCS)	"[We] estimate the impact of state-imposed quality regulations on equilibrium market prices..."
Expectations of brilliance underlie gender distributions across academic disciplines	Science	Regression and mediation analysis on survey responses of 1800 professors, post docs, and graduate students across disciplines and across the USA.	"Expectations of brilliance underlie gender distributions...because women are stereotyped as not possessing such talent."
Revealing the zone of possible agreement between parties in conflict: An application to Israeli-Palestinian peace agreements	PNAS Nexus	Group comparisons of opinions for those who self-reported being the victims of violence in the Israel-Palestine conflict in representative sample from both locations	"Violence exposure hampers compromise among Israelis..."

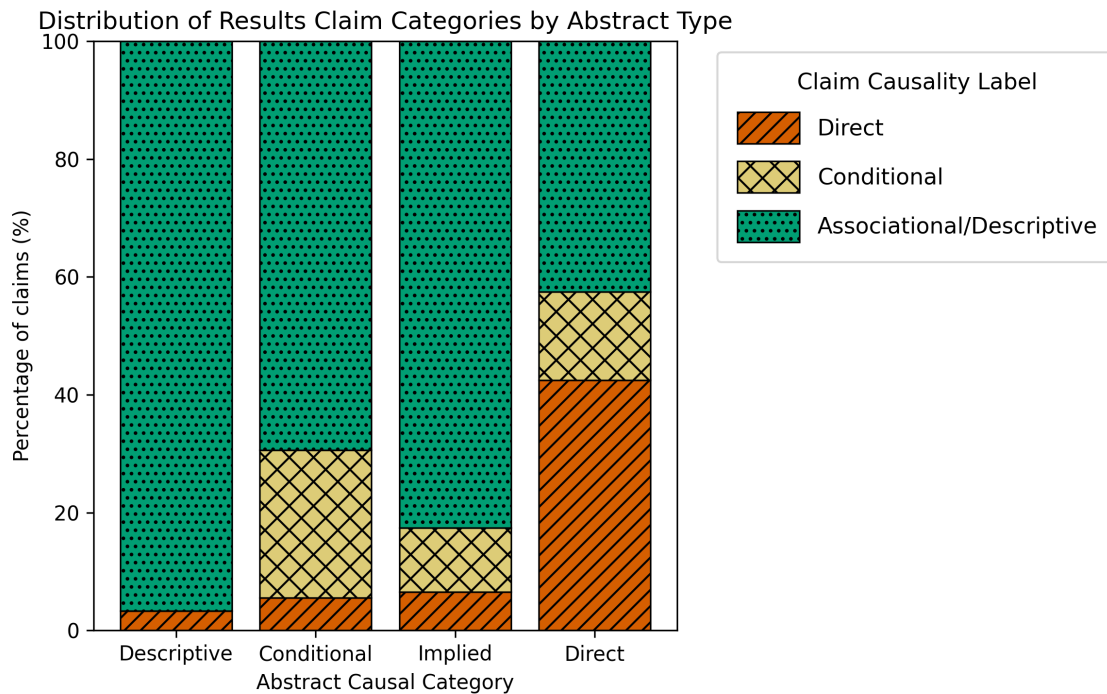
Among the 50 cross-sectional articles, manual review verified that 28 contained causal language, moderately more than identified by GPT (N = 23). In the subset, GPT's causal language detection showed higher precision than recall (F1: 0.82; Precision: 0.91; Recall: 0.75; Accuracy: 0.82), suggesting that the proportions reported in Supplementary Information Table 1 and Main Text Figure 3 likely underestimate the true prevalence of causal language in these venues. Thus, our main text estimates, showing that these journals contain levels of causal language comparable to the broader sample, remain well supported (GPT estimate = 43% causal; Human estimate = 56% Causal). We provide a sample of the articles confirmed to be cross-sectional with brief explanations about their study designs and causal language in Supplementary Information Table 4. For similar descriptions and causal snippets of all cross-sectional articles in this sample, refer to supplemental file "Good-Journals-Rand-Strat-Check-Final.xlsx".

Full Text Claim Level Analysis - Additional Results

Here, we report additional results from our full-text, claim-level analysis. Please refer to that Methods subsection for details on how we identified and matched claims within the texts. First, Supplementary Information Figure 4 shows the distribution of causal claim types among sentences classified as reporting results in the abstracts. As expected, most claims were associational/descriptive, and all causal abstract-level categories (Direct, Implied, Conditional) contained substantially more causal claims than abstracts labeled as purely associational.

We next turn to matched claims within the full text. The majority of full-text claims were associational/descriptive (N = 2,287; 73.0%), followed by unhedged causal (N = 487; 15.%) and conditional causal (N = 360; 11.5%) claims. When matched to abstract-level claim types, we observed that associational abstract claims (N = 138) had several associational matches in the full text (M = 11.2), but also included conditional causal (M = 1.3) and unhedged causal (M = 1.5) counterparts. Conditional abstract claims (N = 20) showed a similar pattern but with more matches across all types (Associational/Descriptive = 17.3; Conditional = 5.1; Unhedged Causal = 4.7). Direct abstract claims (N = 24) produced even more unhedged causal matches (Associational/Descriptive = 18.6; Conditional = 4.2; Unhedged Causal = 8.0).

