https://doi.org/10.51408/issi2025_062

Investigating Information Propagation in Biomedical Literature through Citations: A Case Study

M. Janina Sarol[1], Halil Kilicoglu[2]

[1]*mjsarol@illinois.edu*
Informatics Programs, University of Illinois Urbana-Champaign, Champaign, Illinois (USA)

[2]*halil@illinois.edu*
School of Information Sciences, University of Illinois Urbana-Champaign, Champaign, Illinois (USA)

Abstract

Scientific growth is iterative, with existing knowledge serving as the foundation for new discoveries. Citations serve as the primary channel for information propagation in science, shaping which ideas and findings persist in the literature and which do not. While natural language processing (NLP) is increasingly used in citation context analysis, it is underutilized in studies that examine the actual scientific content of citations. In this pilot study, we explored how NLP can be used to track the propagation of scientific findings by replicating a prior citation context study that relied on manual extraction. We compared two approaches: a traditional NLP pipeline (named entity recognition and relation extraction) and a generative large language model (LLM). We formulated a two-step automated pipeline: (1) extracting findings from a reference paper and (2) mapping citation contexts to the findings they reference. Our findings indicate that LLMs are superior to traditional NLP techniques in both steps of the pipeline. However, they are also more prone to errors, mapping citation contexts to findings they do not reference. While the two-step automated pipeline was effective, integrating manual annotation of findings with LLM-based mapping of citation contexts yields the best results. To our knowledge, this study is one of the first to explore how NLP, particularly LLMs, can be leveraged to track the flow of information in science. Future research should further evaluate the application of LLMs and other NLP techniques on a larger scale to assess their effectiveness in supporting citation-focused scientometric and informetric studies.

Introduction

Scientific progress is fundamentally cumulative, with each new discovery advancing upon prior knowledge. Citations serve as the primary means by which previous work is acknowledged, disseminated, and built upon (Cronin, 1984). They are the channels through which scientific information flows. As a result, analyzing citations can provide us with valuable insights into various aspects of science – the dynamics of scientific progress (Yang & Deng, 2024), influential research endeavors (Herrmannova et al., 2018), emerging trends (Schneider & Costas, 2017), and even gaps in current research (Farooq, 2017). It is therefore unsurprising that citations represent a core unit of analysis in science of science subfields such as bibliometrics, scientometrics, and informetrics.

Analyzing the scientific content within citation contexts could allow us to observe which ideas and findings continue to shape the literature, identify the most impactful discoveries within research domains, follow the emergence of new ideas, and track when and where scientific claims become generally accepted as facts. From an acknowledgment perspective, we can trace when and where scientific claims originated and ensure proper recognition. Finally, it is important to note that not all citations are accurate (Jergas & Baethge, 2015), and unfounded information can make its way into the scientific record (Greenberg, 2009). With timely analysis of citation content, we could uncover and prevent misinformation from spreading in the scientific community.

A number of studies have explored tracing the information propagated from reference to citing articles. In early work, Cozzens (1985) investigated how citations to knowledge claims differed across two papers in Neuropharmacology and Sociology of Science. Most citations to the Neuropharmacology paper were about its methodology and findings, whereas a smaller percentage of citations to the Sociology of Science paper focused on its claims. Anderson and Lemken (2019) reviewed 1,400 citations to *Organizations*, a highly influential publication in management, and classified them into 7 thematic categories (classical organization theory, motivation and decisions, participation, conflicts, cognitive limits, routines and programs, and planning). Leng (2022) examined 343 papers citing a study about coronary heart disease and found that research communities tend to cite the findings most related to their communities.

While some citation context studies focus on the topical content of citations, they are often overshadowed by other higher-level analyses such as sentiment analysis and citation function classification. This is largely because these analyses are inherently easier to conduct, relying on classifying citations into a predefined set of categories that are applicable to all papers. For instance, for citation sentiment analysis, the goal is to classify citations into positive, negative, and neutral (Yu, 2013), while citation function classification categorizes citations based on their rhetorical purpose in the citing paper (Teufel et al., 2006). The relative ease of annotating data for these types of analyses has also contributed to the increasing availability of tools, particularly those leveraging natural language processing (NLP) techniques, which in turn makes it easier to conduct these types of studies at scale.

