Sarol, M.J., Schneider, J. and Kilicoglu, H. (2025), Automatic Identification of Citation Distortions in Biomedical Literature: A Case Study. Proceedings of the Association for Information Science and Technology, 62: 595-606. https://doi.org/10.1002/pra2.1281

Automatic Identification of Citation Distortions in Biomedical Literature: A Case Study

AUTHORS SECTION

Sarol, M. Janina University of Illinois Urbana-Champaign, USA | mjsarol@illinois.edu

Schneider, Jodi University of Illinois Urbana-Champaign, USA | jodi@illinois.edu

Kilicoglu, Halil University of Illinois Urbana-Champaign, USA | halil@illinois.edu

ABSTRACT

Citations are central to the propagation of scientific information. Ensuring the accuracy of citations is essential to maintain the credibility of scientific knowledge. However, assessing citations is a significant challenge, especially at large scale. This study examines the utility of natural language processing (NLP) in identifying poor citation practices. Specifically, we replicate Greenberg's 2009 study on citation distortions in Alzheimer's research, which demonstrated how poor citation practices can contribute to the establishment of unsubstantiated claims as facts. We explored two approaches: one that utilizes large language models (LLMs), and another that relies on existing publicly available NLP tools and publication metadata. Our findings suggest that, among Greenberg's three types of citation distortion – citation bias, amplification, and invention – automated approaches are most effective at detecting amplification, with more limited success in automatically replicating Greenberg's results for the other citation distortion types. Further refinements to LLM pipelines are needed to better capture the subtleties of citation bias and invention in biomedical publications.

KEYWORDS

Natural language processing; large language models; citation distortion; citation accuracy

INTRODUCTION

Citations serve an important role in the scientific ecosystem, facilitating dissemination of knowledge and the flow of information (Zhuge, 2006). For the scientific ecosystem to function effectively and for science to be trustworthy, it is important that the information being propagated is accurate. Citations must be backed by valid evidence from the literature being referenced, and they must faithfully represent the cited work's ideas and findings (Ngatuvai et al., 2021). It is crucial to detect issues stemming from poor citation practices early, as misinformation can make its way into the body of scientific literature and ultimately undermine the integrity of scientific knowledge.

While poor citation practices can greatly impact any field, their effects can be particularly alarming in the biomedical field. The consequences of poor citation practices can have long-lasting effects on human health. One striking example is the pivotal role they may have played in the opioid crisis (Leung et al., 2017). A single paragraph letter published in *The New England Journal of Medicine* suggested that the likelihood of developing addiction was rare among patients without history of substance use disorders (Porter & Jick, 1980); this letter, despite not presenting rigorous evidence, was widely cited, shaping pain management practices for years and contributing to widespread overprescription (Leung et al., 2017).

In another case, Greenberg (2009) documented three questionable citation practices, collectively termed *citation distortions*: (1) *citation bias*, the selective citation of primary data papers that supported a particular claim, while ignoring papers critical of the claim; (2) *amplification*, the growing acceptance of the claim through citation of secondary sources, mostly review papers, instead of direct evidence; and (3) *invention*, where authors altered the nature of the claim and its evidence in various ways, such as citing meeting abstracts as if they were peer-reviewed publications. Greenberg (2009) stated that these citation distortions lead to the establishment of an unsubstantiated claim in Alzheimer's research, specifically the claim that individuals with inclusion body myositis produce β-amyloid (a protein linked to Alzheimer's disease), which damages their skeletal muscles.

Greenberg's is not the only study of this nature. Ravnskov (1992) observed citation bias: clinical trial publications supporting the claim that lowering cholesterol prevents coronary heart disease were cited at a higher rate than those presenting contrary findings. While citing supporting claims is not inherently problematic, it becomes concerning when the consensus is shaped by preferential citation, especially in systematic review and meta-analysis papers, which are expected to offer unbiased syntheses of existing evidence (Ahn & Kang, 2018). Such examples raise an important question: *how many other instances of unsubstantiated claims exist within the scientific literature due to poor citation practices?* Ideally, every claim could be systematically traced and verified, but this task is daunting even for domain experts because manually verifying citations is a labor-intensive process that becomes impractical when dealing with thousands of citations.

Advances in computational methods, particularly natural language processing (NLP), offer significant potential for identifying citation distortions and tracing unsubstantiated claims. NLP methods have been developed for automated fact verification (e.g., Thorne et al., 2018; Krishna et al., 2022), including approaches specifically tailored for biomedical publications (e.g., Wadden et al., 2022). These methods may be used to detect certain types of citation distortion, particularly cases where the citing text alters or misrepresents claims from the cited publication (Sarol et al., 2024). In addition, large language models (LLMs) have demonstrated effectiveness across various tasks without task-specific training, including applications in the biomedical domain (Jahan et al., 2024), offering a promising approach for detecting citation distortions.