In all cases, associational claims were most common, yet papers frequently used causal language to describe results. When examining whether each abstract claim had at least one unhedged causal match in the full text, a binary indicator of causal overclaiming, we found that 38% of descriptive claims (N = 50), 75% of conditional claims (N = 15), and 92% of unhedged claims (N = 22) did so. Because abstracts often contained multiple result claims, aggregating to the paper level produced even



Supplementary Figure 4. Distribution of Results-Claim Categories, within the abstract, by Abstract Type. Total counts by abstract type: Descriptive = 60, Conditional = 36, Implied = 46, Direct = 40. All abstracts contained a large proportion of associational/descriptive claims about their results, though those with a causal abstract-level label included substantially more causal and conditional causal claims.

higher proportions (as discussed in the main text). There we also analyzed the distribution of causal claims across sections of the full text, finding they were overrepresented in the Introduction and Discussion and comparatively rare in the Methods section. Note: Given the small sample sizes for each claim type, these proportions have wide margins of error.

Experiment Stimuli

Our experiment included 28 unique articles (from the stratified random sample of elite journals), each presented across 4 experimental conditions, resulting in a total of 112 distinct stimuli. To prevent participants from copying and pasting the text, we converted each stimulus into a PNG image. These images were uploaded to an OSF folder, “Experiment/Stimuli”, with filenames corresponding to the GOID from our database and condition-specific labels. For example, “123_O_A” refers to the original version (O) of article 123 in the AI feedback condition (A); and “123_R_O” refers to the rewritten abstract which did not contain a Note (second O). You can see abstracted examples of the four experimental conditions in Extended Data Figure 5

Each article was paired with a set of dedicated questions. The exact wording of all questions is provided in an Excel file in the OSF “Experiment” folder (“Experiment-Stimuli-2025-07-03.xlsx”). We used JavaScript to automate the randomization and presentation of questions, ensuring that all participants received the appropriate set. A QSF file containing the full Qualtrics project is also available in the same OSF folder: . <https://osf.io/zpua6/>.

Exploratory analysis for the experiment

We were interested in how both stimulus features and participant characteristics moderated our experiment findings. Specifically, we examined whether certain attributes corresponded to differences in participants’ interpretations of causality. On the stimulus side, we considered: (1) the plausibility of the causal relationship implied in the abstract; (2) the strength of causal language used in the original version of the abstract, as assessed by statistical experts prior to the experiment; and (3) the complexity of the abstract’s language, measured both by objective metrics (e.g., Flesch-Kincaid reading level) and subjective ratings from expert judges. On the participant side, we explored whether individuals’ statistical and general education, captured through brief self-report items, moderated their susceptibility to causal inferences.

To evaluate the strength of causal language, three of the authors independently provided two judgments for each of the 28 abstracts. First, they categorized the type of causal language used, based on the categories described in the main text: Direct, Implied, and Conditional. The majority label was used for each abstract (6 were labeled Implied, and the remaining 22 were

Direct). Second, the coders rated the strength of the causal language on a three-point scale: Strong, Moderate, or Weak. Again, we used the majority label for each abstract (14 were rated Strong, 12 Moderate, and 2 Weak).

To assess the plausibility of the causal claims, we first extracted the main causal claim from each article; these same claims informed the independent and dependent variables used in our question design. We then queried GPT-4.0 to rate the plausibility of each claim on a scale from 0 (Extremely Implausible) to 100 (Extremely Plausible). We collected 10 responses and used the mean rating for each claim as its plausibility score.

To measure the complexity of the abstract language, we used two approaches. First, we calculated the automated Flesch-Kincaid grade level for each abstract. However, because these scores did not appear to reflect the perceived complexity of the texts well, the same three coders also provided independent assessments of readability on a four-point scale: Clear and straightforward, Somewhat challenging, Moderately challenging, and Too complex. We used the majority rating for each abstract (8 were rated Clear and straightforward, 13 Somewhat challenging, 5 Moderately challenging, and 2 Too complex).

Participant education levels were determined based on self-reported responses collected at the end of the Qualtrics survey. General education was assessed with the question, “What is the highest level of education you have completed?” which included seven options to capture educational attainment. Statistical training was assessed with the question “Which best describes your background in statistics or data analysis?” Participants selected from five predefined response options. Full question wording and response categories are available in the Qualtrics files on our OSF page.

Finally, we re-estimated the models reported in the main text, this time including interaction terms for each moderator (treated as numeric). For the Causality and Implied outcomes, coefficients are unstandardized and can be interpreted as the expected change (on a 1–7 scale) in the outcome for a one-unit increase in the moderator in that condition. For the Summary outcome, coefficients have been exponentiated and should be interpreted as odds ratios. For these exploratory analysis, we use unadjusted $p = 0.05$ as significance thresholds. Notably, we ran 63 tests and found only two effects that crossed this significance threshold, slightly lower than the number expected by chance. We take this as an indication that our moderators did not strongly interact with the experiment conditions, but it is important to note that our power analysis was conducted for main effects, so the study is underpowered to detect these interactions. We report the results of this moderation analysis in Supplementary Information Table 5.

Supplementary Table 5. Moderators for the Human Subject Experiment

	Moderator	Outcome	Condition	Estimate	p
1	Causal Strength	Causal	Rewrite	0.01	0.974
2	Causal Strength	Causal	Note	0.11	0.613
3	Causal Strength	Causal	AI	-0.17	0.457
4	Causal Strength	Implied	Rewrite	-0.14	0.443
5	Causal Strength	Implied	Note	0.08	0.648
6	Causal Strength	Implied	AI	0.02	0.913
7	Causal Strength	Summary	Rewrite	0.62	0.157
8	Causal Strength	Summary	Note	0.70	0.323
9	Causal Strength	Summary	AI	0.40	0.011
10	Causal Type	Causal	Rewrite	-0.02	0.941
11	Causal Type	Causal	Note	-0.03	0.941
12	Causal Type	Causal	AI	0.12	0.739
13	Causal Type	Implied	Rewrite	0.03	0.908
14	Causal Type	Implied	Note	0.32	0.280
15	Causal Type	Implied	AI	0.52	0.090
16	Causal Type	Summary	Rewrite	0.45	0.117
17	Causal Type	Summary	Note	1.92	0.233
18	Causal Type	Summary	AI	0.76	0.620
19	Plausible	Causal	Rewrite	-0.01	0.493
20	Plausible	Causal	Note	-0.02	0.286
21	Plausible	Causal	AI	-0.02	0.314
22	Plausible	Implied	Rewrite	-0.02	0.331
23	Plausible	Implied	Note	0.00	0.906
24	Plausible	Implied	AI	-0.02	0.238
25	Plausible	Summary	Rewrite	1.05	0.124
26	Plausible	Summary	Note	1.03	0.278