In contrast, (scientific) citation content varies from one paper to another, based on the source being cited, making it more challenging to develop generalizable NLP approaches tailored for this task. This is why content-focused citation studies typically involve manual analysis, which limits the number of citing publications researchers can feasibly examine. Coupled with the rapid growth of scientific literature and its citations, conducting generalizable content-focused citation studies is increasingly difficult. For example, as of February 2025, *Organizations* has accumulated over 40,000 citations according to Google Scholar. Extending the study by Anderson and Lemken (2019) to cover all citations would be a daunting task.

Although there is a lack of NLP-based approaches specifically developed for extracting and analyzing the scientific content of citations, various other NLP techniques may be useful for this task. Information extraction methods, in particular, can assist in automatically retrieving the scientific content of citations. By applying well-established tasks such as named entity recognition (NER) and relation extraction (RE), we can identify scientific concepts and their relationships mentioned in a reference paper and determine whether this information is cited by subsequent publications. For instance, Leng (2022) analyzed a paper by Paul et al. (1963), which explored various factors associated with coronary heart disease. Leng (2022) identified 34 distinct findings, noting uneven citation distributions across these findings, with research communities typically citing the findings most relevant to their fields. It is possible to apply NER to determine the factors referenced in each citation, while RE could pinpoint which of the 34 findings were cited.

In this pilot study, we explore the potential of current NLP tools in tracking the flow of scientific information through citations. To showcase a real-world application, we aim to replicate the Leng (2022) citation context study. Our key research question is, "*How can we utilize NLP methods to effectively and efficiently track the propagation of information through citations?*" If full automation is not yet feasible, we assess which steps can be automated and which ones still require human intervention. We approach this by testing two methodologies: one that uses established NLP methods in NER and RE, and another that applies generative large language models (LLMs), which have recently attracted significant attention as a promising tool (Google DeepMind, 2024). Our study shows that NLP techniques, particularly LLMs, could help in understanding the flow of scientific information at scale while also suggesting that problems such as hallucinations need to be addressed to do this reliably.

Related Work

Tracking the propagation of information through citations is a well-explored research area, but it has primarily been approached from a network analysis perspective. For instance, della Briotta Parolo (2020) examined forward chains of citations to measure persistent influence, which describes how a paper impacts subsequent works in its citation chain, finding that publications linked to Nobel Prize winners have higher persistent influence. In contrast, Min et al. (2021) focused on the backward chain of citations, or references of references, to map the knowledge ancestry of papers. While these studies provide valuable insights, they overlook the actual content of citations, treating each citation as equally informative and important to the citing paper.

Automatically linking the citing text with the corresponding statements from reference articles has been explored, primarily for the task of scientific document summarization (Jaidka et al., 2016). Ou and Kim (2019) proposed similarity- and ranking-based methods for this task and suggested their use in conducting citation analysis studies. More recently, Sarol et al. (2024) connected citing texts with reference article statements to assess the accuracy of citations.

Methods

In this section, we give a thorough overview of the Leng (2022) citation context study, detail the specifics of our replication efforts, and describe the NLP solutions we evaluate.

The Leng Citation Context Study

Leng (2022) examined 343 publications that cited Paul et al. (1963), hereafter referred to as the *original study*, a prospective cohort study that examined several factors linked to coronary heart disease (CHD). Leng (2022) identified 34 different findings from the original study. The citation contexts from each of these publications were manually extracted and classified based on the finding they referenced. 304 papers cited at least one finding, while 38 merely mentioned the original study without discussing any of its findings. One paper was found to cite incorrect information that did not appear in the original study. With its focus on citation context analysis to investigate how information from a single paper was propagated, Leng (2022) provides a strong foundation for our pilot study.