In this study, we examine how NLP, including LLMs, can be leveraged to identify citation distortions. Specifically, we aim to automatically replicate Greenberg's (2009) manual analysis. We seek to identify suitable tools and evaluate their effectiveness in identifying the three types of citation distortions Greenberg (2009) documented. Our broad research question is: *can NLP tools help identify citation distortions in scientific literature?*

THE GREENBERG STUDY

Greenberg (2009) constructed a citation network of 242 publications that addressed a widely accepted claim about Alzheimer's disease: *β-amyloid, a protein accumulated in the brain in Alzheimer's disease, is produced by and injures skeletal muscle of patients with inclusion body myositis*, which we refer to below as "the claim". Although this claim was widely accepted in Alzheimer's literature, Greenberg demonstrated that the claim lacked rigorous empirical foundation and that its acceptance in the Alzheimer's literature was ultimately unwarranted. This distortion was largely driven by three citation-related factors, detailed below.

Citation Bias

Greenberg (2009) found that papers supporting the claim were heavily cited, whereas papers that were critical of the claim received far fewer citations. 12 papers were classified as primary data; these papers contained experimental data that directly addressed the claim. Half of the primary data papers were critical, and half were supportive, but the 6 critical papers each received fewer than 5 citations, while the 6 supporting papers received at least 20 citations. In total, the supporting papers received 94% of the 214 citations to primary data papers.

Citation Amplification

Greenberg (2009) defined citation amplification as the phenomenon in which an unsubstantiated claim gains more traction through citation to papers that do not directly contain evidence addressing the claim. Amplification happens when secondary sources that mention the claim, such as review papers, are cited instead of citing primary data papers, and when supportive claims are more frequently reiterated in citations. In the Greenberg study, 95% of citation paths include four review papers, with one review paper having 63% of the citation paths flow through it. That review paper became an authoritative source with respect to the claim, even though it did not report any experimental findings about the claim.

Citation Invention

In citation invention, the citing authors misrepresent a claim as being supported by a peer-reviewed publication, either through distorting the publication's findings and speculations or through misleading the reader that the cited publication has undergone rigorous peer-review. Greenberg (2009) identified five different ways in which information about a claim may be invented (number of cases for each invention type is provided in parentheses):

1. *Citation diversion*— the source publication's content is altered in a manner that changes its implications (4 cases)
2. *Citation transmutation*— hypotheses or speculations become stated as facts through citations (17 cases)
3. *Back door invention*— abstracts are misrepresented as peer-reviewed publications creating a misleading impression that the findings have been subject to peer review (12 cases)
4. *Dead end citation*— the source publication is cited in support of a claim, but it lacks any content acknowledging the claim (9 cases)
5. *Title invention*— results are reported in the source publication's title, but no evidence for these results is presented in the publication's main text (1 case)

METHODS

Data

The Greenberg (2009) study served as our ground truth for evaluating whether NLP tools can effectively identify citation distortions. We sought to collect the following information for all 242 papers: publication metadata, citation information, and full text of the papers. Publication metadata (title, abstract, publication type, and MeSH terms) were retrieved from PubMed using the Entrez Programming Utilities API. We obtained the citation information (citing paper, cited paper, citation context – the text describing the cited work) from the supplementary material of

Greenberg (2009). We manually collected the full text of all papers in plain text (preferred) or PDF format. Out of 242 papers, we collected the full text of 241 (237 in PDF format) and were unable to locate one. As this paper was not cited by any of the other papers in our list and because the full text was only needed for one of our NLP tasks, its absence had minimal impact on the results.

PDF documents were converted into plain text format using the Adobe PDFExtract API. We encountered conversion issues for 32 documents. We first converted these 32 documents from PDF to rich text format (RTF) using Adobe Acrobat and then converted the RTF to plain text format. We manually removed any text that was not part of the main article text content, i.e., author list, references, acknowledgements, tables and figures (including captions). We found two common PDF-to-text conversion issues in our dataset, which we manually corrected: (1) paragraphs that were incorrectly combined and (2) paragraphs that were incorrectly split (which typically happens when the paragraph continues to the next page). In rare cases, we also manually corrected the order of paragraphs.

There were 808 total citation instances across 672 unique citing-cited publication pairs. 212 papers cited at least one other paper within the network, and 96 papers were cited by at least one other paper within the network.

Replication Approaches

We identified four key tasks that are needed to detect types of citation distortion:

1) Publication Type and Topic Classification: The goal of this task is to classify each of the 242 papers into one of four types, following Greenberg's (2009) categorization: primary data papers (n=12), myositis review papers (n=63), animal/cell culture model papers (n=17), and other (n=150). This process also involved topic classification (e.g., identifying which review papers are about myositis). This categorization is needed to identify citation bias and amplification. To assess citation bias, one must first identify the 12 primary data papers presenting experimental evidence and classify them based on whether they support or refute the claim. Only after identifying the primary data papers can we perform a comparison of the citation counts of the supportive and critical primary data papers. To detect the presence of amplification, we focus on identifying secondary sources, particularly myositis review papers.
2) Stance Detection: Given a piece of text, stance detection aims to identify whether the text's author supports, opposes, or remains neutral toward a particular subject (Mohammad et al., 2016). We used stance detection with respect to the claim to detect amplification. Amplification occurs when citations disproportionately reference statements that support the claim. Citations to these statements tend to predominantly have a supportive stance, rather than critical or neutral.
3) Citation Accuracy Classification: The goal of this task is to classify citations as accurate or inaccurate (Sarol et al., 2024). Inaccurate citations are also known as "quotation errors" (Jergas & Baethge, 2015). Classifying citations for accuracy can help detect whether three of the five invention issues occur: citation diversion, citation transmutation, or dead end citation.
4) Scientific Claim Verification: Originally developed for retrieving relevant scientific abstracts, this task aims to identify literature that contains evidence – either supporting or refuting evidence – for a given claim (Wadden et al., 2020). We repurpose this task to detect title invention issues by treating the title as the claim and the remainder of the article as the potential body of evidence. If no relevant statements are found within the article, the article is considered an instance of title invention.