Supplementary Table 5. Moderators for the Human Subject Experiment

	Moderator	Outcome	Condition	Estimate	p
27	Plausible	Summary	AI	1.04	0.263
28	Readable	Causal	Rewrite	0.02	0.904
29	Readable	Causal	Note	0.08	0.620
30	Readable	Causal	AI	0.22	0.156
31	Readable	Implied	Rewrite	-0.21	0.116
32	Readable	Implied	Note	-0.11	0.375
33	Readable	Implied	AI	-0.12	0.353
34	Readable	Summary	Rewrite	1.03	0.913
35	Readable	Summary	Note	0.87	0.579
36	Readable	Summary	AI	0.77	0.298
37	FK Grade	Causal	Rewrite	-0.03	0.105
38	FK Grade	Causal	Note	0.00	0.898
39	FK Grade	Causal	AI	-0.03	0.078
40	FK Grade	Implied	Rewrite	0.00	0.819
41	FK Grade	Implied	Note	0.01	0.332
42	FK Grade	Implied	AI	-0.02	0.183
43	FK Grade	Summary	Rewrite	1.08	0.006
44	FK Grade	Summary	Note	1.02	0.429
45	FK Grade	Summary	AI	1.03	0.286
46	Stat Ed.	Causal	Rewrite	0.01	0.897
47	Stat Ed.	Causal	Note	-0.05	0.684
48	Stat Ed.	Causal	AI	-0.09	0.455
49	Stat Ed.	Implied	Rewrite	-0.08	0.378
50	Stat Ed.	Implied	Note	-0.12	0.220
51	Stat Ed.	Implied	AI	-0.16	0.098
52	Stat Ed.	Summary	Rewrite	1.05	0.794
53	Stat Ed.	Summary	Note	1.22	0.280
54	Stat Ed.	Summary	AI	0.9	0.573
55	Education	Causal	Rewrite	-0.04	0.756
56	Education	Causal	Note	-0.18	0.106
57	Education	Causal	AI	-0.13	0.267
58	Education	Implied	Rewrite	-0.05	0.603
59	Education	Implied	Note	-0.11	0.209
60	Education	Implied	AI	-0.15	0.129
61	Education	Summary	Rewrite	0.96	0.831
62	Education	Summary	Note	1.29	0.165
63	Education	Summary	AI	1.0	0.994

Sensitivity analysis without flagged participants

As preregistered, we repeated all analyses after excluding participants who (i) completed the experiment in less than one-third of the median completion time or (ii) failed the attention check (total excluded $N = 36$). All model specifications were identical to those in the main text.

In the primary analysis, participants were less likely to judge that a study provided causal evidence when (a) the write-up used strictly associational language ($\beta_{\text{Rewrite}} = -0.25, p = 0.003$), (b) a methodological note was included ($\beta_{\text{Rewrite}} = -0.40, p < 0.001$), and (c) AI feedback was present ($\beta_{\text{AI}} = -0.60, p < 0.001$). Each effect was equal to or larger in magnitude than in the full-sample analysis. For secondary outcomes assessing subtle and implied causal inference, effect sizes remained small and non-significant after adjustment for multiple comparisons (β s range from -0.11 to -0.18). For the free-response outcomes, the odds ratios were similar to or larger than those in the main text. ($OR_{\text{Rewrite}} = 0.37, p < 0.001$; $OR_{\text{Note}} = 0.73, p = 0.082$; $OR_{\text{AI}} = 0.58, p = 0.019$). Overall, the pattern of results and our substantive conclusions remained similar when the flagged participants were excluded.

Residual Diagnostics

We assessed model assumptions for the primary human-experiment model using `performance::check_model` in R. The diagnostics indicated approximate homoscedasticity, no systematic nonlinearity in residuals vs. fitted values, no evidence of influential cases, and normality of the random effects. The residual Q-Q plot showed mild departures from normality (Supplementary Information Fig. 5). To gauge any impact on inference, we re-estimated the model with cluster-robust (CR2) standard errors; fixed-effect estimates remained statistically significant: ($b_{\text{Rewrite}} = -0.36, p = 0.001$; $b_{\text{Note}} = -0.54, p < 0.001$; and $b_{\text{AI}} = 0.81, p < 0.001$).

Power Analysis

We ran a pilot experiment with 5 randomly selected articles and 200 participants. This pilot revealed moderate effect sizes for our primary outcome (Cohen’s $D = 0.5$) and smaller effects for our secondary items ($D = 0.2\text{--}0.3$). With the pilot data, we performed two forms of a priori power analysis.

The first utilized the minimum effect of interest ($D = 0.25$ or $F = 0.18$) and looked for a difference for one of the groups in anova using the “pwr” package in R. With a Bonferroni-adjusted significance level of $\alpha = .0333$ (for three planned contrasts and one sided p-values) and 80% power, `pwr.anova.test` returned $n = 98$ participants per group. Multiplying by the four conditions and correcting for clustering within the five article stimuli used in the pilot (with an ICC design effect of 1.17) yielded a final target of 457 participants.