We categorized the findings discussed in Paul et al. (1963) into four sets of categories:

1) Association Relations

The original study found 15 factors associated with CHD: *cholesterol*, *blood pressure*, *coffee*, *smoking*, *body fatness*, *electrocardiogram findings* (particularly ST-segment or T-wave abnormalities), *somatotype* (primarily endomorphic dominance), *heart rate*, *chest discomfort*, *peptic ulcer*, *age*, *early death of father*, *chronic cough*, *shortness of breath*, and *arteriovenous nicking* (Paul et al., 1963). Although *diet* was not directly linked to *CHD*, a positive association was found between *diet* and *cholesterol levels*. There were 302 references to these association findings, with 195 papers citing at least one of them, representing over half (56.85%) of the citing papers. The association between *arteriovenous nicking* and *CHD*, however, was never cited.

2) Lack of Association

Paul et al. (1963) discovered 12 factors – *diet*, *alcohol*, *physical activity*, *body weight*, *job role*, *blood glucose*, *height*, *hemoglobin*, *gallbladder disease*, *lipoprotein lipase*, *non-paternal family history*, and *arcus senilis* – that appeared unrelated to CHD. These non-association findings were cited 124 times across 110 citing papers (32.07%). The non-associations between *CHD/family history* and *CHD/arcus senilis* received no citations.

3) Comparison

Paul et al. (1963) noted differences in dietary information based on the collection method. Dietary information collected using food diaries showed lower food intake than data from participant interviews. This finding was cited 7 times. 5 citing papers also compared the dietary intake between the original study participants and other population groups.

4) Other Findings

13 citing papers discussed the general incidence of *CHD* in the original study, without specifying its association to the factors. The seasonal fluctuations in serum cholesterol, seasonal fluctuations in blood pressure, and participation rate in the original study were cited by 6, 2, and 5 papers, respectively.

The citation counts of each finding are shown in Appendix Table 1. 231 papers cited a single finding, while 73 papers cited two or more findings. The most cited findings are the associations between *CHD* and *cholesterol* (85 citing papers), *blood pressure*

(57), and *coffee* (54). Additional analysis of the categorized citation contexts was conducted to determine which findings were cited together and how findings varied over time.

Finally, Leng (2022) constructed a citation network among the citing papers and partitioned the network into nine clusters, each representing a research community as inferred from the papers' titles. The network analysis revealed that (1) the distribution of findings highly varied, with no single finding being referenced by more than 25% of the papers, and (2) research communities primarily cited findings that aligned with their own research interests.

Replication

As the study focuses on the utility of NLP, our main goal is to automatically replicate the manual process of linking the content of each citation in a citing paper to the findings in the reference article. Specifically, given that the original study aimed to identify factors associated with *CHD*, we seek to identify citation contexts that referenced the association and lack of association findings. This replication will allow us to assess the feasibility of conducting such studies on a larger scale.

A total of 268 papers referenced at least one of these two groups of findings. We used the citation contexts extracted by Leng (2022), available in the supplementary material of this citation context study. Our automated process is as follows: we begin by extracting the findings from the original study, then classify the citing papers in accordance with those findings. This process simulates a scenario where a researcher fully relies on NLP, eliminating the need to manually read and extract findings from the reference paper.

Natural Language Processing Methods

We examined two methods: one that uses a combination of NER and RE, and another that solely relies on a large language model.

1) Named Entity Recognition and Relation Extraction

NER followed by RE is a common approach to extract knowledge from scientific literature in the form of concepts and their relationships, respectively. scispaCy is a Python library designed for processing biomedical and scientific texts (Neumann et al., 2019). It offers tools for biomedical NER, which is the NLP task of extracting biomedical concepts (entities) from unstructured text. Additionally, scispaCy supports entity linking, which normalizes different mentions of the same concept to standard identifiers in knowledge bases (French & McInnes, 2023). We mapped the concept mentions to their identifiers in the Unified Medical Language System (UMLS) (Bodenreider, 2004). For instance, the mentions *clinical coronary disease* and *coronary disease* both map to the same UMLS concept *coronary heart disease* (concept unique identifier: C0010068).

RE involves identifying related concepts based on the text and the nature of their relations. We performed relation extraction using the BERT-based model developed by Sarol et al. (2024) to identify associations and non-associations. This model was trained on the BioRED corpus (Luo et al., 2022) and extracts eight relation types:

association, positive correlation, negative correlation, binding, drug interaction, cotreatment, comparison, and conversion between six types of entities: diseases, chemicals, species, genes/proteins, mutations, and cell lines. Since the original study focused on associations, such as the link between *elevated cholesterol levels* and *CHD*, we broadened our definition of association to include both association and positive correlation predictions from the model. To determine lack of association, if scispaCy identified a pair of concepts (e.g., *CHD* and *diet*) in a citation context but the model did not detect a relation, we classified this as a lack of association.