In this replication study, we compared two distinct approaches: one that primarily leveraged LLMs and another that combined existing available tools and data. For the LLM-based approach, we used GPT-4o (*gpt-4o-2024-08-06*; OpenAI et al., 2024). Prompts were manually crafted to align as closely as possible with the original definitions and wording used in the Greenberg study. Each prompt was then refined by submitting it to the LLM model with the instruction: *"Please improve this prompt:"* The resulting output was then used as the prompt for this study. Non-LLM approaches for each of the key tasks are described in detail below. We did not attempt to replicate the backdoor invention results, as that would require checking against external bibliographic databases.

Key Task 1: Publication Type and Topic Classification

In both LLM and non-LLM approaches, we followed the flowchart shown in Figure 1. We first classified papers into myositis review papers or non-review papers, then identified model papers, followed by primary data papers; the remaining papers were assigned to the "other" category.

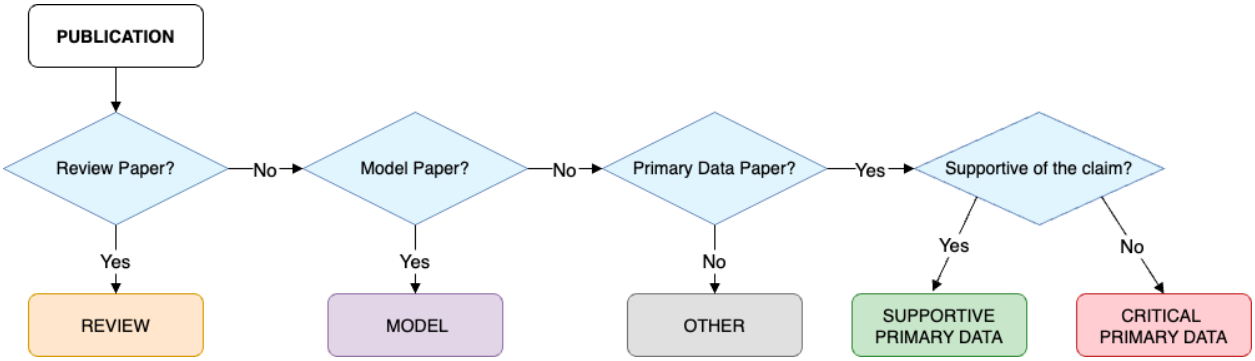

Figure 1. Decision Tree for Determining Paper Type

Identifying myositis review papers

Greenberg (2009) defined myositis review papers as '*review papers with "myositis" or the equivalent in their title*'.

Non-LLM approach: Classification primarily relied on publication metadata. We first identified review papers by examining the publication types listed in the metadata. If one of the publication types is *Review*, then the paper was classified as a review paper. To narrow to myositis review papers, we used the following criteria: whether at least one of the following MeSH terms was assigned: *Myositis* and *Myositis, Inclusion Body*, *Polymyositis,* or *Dermatomyositis* (i.e., all the myositis-related terms mentioned in Greenberg (2009)).

LLM approach: We prompted the LLM to classify whether each paper is a myositis review paper based on its title and abstract. Each paper was submitted to the model separately. The prompt is shown in Box 1.

```
You are an expert in identifying whether a paper is a myositis review paper.

Definition of a myositis review paper: A review paper that includes "myositis" or an
equivalent term in its title. A review paper summarizes, analyzes, and synthesizes
existing research on a particular topic without presenting original experimental
data. It typically provides an overview of findings from various studies and offers
insights, conclusions, or directions for future research.

Analyze the provided title and abstract to determine if the paper meets the criteria
of a myositis review paper.

Output only one of the following:

1 - The paper is a myositis review paper (a review paper with "myositis" or an
equivalent term in the title).

0 - The paper is not a myositis review paper (it presents original research or does
not have "myositis" or an equivalent term in the title).

Input Format:

The title will be enclosed in <title></title>.

The abstract will be enclosed in <abstract></abstract>.

Output only 1 or 0, with no additional text or explanation.
```

Box 1. LLM Prompt for Classifying Review Papers

Identifying animal/cell culture model papers

Greenberg (2009) defined animal or cell culture model papers as those "*reporting cell culture or animal model experiments*".

Non-LLM approach: We deemed a paper to be an animal/cell culture model paper if it was assigned the following MeSH terms: *Animals* and *Cells, Cultured*.