For the second, simulation-based a priori power analysis, we employed the “simr” package in R to run 1,000 simulations of the mixed-effects model fitted to the pilot data. The design was expanded from the pilot’s five articles to the full set of 28 stimuli by extending the random-intercept term and allocating an equal number of participants to each stimulus. Fixed-effect coefficients for the three experimental conditions (Rewrite, Note, AI) were set to the smallest effect of interest (Cohen’s $d = 0.25$) in the predicted negative direction. Power was defined as obtaining significance for any one of the three contrasts against the Original abstract, using one-sided p-values with a Holm–Bonferroni correction ($\alpha = .05$) to control the family-wise error rate. The resulting power curve (Supplementary Information Figure 6) indicated that a total sample of $N = 784$ (196 per experimental condition) would provide approximately 81% power to detect the target effect. Given this power analysis, we used one-tailed tests with Holm–Bonferroni correction to increase power for detecting effects in the pre-specified direction; this decision, along with our directional hypotheses, was pre-registered, and all conclusions remain substantiated with a two-tailed test.

Given these two methods, we went with the higher estimated N from our second power analysis in establishing the appropriate recruitment size. To accommodate an additional no-information control arm required for ancillary analyses, this figure was multiplied by 1.25, yielding $N = 980$. Finally, an extra 120 participant buffer was added to offset potential attrition, producing a final recruitment target of $N = 1,100$ participants.

First Experiment with GPT-4.1

We conducted a parallel experiment using GPT-4.1 in place of human participants, replicating the design of our original human subjects study. For each trial, GPT-4.1 was provided with the question prompt followed by the title and abstract of all articles across all experimental conditions (presented one per prompt). In the methodological note and AI feedback conditions, the corresponding comments were presented before the title and abstract but after the question text. Text highlights were indicated using LaTeX syntax from the soul package (e.g., `\hl{Highlighted text}`).

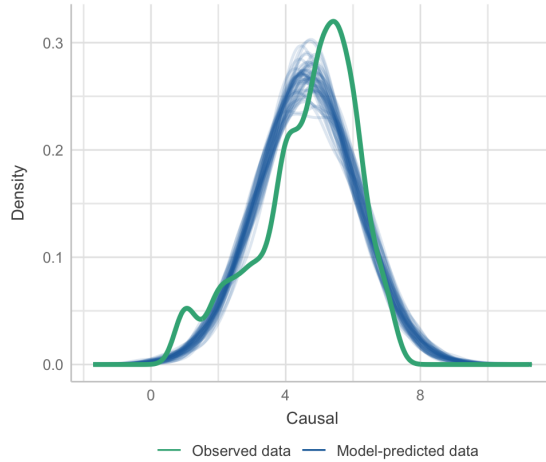
For each condition, we queried GPT-4.1 for 10 responses per article-question pair. For questions with Likert responses, responses were capped at a maximum of 5 tokens; for free-response items, the maximum was set to 100 tokens. We used a temperature setting of 1.0. All responses were generated on July 3, 2025.

We processed GPT responses by extracting only numeric values for the Likert questions, computing a single average score per article. For the free-response items, we applied a slightly updated causal language classification prompt to generate a binary outcome indicating whether each summary contained causal language (this prompt was also used to evaluate human summaries). We then calculated the proportion of responses (out of ten) classified as containing causal language for each article. As a result, we obtained one continuous measurement per article per condition (total $N = 168$). We conducted all analyses using Python and R.

We employed parallel models to those used in the human participant analysis. Specifically, we fit mixed-effects models with random intercepts for article ID and fixed effects for the three experimental predictors: writing condition, methodological label, and AI-generated feedback. For the model predicting summary-based outcomes, we used a logistic regression framework with the response specified as the count of causal versus non-causal claims across GPT responses.