2) Large Language Model

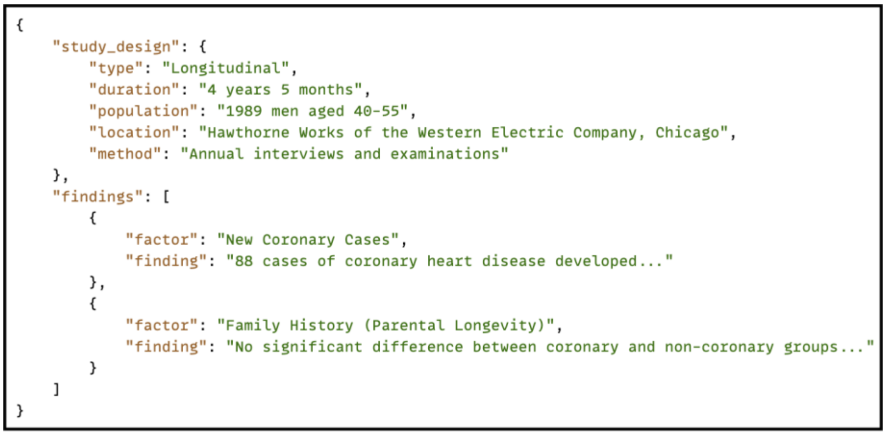
The NER + RE approach above is limited to some extent, as it can only consider the entity types included in UMLS, which, while extensive, is not exhaustive, and can only identify relations similar to those expressed in the BioRED corpus. Large language models have been shown to be capable of handling tasks they were not specifically trained on (Yang et al., 2024), making them a promising approach for this study. We designed two prompts: one to extract the findings from the original study and another to determine which findings were referenced in each citation context. In the first prompt, rather than instructing the LLM to identify concepts and their relations, we directly prompted the LLM to extract the original study's findings. The second prompt was applied individually to each citation context. We used Google Gemini 1.5 Pro as the LLM for this study, as it has demonstrated strong performance on long context documents (Google DeepMind, 2024), which makes it appropriate for processing scientific articles.

Table 1 shows the prompt used for the first step, with the input text truncated for readability. The input text contains the full text of the original study.

Table 1. Prompt for Identifying Findings in the Original Study.

Instruction	*The text below is a research publication. Please extract and summarize all the findings of this paper and present them in a structured JSON format. Ensure that each finding is concise, clearly worded, and reflects the main conclusions of the study.*
Input Text	*A Longitudinal Study of Coronary Heart Disease* *SINCE the Fall of 1957, a long-term study of coronary heart disease has been in progress at the Hawthorne Works of the Western Electric Company in Chicago under the auspices of the University of Illinois College of Medicine and Presbyterian-St. Luke's Hospital. The study was undertaken in the belief that coronary heart disease was a disease resulting from the interplay of multiple factors and that there was need to delineate these factors further…*

Figure 1 illustrates the JSON format of the output produced by the given prompt.



```
{
    "study_design": {
        "type": "Longitudinal",
        "duration": "4 years 5 months",
        "population": "1989 men aged 40-55",
        "location": "Hawthorne Works of the Western Electric Company, Chicago",
        "method": "Annual interviews and examinations"
    },
    "findings": [
        {
            "factor": "New Coronary Cases",
            "finding": "88 cases of coronary heart disease developed..."
        },
        {
            "factor": "Family History (Parental Longevity)",
            "finding": "No significant difference between coronary and non-coronary groups..."
        }
    ]
}
```

Figure 1. Output of the Prompt for Identifying Findings in the Original Study.

We constructed the second prompt based on the output of the first prompt. Table 2 shows an example prompt, which contains the instruction, the JSON-formatted list of findings obtained from the first prompt, and the citation context.

Table 2. Example Prompt for Identifying Cited Findings.