LLM approach: We instructed the LLM to determine whether a paper is an animal or cell culture model paper by passing its title and abstract to the prompt. We did not attempt to distinguish between animal and cell culture studies. The prompt used for this task is shown in Box 2.

```
You are an expert in identifying whether a paper involves animal or cell culture
experiments. A paper involves animal or cell culture experiments if it includes
studies conducted on non-human living organisms (animals) or in vitro cell cultures
to investigate biological processes, disease mechanisms, or treatment effects.

Analyze the provided title and abstract and determine if the paper involves animal
or cell culture experiments. Output only one of the following:

1 – The paper involves animal or cell culture experiments (excluding human studies).

0 – The paper does not involve animal or cell culture experiments.

Input will include the title and abstract. Output only 1 or 0, with no additional
text or explanation.
```

Box 2. LLM Prompt for Classifying Animal and Cell Culture Model Papers

Identifying primary data papers addressing the claim

Greenberg (2009) defined primary data papers as papers "*containing experimental data addressing the specific and abnormal presence of these molecules in inclusion body myositis muscle*".

Non-LLM approach: To our knowledge, there are no existing tools specifically designed to determine whether a paper is a primary data paper. Therefore, we adapted a tool developed for a different purpose. We used the MultiVerS (Wadden et al., 2022) model, which was originally developed for scientific claim verification. Given two inputs—a claim and a set of abstracts—MultiVerS tries to identify evidentiary sentences that support or refute the claim. In our adaptation, we replaced abstracts with all paragraphs from the full text. We also excluded citation sentences from the output using regular expressions, to ensure that the evidence came from the paper itself. A paper was labeled as a supportive primary data paper if most evidence sentences support the claim; a paper was labeled as a critical primary data paper if most evidence sentences refute it. If no evidence sentences were found, or if there was a tie between supporting and refuting evidence, then the paper was not classified as a primary data paper. Only papers that included *Journal Article* as one of their publication types in the metadata were considered eligible for primary data classification.

LLM approach: Because the title and abstract likely do not provide enough context to determine whether a paper is a primary data paper, we provided the model with the full text. We also included the claim in the prompt, which is shown in Box 3.

```
You are an expert in evaluating scientific literature to determine whether a paper
presents primary experimental data relevant to a specific biomedical claim. The
claim under investigation is: Beta-amyloid precursor protein (APP) mRNA or protein,
or β-amyloid protein, is abnormally present in inclusion body myositis (IBM) muscle.

A primary data paper is one that includes original experimental data directly
assessing the presence of these molecules in IBM muscle. You will analyze the given
paper to determine:

Whether it is a primary data paper (i.e., it contains experimental results
addressing the claim).

If it is a primary data paper, whether its findings support or refute the claim.

Your response must be one of the following, with no additional explanation:

SUPPORT – The paper presents experimental data confirming the abnormal presence of
these molecules in IBM muscle.

REFUTE – The paper presents experimental data contradicting the claim.

NOT A PRIMARY DATA PAPER – The paper does not contain original experimental data
addressing the claim.

The full text of the paper is provided below. Only output the category, without any
explanation.
```

Box 3. LLM Prompt for Classifying Primary Data Papers

Key Task 2: Stance Detection

Non-LLM approach: Currently, there are no stance detection tools specifically developed for biomedical texts. Therefore, we evaluated two alternative approaches: (1) a stance detection model developed by Riedel et al. (2017) for fake news detection, and (2) a clinical citation sentiment classification model by Kilicoglu et al. (2019).

Although the latter was not specifically designed for stance detection, it performs a similar task by analyzing the tone of citation contexts and classifying them as positive, negative, or neutral.

The stance detection model from Riedel et al. (2017) uses bag-of-words features passed to a multi-layer perceptron to predict the stance of a news article given its headline and main article body. We treated the claim as the headline and the citation context as the article body. The model outputs one of four labels: *agree*, *disagree*, *discuss*, or *unrelated*. We mapped *agree* to supportive papers, *disagree* to critical papers, and both *discuss* and *unrelated* to neutral papers.

The citation sentiment classification model from Kilicoglu et al. (2019) uses a convolutional neutral network model to predict the sentiment of a citation. Each citation context was passed to the model, which classifies sentiment into one of three categories: *POSITIVE*, *NEGATIVE*, or *NEUTRAL*. We mapped *POSITIVE*-labeled instances to supportive papers, *NEGATIVE* to critical papers, and *NEUTRAL* to neutral papers.

LLM approach: The LLM was instructed to assess the stance of the citation context toward the claim. We provided both the claim and citation context in the prompt (shown in Box 4), along with the definition of stance detection. We used the label *CRITICAL* for opposing texts to better align with the terminology used in Greenberg's study.