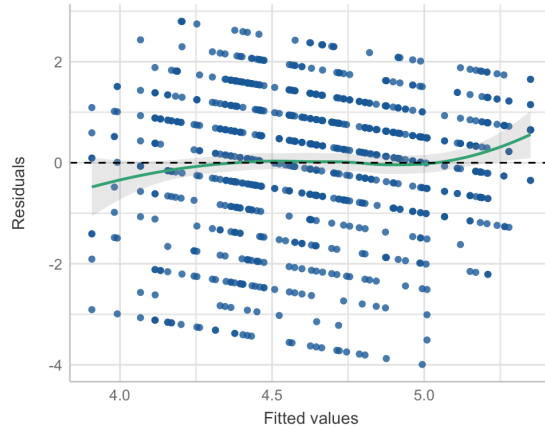
Posterior Predictive Check

Model-predicted lines should resemble observed data line



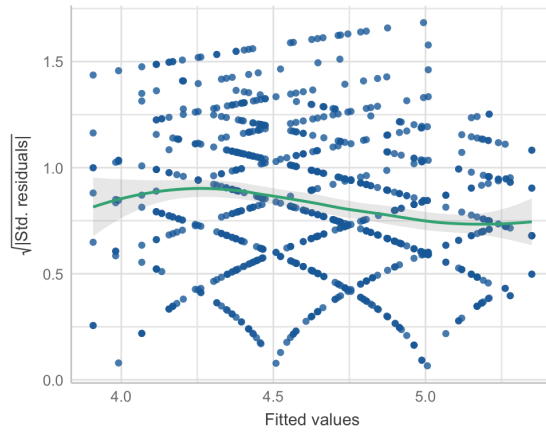
Linearity

Reference line should be flat and horizontal



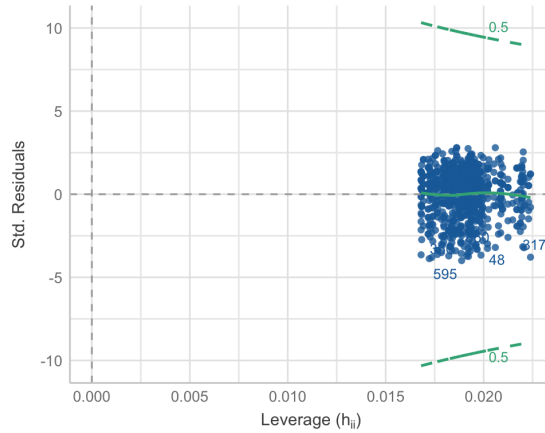
Homogeneity of Variance

Reference line should be flat and horizontal



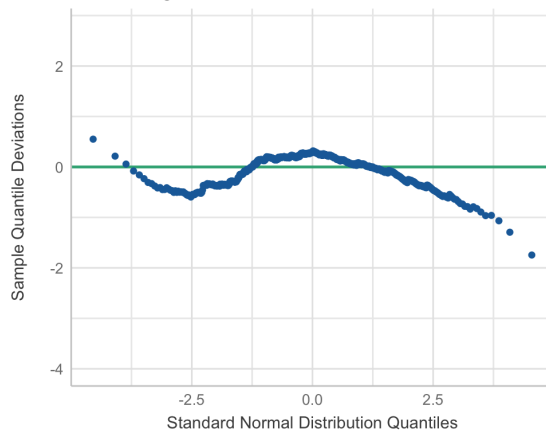
Influential Observations

Points should be inside the contour lines



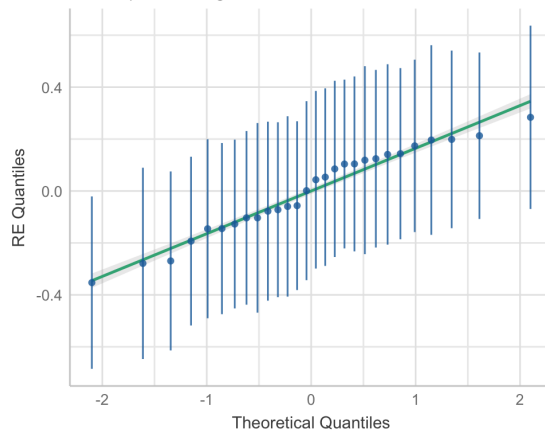
Normality of Residuals

Dots should fall along the line

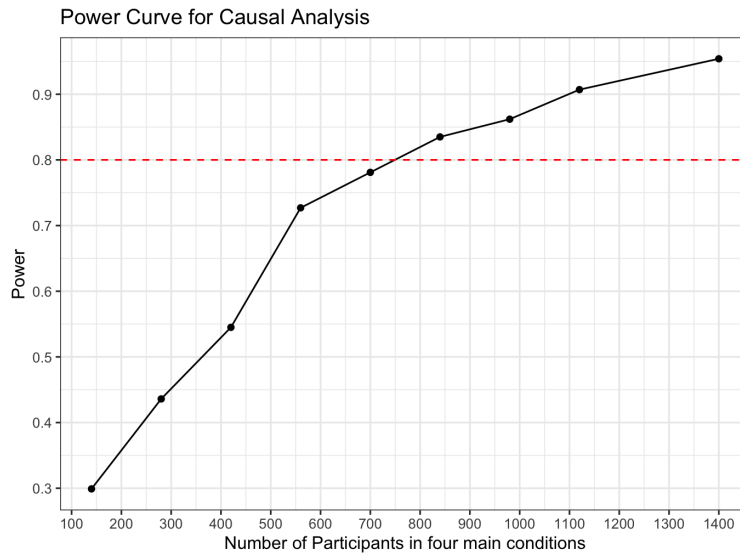


Normality of Random Effects (Stimuli)

Dots should be plotted along the line



Supplementary Figure 5. Residual diagnostics for the primary mixed-effects model. Panels display residuals vs. fitted values, scale–location, Q–Q plots for residuals and random effects, and influence diagnostics. Assumptions were broadly supported, with only mild residual non-normality.



Supplementary Figure 6. Power curve with estimated experimental power for different counts of participants, based on 1,000 simulations per level.

Here is the updated prompt for causal language classification of these research summaries from participants:

Causal language can appear in research summaries in different forms:

- 1 Direct causal statement - The authors explicitly claim one variable affects another (e.g., ``influences,' ' ``causes,' ' ``impacts,' ' ``mechanisms'').
- 2 Implied causal introduction - The authors state that the study investigates a causal relationship (e.g., ``this study examines the influence'').
- 3 Conditional causal language - The authors make causal claims but hedge their language (e.g., ``may cause,' ' ``suggests a role of'').
- 4 Implied causal conclusion - The authors imply causality by recommending interventions based on their findings.
- 5 None - The authors use purely associational language when discussing their findings.

Instructions:

Code the presence of causal language in this summary based on the categories above. Ignore causal language referring to prior studies or general background. Focus only on how they describe the novel, empirical results.

If the summary explicitly states the results are purely associational or that future work is necessary to establish causality, code as 5.

Respond with a single number. If multiple types are present, report the lowest applicable number.