Instruction	*The JSON text below lists the findings of a reference paper. Each finding is described in the 'finding' field, with its shorthand provided in the 'factor' field. The text enclosed in $CITATION$ contains a citation to this reference paper.* *Identify which findings from the JSON are referenced in the citation text. A finding is considered cited if the information it conveys is consistent with the text in the 'finding' field.* *Output only the 'factor' values of the relevant findings as a comma-separated list. If no findings are cited, return an empty string.*
List of Findings	*[* *{* *"factor": "New Coronary Cases",* *"finding": "88 cases of coronary heart disease developed..."* *},* *{* *"factor": "Family History (Parental Longevity)",* *"finding": "No significant difference between coronary and non-coronary groups..."* *},* *...* *]*
Input Text	*$CITATION$* *A relationship between the serum cholesterol level and the relative risk of developing clinical coronary heart disease has been reported by many investigators [4, 6, 9, 11, 12, 14-16]". (p. 358)* *$CITATION$*

Evaluation

Recall served as our main evaluation metric for this study, as our goal was to determine if NLP could capture the same data as the manual approach. For each finding, we calculated the proportion of citing publications correctly identified by the NLP methods. Precision was less suitable, particularly in the NER + RE approach, since it may extract valid biomedical concepts that were not part of the cited findings.

Results

Table 3 presents a comparison of the results of the two NLP methods. Overall, the LLM-based pipeline outperformed the traditional NLP approach. It identified 80% of the total citations to the findings, while the NER+RE approach only succeeded in mapping 23%. Further, out of the 268 citation contexts, the LLM correctly found all cited findings in 184 (69%) citation contexts compared to just 47 (18%) for the NER+RE method. The LLM successfully extracted 26 out of 28 findings from the original study (93%), whereas the NER+RE approach managed to retrieve only 16 (54%). We also examined the scenario in which findings are manually extracted (data was collected from the Leng (2022) study's supplementary material), automating only the process of mapping each citation context to the corresponding finding. The NER+RE approach had similar performance to the full 2-step pipeline, but the LLM method yielded better results when given manually annotated findings. A detailed list of recall results for each finding is available in Appendix Table 2.

Table 3. Summary of Results.

Task	*NLP Method*	*Recall*
Full Pipeline	NER + RE	23%
	LLM	80%
Step 1 Only: Extracting Findings from Original Study	NER + RE	57%
	LLM	93%
Step 2 Only: Mapping Findings to Citation Contexts	NER + RE	23%
	LLM	86%

Named Entity Recognition and Relation Extraction

Out of the 28 key concepts that scispaCy was tasked with identifying from the original study – *CHD* and the 27 studied factors – only one factor, *early death of father*, was not recognized. This factor does not correspond to a single UMLS concept. We found that despite the entity linking capabilities of scispaCy, manual entity linking was still necessary, as concepts in the original study could further be mapped to multiple UMLS concepts. For instance, the term *coronary heart disease* was used loosely in the original study, covering related concepts such as *angina pectoris* and *myocardial infarction*. Thus, we had to add these UMLS concepts to ensure that references to *CHD* were properly covered. The complete mapping is provided in Appendix Table 1. We used the UMLS identifiers to identify the concepts mentioned in the citation contexts.

The BERT-based model successfully extracted 4 of the 16 association relations from the original study, correctly linking *CHD* to *cholesterol*, *blood pressure*, *smoking*, and *coffee*, coincidentally the four most cited findings. However, it erroneously detected an association between *CHD* and *hemoglobin*. The model also incorrectly mapped 8 findings to citation contexts that did not mention them – for example, the association between *CHD* and *height* was linked to a citation context that discussed the relationship between *CHD* and *cholesterol*.

Large Language Model

The LLM found 28 total findings (shown in Appendix Table 3). The lack of association between *CHD* and both *height* and *weight* were combined into a single finding. One of the findings is about the general incidence of *CHD* and another about the relation between *CHD* and *perceived tension*, which was not on the list of findings extracted by Leng (2022). However, it is indeed reported in the original study. All findings extracted by the LLM are accurate; the LLM did not hallucinate any findings. The LLM identified 14 association and all 12 lack of association findings. It failed to identify the associations between *diet* and *cholesterol*, and *CHD* and *age*. The LLM incorrectly attributed 46 incorrect findings to citation contexts that did not discuss them.