Citation contexts were passed individually to the model. Some papers cited the same publication multiple times; in these cases, we applied majority voting. If the majority of the contexts were supportive, the stance was labeled as supportive; if the majority were critical, the stance was labeled as critical. In case of ties or when all citation contexts were labeled as neutral, the stance was classified as neutral.

```
You are an expert in stance detection for citation contexts with respect to a
specific scientific claim.

Definition of stance: Stance refers to the position or attitude expressed toward a
particular claim—whether the text supports, challenges, or remains neutral toward
it.

Claim: Beta-amyloid precursor protein (APP) mRNA or protein, or β-amyloid protein,
is abnormally present in inclusion body myositis (IBM) muscle.

Your task is to determine the stance expressed in the citation context toward the
claim above.

SUPPORT – The citation context supports or agrees with the claim.

CRITICAL – The citation context refutes, contradicts, or challenges the claim.

NEUTRAL – The citation context is descriptive, does not take a stance, or does not
clearly support or challenge the claim.

Input Format:

The citation marker will be enclosed in <citation_marker></citation_marker>.

The citation context will be enclosed in <citation_context></citation_context>.

Analyze the citation context in relation to the claim and output only one of the
following: SUPPORT, CRITICAL, or NEUTRAL, with no additional text or explanation.
```

Box 4. LLM Prompt for Stance Detection

We note that not all 808 citation instances were assigned stances by Greenberg (2009). Three citation contexts labeled as citing incorrect information were not assigned labels. Of the remaining 669 unique citing–cited publication pairs corresponding to 805 citation contexts, Greenberg labeled 631 as supportive, 20 as critical, and 18 as neutral.

Key Task 3: Citation Accuracy Classification

Non-LLM approach: For citation diversion and dead end citation, we used a publicly available citation accuracy classification model that we previously developed (Sarol et al., 2024). This model takes as input the citation context along with the full text of the cited paper and classifies the citation context into one of three categories: *ACCURATE*, *INACCURATE*, and *IRRELEVANT*. A citation is classified as *INACCURATE* when the citation context alters or misrepresents the content of the cited work. *IRRELEVANT* label indicates that the citation context refers to content not found in the cited paper. Based on these definitions, we deemed *INACCURATE* predictions as citation diversion cases and *IRRELEVANT* predictions as instances of dead end citations. We did not attempt to replicate citation transmutation using non-LLM methods, as we did not find any suitable, publicly available tools for this task.

LLM approach: For citation diversion, citation transmutation, and dead end citation, we included the citation context, the full text of the cited paper, and the definition of the particular citation invention issue in the prompt. A sample prompt used for one of the citation invention cases (citation diversion) is shown in Box 5. We preserved the exact wording of the definitions in Greenberg (2009) in our prompts.

```
You are an expert in identifying citation diversion cases.
Definition of citation diversion: Citing content but claiming it has a different meaning, thereby diverting its implications.
Task:
Analyze the provided citation context and full text of the referenced article.
Determine if the citation accurately reflects the original meaning of the cited content or if it distorts its implications.
Output only one of the following:
1 – There is a citation diversion issue (the citation misrepresents the meaning or implications of the original content).
0 – There is no citation diversion (the citation correctly reflects the original meaning or implications).
Input Format:
The citation context will be enclosed in <citation_context></citation_context>.
The full text of the referenced article will be enclosed in <paper></paper>.
Output only 1 or 0, with no additional text or explanation.
```

Box 5. LLM Prompt for Citation Diversion

Key Task 4: Scientific Claim Verification

Non-LLM approach: To identify title invention cases, we used the MultiVerS (Wadden et al., 2022) model, with the paper title as the claim and each paragraph of the paper, including the abstract, as the set of potential evidence. If any of the content in the paragraphs are deemed supportive, then the title accurately reflects the content of the paper. Otherwise, we deemed the paper to be an instance of title invention. We applied this model to all 241 papers with full text data.

LLM approach: We passed the title, full text of each paper (title excluded), and definition of title invention – as described in Greenberg (2009) – to the model. The prompt used for this task is shown in Box 6.

```
You are an expert in detecting title invention in scientific papers.
Definition of title invention: Reporting of "experimental results" in a paper's title, even though the paper does not report the performance or results of any such experiments.
Task:
Analyze the given title and paper content.
Determine whether title invention occurs based on the provided definition.
Output only one of the following:
1 – Title invention occurs (the title claims experimental results, but the paper does not report them).
0 – No title invention (the title accurately reflects the paper's content).
Input Format:
The title will be enclosed in <title></title>.
The paper will be enclosed in <paper></paper>.
Output only 1 or 0, with no additional text or explanation.
```

Box 6. LLM Prompt for Title Invention

For publication type and topic classification (Key Task 1) as well as stance detection (Key Task 2), we evaluated performance using precision, recall, and F1-score for each class, and overall accuracy. For citation accuracy classification and scientific claim verification, Key Tasks 3 and 4, we prioritized accuracy and recall given the low number of positive cases.

RESULTS

Key Task 1: Publication Type and Topic Classification

Publication Type	Count	Non-LLM Approach			LLM Approach		
		P	R	F1	P	R	F1
Myositis Review	63	0.75	0.92	0.83	0.85	0.79	0.82
Animal/Cell Culture Model	17	0.35	0.94	0.51	0.38	1.00	0.55
Primary Data	12	0.35	0.58	0.44	0.21	0.58	0.30
Other	150	0.91	0.60	0.72	0.88	0.61	0.72

Table 7. Publication Type Classification Results

Table 7 presents the precision, recall, and F1-scores for both the non-LLM-based and LLM approaches in publication type and topic classification. Overall, the two methods performed similarly, with the non-LLM approach achieving 71% accuracy and the LLM approach slightly lower at 69%. Performance on the myositis review paper categorization was comparable across both methods in terms of F1-score; however, the non-LLM approach, metadata-driven and independent of NLP, demonstrated higher recall, while the LLM approach showed higher precision. For animal/cell culture model papers, the non-LLM approach underperformed the LLM method across all three metrics. The non-LLM approach missed one animal/cell culture model paper, while the LLM correctly identified all papers in this category. Despite this, the low precision in both approaches indicates that several papers were incorrectly classified as animal/cell culture model papers.