Second Experiment with LLMs - Prompts

In the second experiment, we examined whether the likelihood of causal claims in GPT-generated summaries varied by prompt type. Four prompt strategies were tested, as shown below:

Simplified:

""Summarize the text for a general reader (~8th-grade level). Use plain, everyday words and short sentences. Focus on the most important idea(s), avoiding jargon and numbers unless essential. Output one paragraph ~50 words (45-60).

No preamble or bullets.

TEXT:""

Practical:

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

TEXT:""

Basic:

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

TEXT:""

Careful:

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

TEXT:""

Second Experiment with LLMs – Causal Classifier

In reviewing the summaries generated by GPT-4.1, we observed that the language differed substantially from that of academic texts, particularly in the Simplified and Practical summary conditions. Therefore, we developed a new causal language classifier specifically tailored to these summaries.

We began by hand-coding 40 summaries as a training sample to explore how causal language appeared in these summaries and to refine our prompt for GPT-4.1. Three categories remained conceptually clear: Direct, Conditional, and Associational/Descriptive. We dropped the “Implied” category because, within the summaries, such claims could consistently be recoded into one of the other three: they were either phrased conditionally, articulated strongly enough to qualify as direct, or embedded in careful associational language. For any remaining cases that resembled the original “Implied” category, we grouped them with Direct. We therefore changed the label of direct to “Unhedged Causal” to reflect this difference. Although each summary was brief (approximately 50 words), some contained multiple statements about results; each summary was assigned a single label corresponding to the strongest causal claim present. We then developed a GPT-based classifier to reproduce human judgments, iteratively refining the prompt on the training sample. In our final prompt, GPT-4.1 received the classification prompt and the summary text (temperature = 1, max_tokens = 5, n = 5, all other parameters set to the default Batch API values as of November 5, 2025), producing 5 categorization labels. The majority label across the five outputs was assigned as the final classification for each summary.

Next, we drew a stratified random sample of summaries from each prompting strategy (Total N = 240; N = 60 per prompt), further stratified by the GPT-assigned causal category (N = 80 each) to account for the scarcity of conditional claims. Two human coders independently classified each summary into one of the three categories (Unhedged Causal, Conditional Causal, or Associational/Descriptive), blind to GPT’s labels. Inter-coder agreement was moderate (Krippendorff’s $\alpha = 0.68$) on this stratified sample. The automated GPT-based labels showed moderate agreement with the consensus label from human coders on the held-out validation set (Macro F1 = 0.75, Precision = 0.75, Recall = 0.78, Accuracy = 0.79). Notably, in cases of mismatch, the human coders were substantially more likely to label summaries as containing a direct claim than GPT. There were 44 cases where the human’s labeled a sample as “Direct” when GPT said either conditional or associational, and only 8 with the reverse. The full classification prompt is provided below, and was applied to the full sample of 4,000 summaries.

GPT-4.1 Prompt to Identify Causal Claims in LLM Summaries

You will read a summary of an academic paper that may describe the results in simplified language and discuss practical implications. Your task is to decide whether the summary conveys a causal relationship or implied mechanism.

A causal relationship may be expressed through words such as ``influences,'' ``affects,'' ``helps,'' ``boosts,'' or ``mediates an impact,'' among many others. It may also be implied by statements linking actions or policies to outcomes (e.g., ``if people do X, Y will happen'') or by describing a study as investigating the impact of one variable on another. If one variable increases/decreases another or if X leads to Y, this is also to be treated as causal.

If the only causal language is qualified with words like ``may,'' ``might,'' ``possibly,'' ``could,'' or ``suggests,'' classify the summary as ``Conditional.'' If it includes both hedged and unhedged causal language, classify it as ``Causal.''

Do not consider participants' reported beliefs about causes as causal claims. If it is unclear whether causality is implied, err on the side of ``Non-Causal.'' Similarly, references to one variable 'affecting/moderating/mediating a relationship' (e.g., interaction effects) should be interpreted as non-causal as this language describes a statistical interaction, not a causal mechanism.

If a study makes a causal claim and later mentions limitations with regard to causality, it should still be classified as causal. If it only mentions limitations, it should be "Non-Causal."

Return only one of the following labels: Causal, Conditional, or Non-Causal.

Second LLM Experiment - GPT-4.1 Robustness Checks

To evaluate the sensitivity of our second experiment with LLMs to the measurement strategy, we repeated the analysis with two other causal language classification strategies and found similar results, looking exclusively at the data produced by GPT-4.1. We present these approaches below.

We take two similar approaches that mirror the full text analysis conducted in the main text. Specifically, for each summary, we split the text into sentences using the sentence tokenizer from the Python package spaCy. We then ran all sentences through our BERT-based causal language classifier, as well as the causal language classifier developed by Yu et al., (2019) trained on the conclusion section of biomedical abstracts. We searched for the same set of hedges used in the full text analysis, labeling any causal claims with a hedge as Conditional, and those without as Causal. For every summary, we collapsed across sentences, assigning the strongest level of causal language present. Because 5 summaries were generated for each abstract, prompt, content (full text vs abstract), we assigned the majority label from these 5 summaries, and randomly choose between the majorities in the event of a tie.

Across the prompting strategies, we find similarly high amounts of causal claims for both strategies. Our classifier labeled the summaries as containing a direct, unhedged causal claim in 44.1% of the abstracts (95% CI = [40.7%, 47.6%]), and Yu's found it in 53.1% (95% CI = [49.7%, 56.6%]). For both, this varies substantially by prompting strategy (Our classifier $\chi^2(3) = 103.6, p < 0.001$; Yu $\chi^2(3) = 62.2, p < 0.001$) and by the abstract level label (Our classifier $\chi^2(3) = 139.5, p < 0.001$; Yu $\chi^2(3) = 146.0, p < 0.001$). Proportions and confidence intervals for each prompt and abstract-level causal label are presented in the table below. Most notably, summaries of articles with abstracts that only contained conditional causal claim had a direct causal claim in nearly half of the abstracts for both approaches (40.5%, 95% CI = [33.7%, 47.3%] for our classifier, 50.0%, 95% CI = [43.1%, 56.9%] for Yu's). And summaries of abstracts that contained only descriptive claims, with no causal language, still frequently had causal claims (14.5%, 95% CI = [9.6%, 19.4%] with our classifier; 22.0%, 95% CI = [16.3%, 27.7%] for Yu's).