For primary data papers, both methods correctly identified 7 out of 12 papers, but the non-LLM approach had higher precision. The non-LLM approach identified 19 primary data papers in total, all labeled as supportive. Of these, 6 were correctly identified as supportive primary data papers, while a single critical primary data paper was misclassified as supportive. Similarly, the LLM approach identified more supportive than critical primary data papers (32 vs. 2). All 7 of the correctly identified primary data papers by the LLM approach were supportive.

Key Task 2: Stance Detection

Method	Supportive (n=631)			Neutral (n=18)			Critical (n=20)		
	P	R	F1	P	R	F1	P	R	F1
Stance Detection	0.93	0.41	0.57	0.02	0.33	0.04	0.03	0.15	0.06
Sentiment Classification	0.96	0.04	0.08	0.03	1.00	0.05	0.00	0.00	0.00
LLM	0.96	0.95	0.96	0.03	0.06	0.04	0.75	0.45	0.56

Table 8. Stance Detection Results

Stance detection results, including precision, recall, and F1-scores, are summarized in Table 8. The LLM approach had the best performance. The overall accuracies for the non-LLM stance detection, non-LLM citation sentiment classification, and LLM stance detection were 40%, 7%, and 90%, respectively. The citation sentiment classification approach performed poorly, failing to predict any critical stance instances. We hypothesize that this is because the stance detection models were given claim information, whereas the citation sentiment classification model only was provided with the citation context. To test this, we evaluated all citation contexts against two nearly identical LLM prompts for citation sentiment classification. Providing both the claim and citation context (89% accuracy) significantly outperformed providing only the citation context (13% accuracy).

Key Task 3: Citation Accuracy Classification

Citation Invention	Cases	Non-LLM Approach			LLM Approach		
		Accuracy	Recall	Cases Detected	Accuracy	Recall	Cases Detected
Citation Diversion	4	0.71	0.25	230	0.75	0.50	200
Citation Transmutation	17	NA	NA	NA	0.41	0.65	485
Dead End Citation	9	0.67	0.44	268	0.80	0.56	164

Table 9. Citation Accuracy Classification Results

Table 9 shows the accuracy, recall, and total number of detected cases for each type of citation invention. Both non-LLM and LLM approaches made a substantial number of overpredictions.

Citation Diversion: The LLM approach slightly outperformed the non-LLM approach, achieving 4% higher accuracy, correctly identifying one additional case, and detecting 31 fewer incorrect cases. The one true case identified by the non-LLM approach did not overlap with the two identified by the LLM. While both methods labeled at least 200 citation contexts as citation diversion, only 51 citation contexts were common to both approaches. Overall, the non-LLM and LLM approaches agreed on the classification labels for 480 out of the 808 total citation contexts. Additionally, the non-LLM approach classified 9 citation transmutation cases and a single dead end citation case as citation diversion, while the LLM classified 5 citation transmutation and 2 dead end citation cases as citation diversion.

Citation Transmutation: The LLM approach correctly identified 11 of the 17 true citation transmutation cases. However, it predicted a total of 485 citation contexts as citation transmutation, which includes all 4 citation diversion cases and 7 out of 9 dead end citation cases.

Dead end citation: The LLM approach had better performance than the non-LLM approach, obtaining higher accuracy (+13%), identifying more true dead end citation cases (5 vs. 4), and detecting substantially fewer cases (164 vs. 268). The non-LLM approach labeled 1 citation diversion and 4 citation transmutation cases as dead end citation, compared to 1 and 8, respectively, for the LLM. Overall, the LLM and non-LLM methods assigned the same labels for 470 citation contexts, including 47 instances where both predicted dead end citation cases. However, they agreed on only one true dead end citation case.

Key Task 4: Scientific Claim Verification

Title Invention: Neither approach was able to detect the lone case of title invention. The non-LLM method misclassified 46 titles, while the LLM approach misclassified 12, with both agreeing on 6 instances.

DISCUSSION

Our study investigated whether NLP tools can help identify citation distortions in scientific literature. Specifically, we sought to replicate Greenberg's results using NLP techniques. We evaluated both LLM-based and non-LLM approaches, finding that the LLM approach outperformed the non-LLM approach in almost all tasks. In contrast, identifying review or model papers may only require metadata such as publication types, MeSH terms, or abstracts, as the use of LLMs did not yield significant improvements over these metadata.

An overarching pattern across all tasks is poor performance when the information needed includes a paper's full text, whether of the paper itself (as in identifying primary data papers and title invention issues) or of the cited article (as in detecting citation diversion, citation transmutation, and dead end citation). Stance detection, which showed the most promising results out of all tasks, with LLM accuracy reaching as high as 90%, used the shortest inputs (roughly 2 sentences), requiring only the claim and citation context.