Thus the key findings, that GPT-4.1 summaries are very likely to contain causal overstatement, with notable variation by prompt strategy and abstract-level causal label, remains substantiated for these alternative measurement approaches. The largest differences emerge for the simplified and practical prompts, both of which were under-estimated by the models trained on academic texts, relative to our updated LLM-classification prompt, likely due to the difference in text characteristics between academic texts and simplified summaries.

Second LLM Experiment - GPT-4.1 Claim Level Analysis

Academic articles typically include multiple claims about their findings. Our experiment provides suggestive evidence that GPT tends to emphasize causal ones, but it is also important to understand how the distribution of claim types may correspond

Supplementary Table 6. Comparison of Proportions Across Key Dimensions

Key Dimension	Category	Our Prop. [95% CI]	Yu Prop. [95% CI]
Prompt	Basic	72.0% [65.8, 78.2]	68.5% [62.1, 74.9]
	Careful	28.0% [21.7, 34.2]	30.5% [24.1, 36.9]
	Simplified	29.0% [22.7, 35.2]	57.0% [50.1, 63.9]
	Practical	47.5% [40.6, 54.4]	56.5% [49.6, 63.4]
Abstract-level Causal Label	Direct	72.5% [66.3, 78.7]	81.5% [76.1, 86.9]
	Implied	49.0% [42.1, 55.9]	59.0% [52.2, 65.8]
	Conditional	40.5% [33.7, 47.3]	50.0% [43.1, 57.9]
	Descriptive	14.5% [9.6, 19.4]	22.0% [16.3, 27.7]

with differences in summaries. Using our sentence-level claim coding, we found that abstracts contained an average of 1.8 distinct claims, 69% of which were descriptive, 14.5% conditional causal, and 17.5% unhedged causal or direct. Full texts contained substantially more claims ($M = 31.3$), of which 73.0% were descriptive, 11.5% conditional, and 15.5% direct.

To examine whether the mix of claim types in an article predicted whether GPT’s summary included an unhedged causal statement, we estimated a binomial GLM. Observations were nested within articles, with repeated measures across prompting strategies and the abstract/full-text condition. Because our focus here is not on differences among prompts, we pooled across them and included fixed effects for input type (full text vs. abstract) and its interaction with the total number of claims, along with predictors for the proportion of unhedged causal, conditional causal, and descriptive claims. A random intercept for article accounted for within-article correlation. We used proportions rather than raw counts to avoid penalizing longer or shorter texts. As the number of claims was not experimentally manipulated, these findings should be interpreted as associational rather than causal. The model was estimated using a Bayesian mixed-effects framework, and we report posterior means and 95% credible intervals for each coefficient.

The results showed several strong effects. The intercept was 0.57 (95% CrI = [0.39, 0.75]), FullText = 0.05 (95% CrI = [-0.19, 0.31]), Unhedged = 1.59 (95% CrI = [0.80, 2.39]), Conditional = -0.13 (95% CrI = [-0.79, 0.52]), Descriptive = -0.89 (95% CrI = [-1.10, -0.67]), TotalClaims = -0.01 (95% CrI = [-0.017, -0.003]), and the interaction between TotalClaims and FullText = 0.02 (95% CrI = [0.01, 0.03]). These coefficients represent log-odds. In general, a higher proportion of direct claims increased the probability that GPT produced an unhedged causal statement, whereas a higher proportion of descriptive claims decreased it. To interpret these results, we generated predictions using dummy data. For an abstract containing three descriptive-only claims, the model predicts a 41.4% probability that GPT’s summary includes an unhedged causal statement. If one of the three claims is direct (33%), the probability rises to 61.6%. For full texts, when ten claims are all descriptive, the predicted probability is 46.1%; if one-fifth are direct, it rises to 58.4%, and if half are direct, to 74.7%.

In sum, GPT’s summaries tend to introduce causal framing even when the underlying claims are primarily descriptive; however, this tendency increases a great deal when even a modest fraction of the article’s claims are explicitly causal.

Second Experiment with LLMs - Additional Tables

Supplementary Table 7. Agreement of Causal Classifications for Summaries Generated when provided with the Full Texts Compared to Only the Abstract

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

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

Supplementary Table 8. Contingency tables comparing abstract and LLM summary causal classifications

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

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

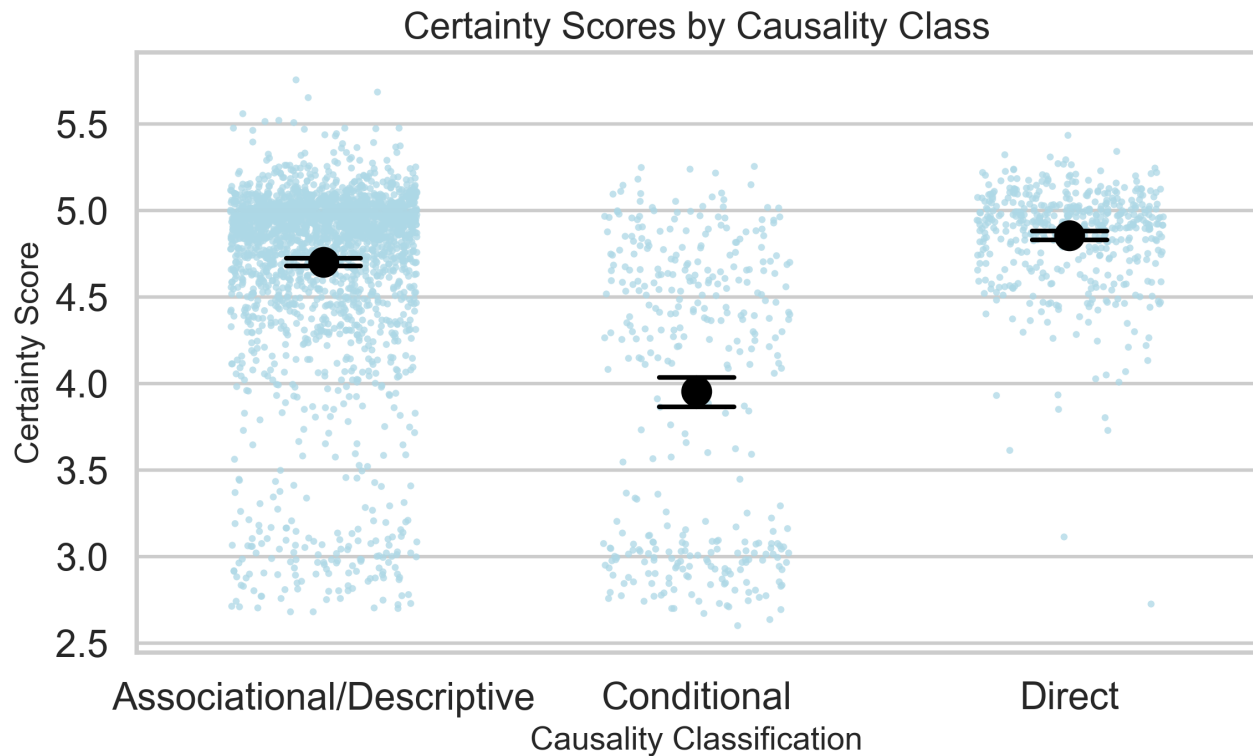
Supplementary Table 9. Multilevel Logistic Regression Predicting Unhedged Causal Claims in Summaries by Model, Prompt, Their Interaction, and Abstract Causal Category

	OR (95% CIs)	p-value
Intercept Odds (Reference: gpt-5.1, Basic prompt, Descriptive abstract; not OR)	0.06 [0.03, 0.16]	<0.001
Model: Claude Sonnet 4.6	0.78 [0.36, 1.71]	0.542
Model: DeepSeek-Chat-V3.2	2.18 [1.01, 4.71]	0.046
Model: GPT-4.1	2.36 [1.09, 5.09]	0.029
Model: GPT-5.4	0.17 [0.07, 0.39]	<0.001
Prompt: Careful	0.04 [0.01, 0.10]	<0.001
Prompt: Practical	1.73 [0.80, 3.74]	0.161
Prompt: Simplified	2.36 [1.09, 5.09]	0.029
Abstract causal category: Conditional	6.04 [2.24, 16.31]	<0.001
Abstract causal category: Implied	11.88 [4.38, 32.22]	<0.001
Abstract causal category: Direct	178.92 [62.16, 514.96]	<0.001
Claude Sonnet 4.6 × Careful	2.57 [0.70, 9.49]	0.157
DeepSeek-Chat-V3.2 × Careful	0.71 [0.19, 2.63]	0.608
GPT-4.1 × Careful	2.34 [0.66, 8.29]	0.188
GPT-5.4 × Careful	1.42 [0.29, 6.89]	0.664
Claude Sonnet 4.6 × Practical	3.79 [1.26, 11.36]	0.017
DeepSeek-Chat-V3.2 × Practical	1.36 [0.46, 4.03]	0.578
GPT-4.1 × Practical	1.36 [0.46, 4.04]	0.575
GPT-5.4 × Practical	4.41 [1.41, 13.79]	0.011
Claude Sonnet 4.6 × Simplified	1.09 [0.37, 3.24]	0.873
DeepSeek-Chat-V3.2 × Simplified	0.79 [0.27, 2.32]	0.665
GPT-4.1 × Simplified	0.62 [0.21, 1.84]	0.392
GPT-5.4 × Simplified	3.24 [1.04, 10.11]	0.043

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

Causal Claims and Uncertainty

Do causal claims carry more or less certainty than non-causal claims? Applying the certainty-estimator package (Citation 77 main text) to all findings from our full-text analysis, we found that causal claims were expressed with reliably lower certainty ($M = 4.47$, $SD = 0.73$, $N = 847$) than non-causal claims ($M = 4.70$, $SD = 0.54$, $N = 2,287$), with a modest but significant effect size (Cohen's $d = -0.36$, $p < .001$). This difference was driven largely by conditionally causal claims ($M = 3.95$, $N = 360$), which showed markedly lower certainty than direct causal claims ($M = 4.86$, $N = 487$). Direct causal claims, in turn, exhibited higher certainty than non-causal claims. The limited overlap in certainty scores across these three categories suggests they represent meaningfully distinct phenomena warranting separate analysis. Supplementary Information Figure 7 illustrates the level of certainty across causal claim types.



Supplementary Figure 7. Uncertainty scores for individual claims (light blue) and aggregated by causal classification (black), with means and 95% confidence intervals. The dataset includes 2,287 Associational/Descriptive claims, 360 Conditional, and 487 Direct Causal statements.