Citation Bias

Greenberg (2009) observed that supportive primary data papers were cited far more frequently than critical papers, with supporting papers receiving 94% of the citations. Although both approaches performed poorly (<0.5 F1-score) in identifying primary data papers, our results still lead to the conclusion that citation bias is present because both methods overpredicted the number of supportive primary data papers. The non-LLM method identified 19 supportive papers (no critical primary data papers), while the LLM approach identified 32 supportive and 2 critical papers. Based on the LLM predictions, supportive papers received 311 citations, while critical papers received only 1, accounting for 99.7% of all citations to primary data paper, aligning with Greenberg's original finding, although our conclusion is based on evidence that is not entirely accurate.

Amplification

Greenberg (2009) found that 95% of citation paths pass through four review papers. Our methods successfully identified all four of these key review papers, allowing us to replicate this aspect of Greenberg's findings. This suggests that our approach is effective in helping capture amplification due to information propagation from secondary sources. Furthermore, our LLM-based stance detection also performed well, enabling us to replicate Greenberg's observation that supportive statements receive more citations. Together, our results suggest that amplification-related citation distortions can be effectively captured using current tools, particularly LLMs.

Invention

Both approaches showed limited effectiveness in detecting citation invention, which presents a greater challenge than citation bias or amplification, as it requires deeper contextual understanding and domain knowledge. We observed substantial overlap among different citation invention issues, suggesting either a need to craft prompts with clearer distinctions or that the differences between the issues are too nuanced for current models to disambiguate.

We conducted a more detailed analysis of the LLM approach, focusing on potential improvements to prompt design. Specifically, we tested two alternative prompts for identifying citation diversion cases: one employing a multi-step reasoning approach that decomposed the task into four sequential steps, and another that required the LLM to justify its decision. Neither approach yielded substantial improvements over the original prompt: both the multi-step reasoning and justification-based prompting achieved 76% accuracy, compared to the original 75% accuracy. The original prompt correctly identified two true positives but generated 200 total predictions. The multi-step reasoning obtained the same number of true positives with fewer predictions (194), while the justification-based prompt identified only one true case, with 193 total predictions. These results suggest that more structured prompting alone may be insufficient to improve performance on citation invention tasks without incorporating example contexts.

Practical Implications

An effective and scalable way to identify citation distortions, particularly through widely accessible LLMs, has the potential to mitigate citation distortion issues. Integrating such tools into the scholarly publication system could help prevent citation issues from making their way into scientific literature. However, our current results raise concerns about the reliability of these tools for detection of citation distortions, given their tendency to overpredict distortions.

LIMITATIONS AND FUTURE WORK

Further investigation is needed into other factors that may have contributed to poor performance on tasks requiring full text input, including input quality. PDF-to-text conversion tools are not perfect (Meuschke et al., 2023). However, input quality alone is not sufficient to explain the low agreement between the non-LLM and LLM approaches regarding detection of citation invention. Another reason may be that using full papers as inputs make it challenging to pinpoint the relevant context needed to identify citation distortions. A more focused strategy, such as retrieval-augmented generation (Lewis et al., 2020), could improve performance by providing only the most relevant context as input.

It is worth noting that neither LLM nor the non-LLM tools we adapted were developed for the tasks in this study. Tools specifically developed for identifying citation distortions might demonstrate better effectiveness. In the future, we plan to explore alternative approaches to automate citation distortion analysis and develop a more generalized pipeline, as the current replication study remains closely tied to the specifics of Greenberg's work. A fully automated pipeline would start by identifying all papers relevant to a claim and extracting the citation contexts. Both tasks are non-trivial, and we did not attempt these steps in this work. Rather, we built upon the manually curated data provided by Greenberg's (2009). While methods for stance detection, citation accuracy classification, and scientific claim verification are likely transferable, the publication type and topic classification task is more context-dependent. In our case, MeSH terms closely aligned with the defined publication types and topics, which simplified the classification process. For future studies, tools like MultiTagger (Cohen et al., 2021; Menke et al., 2024), which automatically tags papers with their publication types and study designs, could be considered.

Finally, the claim-related citation network examined in Greenberg's study (2009) includes papers published only through 2007; this network has likely grown since then. While our goal was to assess whether NLP tools can identify citation distortions, a direct application of this work would be to extend Greenberg's analysis to more recently published literature using automated methods.

CONCLUSION

In this study, we examined the utility of NLP in identifying poor citation practices by attempting to replicate Greenberg's 2009 analysis of citation distortions in Alzheimer's research. We explored two approaches: one leveraging LLMs and another relying on existing, publicly available tools and publication metadata. Although full replication of the original study remains challenging, particularly for citation bias and invention, our results show promise for using NLP to detect amplification.

GENERATIVE AI USE

We used ChatGPT for the following purpose: converting BibTeX formatted citations into APA format. We evaluated the output by confirming that the information in the BibTeX file was consistent with the generated output. The authors assume all responsibility for the content of this submission.

AUTHOR ATTRIBUTION

MJS: conceptualization, data curation, formal analysis, software, writing – original draft; JS: conceptualization, supervision, writing – review & editing; HK: conceptualization, supervision, writing – review & editing.

ACKNOWLEDGEMENTS

This study was supported by the Office of Research Integrity (ORI) of the US Department of Health and Human Services (HHS) (grant number: ORIIR220073). The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement, by ORI/OASH/HHS, or the US Government.

JS: 2024–2025 Perrin Moorhead Grayson and Bruns Grayson Fellow, Harvard Radcliffe Institute for Advanced Study; Alfred P. Sloan Foundation G-2022-19409 Reducing the Inadvertent Spread of Retracted Science II: Research and Development towards the Communication of Retractions, Removals, and Expressions of Concern; NSF 2046454 CAREER: Using network analysis to assess confidence in research synthesis.

REFERENCES

Ahn, E., & Kang, H. (2018). Introduction to systematic review and meta-analysis. *Korean Journal of Anesthesiology*, *71*(2), 103–112. https://doi.org/10.4097/kjae.2018.71.2.103

Cohen, A. M., Schneider, J., Fu, Y., McDonagh, M. S., Das, P., Holt, A. W., & Smalheiser, N. R. (2021). Fifty ways to tag your pubtypes: Multi-Tagger, a set of probabilistic publication type and study design taggers to support biomedical indexing and evidence-based medicine [Preprint]. https://doi.org/10.1101/2021.07.13.21260468

Greenberg S. A. (2009). How citation distortions create unfounded authority: analysis of a citation network. *BMJ (Clinical Research Ed.)*, *339*, b2680. https://doi.org/10.1136/bmj.b2680

Jahan, I., Laskar, M. T. R., Peng, C., & Huang, J. X. (2024). A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine, 171*, 108189. https://doi.org/10.1016/j.compbiomed.2024.108189

Jergas, H., & Baethge, C. (2015). Quotation accuracy in medical journal articles—a systematic review and meta-analysis. *PeerJ*, *3*, e1364. https://doi.org/10.7717/peerj.1364

Kilicoglu, H., Peng, Z., Tafreshi, S., Tran, T., Rosemblat, G., & Schneider, J. (2019). Confirm or refute? A comparative study on citation sentiment classification in clinical research publications. *Journal of Biomedical Informatics, 91*, 103123. https://doi.org/10.1016/j.jbi.2019.103123

Krishna, A., Riedel, S., & Vlachos, A. (2022). ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics, 10*, 1013–1030. https://doi.org/10.1162/tacl_a_00503

Leung, P. T. M., Macdonald, E. M., Stanbrook, M. B., Dhalla, I. A., & Juurlink, D. N. (2017). A 1980 letter on the risk of opioid addiction. *The New England Journal of Medicine*, *376*(22), 2194–2195. https://doi.org/10.1056/NEJMc1700150

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (NeurIPS 2020, pp. 9459–9474). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Menke, J. D., Kilicoglu, H., & Smalheiser, N. R. (2024). Publication type tagging using Transformer models and multi-label classification. *AMIA Annual Symposium Proceedings. 2024*, 818–827.

Meuschke, N., Jagdale, A., Spinde, T., Mitrović, J., & Gipp, B. (2023). A benchmark of PDF information extraction tools using a multi-task and multi-domain evaluation framework for academic documents. In *Information for a better world: Normality, virtuality, physicality, inclusivity. iConference 2023* (Lecture Notes in Computer Science, Vol. 13972, pp. 383-405). https://doi.org/10.1007/978-3-031-28032-0_31

Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., & Cherry, C. (2016). A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3945–3952). https://aclanthology.org/L16-1623/

Ngatuvai, M., Autrey, C., McKenny, M., & Elkbuli, A. (2021). Significance and implications of accurate and proper citations in clinical research studies. *Annals of Medicine and Surgery*, *72*, 102841. https://doi.org/10.1016/j.amsu.2021.102841

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 technical report* [Preprint]. https://arxiv.org/abs/2303.08774

Porter, J., & Jick, H. (1980). Addiction rare in patients treated with narcotics. *The New England Journal of Medicine*, *302*(2), 123. https://doi.org/10.1056/nejm198001103020221

Ravnskov U. (1992). Cholesterol lowering trials in coronary heart disease: frequency of citation and outcome. *BMJ (Clinical Research Ed.), 305*(6844), 15–19. https://doi.org/10.1136/bmj.305.6844.15

Riedel, B., Augenstein, I., Spithourakis, G. P., & Riedel, S. (2017). A simple but tough-to-beat baseline for the Fake News Challenge stance detection task [Preprint]. http://arxiv.org/abs/1707.03264

Sarol, M. J., Ming, S., Radhakrishna, S., Schneider, J., & Kilicoglu, H. (2024). Assessing citation integrity in biomedical publications: Corpus annotation and NLP models. *Bioinformatics*, 40(7), btae420. https://doi.org/10.1093/bioinformatics/btae420

Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1074

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 7534–7550). https://doi.org/10.18653/v1/2020.emnlp-main.609

Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., & Hajishirzi, H. (2022). MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 61–76). https://doi.org/10.18653/v1/2022.findings-naacl.6

Zhuge, H. (2006). Discovery of knowledge flow in science. *Communications of the ACM, 49*(5), 101–107. https://doi.org/10.1145/1125944.1125948