Discussion

In this pilot study, we attempted to replicate a study that focused on citation context analysis to understand how information is propagated from one article to others using automated methods. Our results show that LLMs are superior to more traditional information extraction methods in linking findings from a reference article to their citations.

The Need for Human Intervention

We note that in both methods, we still needed human intervention to complete the tasks. For the NER+RE pipeline, we needed to map *CHD* and each of the 27 factors to UMLS concepts, and this mapping was also limited to the UMLS concepts extracted from the original study. As a result, any concept that was not extracted from the original study, even if correctly identified in the citation contexts, was not included in the mapping. We found several UMLS concepts in the citation contexts that were consistent with those in the original study but were not extracted by scispaCy from the original study. For example, *cholesterol* and *alcohol*, both of which have 5% recall, had more than half of citation contexts containing mappings to *Serum cholesterol measurement* (C0587184) and *Alcohol consumption* (C0001948), respectively. Including these terms in the list of allowed UMLS concepts would raise their recall values to greater than 50%. Not only is there a need to manually map related UMLS concepts within the original study, but there is also a need to review the UMLS concepts extracted from the citation contexts, which may be an infeasible task to perform at scale. In contrast, the LLM only required minor human intervention, primarily for identifying the JSON format produced by the first prompt.

While the two-step automated pipeline using an LLM was shown to be effective, using the manually annotated findings with LLM-based mapping of citation contexts yielded the best results. Of particular note was the boost in results related to *diet* when the manual annotations were used instead of the automatically extracted findings: recall of citation contexts citing findings on the association of *diet* and *cholesterol* and the lack of association of *CHD* and *diet* increased from 0% to 71% and 49% to 73%, respectively. This suggests that the most effective process still requires human intervention.

Human vs Machine Annotation

While the LLM yielded better recall performance, it also made more errors. We manually examined the 46 erroneous citation context-finding mappings and found that 8 were consistent with the citation context (i.e., they were manual annotation errors), 12 resulted from the extra text in the citation context (citation contexts were on a sentence level), and the remaining 26 were incorrect mappings by the LLM. Examples of each case are shown in Table 4. For the 8 incorrect manual annotations, 5 involved confusion between *job role* and *physical activity*. In the original study, *job role* referred to physical activity on the job, while *physical activity* referred to physical activity off the job.

Our analysis demonstrated that manual annotations are not consistently accurate, indicating the potential value of a hybrid approach that integrates both human and machine annotations. LLMs can either serve to supplement and double-check manual annotations or be regarded as independent annotators. However, it remains essential that humans perform the final verification and thoroughly review all annotation outputs.

Table 4. Examples of Erroneous Citation Context-Finding Mappings by the LLM.

Case	*Citation Context*	*CHD-Associated Factor*
Incorrect Manual Annotation	*However, studies are not entirely consistent, and a number of US long-term studies of initially healthy men have failed to show a relationship between incidence of ischemic heart disease and occupational activity [25-28]*	Manual: *physical activity* LLM: *job role* (no association)
Extra Text from the Citation Context	*Cigarette smoking is well established as a CHD-risk factor [17, 18], and caffeine intake has been incriminated recently [**19**]*	Manual: *coffee* LLM: *smoking*
Incorrect LLM Annotation	*Paul et al. [**30**] demonstrated a significant correlation between coffee consumption and the later development of coronary disease, although serum cholesterol levels were normal.*	Manual: *coffee* LLM: *cholesterol*

A Fully Automated Process

While our study aimed to replicate the manual mapping of citation contexts to referenced findings, arguably the most time-consuming part of the study, we skipped some necessary steps for full automation. A full end-to-end pipeline would include automated collection of the full texts of the original study and the citing papers, as well as the citation contexts pertaining to the original study. Both steps are non-trivial. We initially tried collecting the list of citing papers automatically but failed to find most papers. We resorted to manually collecting the PDFs of citing papers and found that conversion from PDF to text is also an issue, especially since the citing papers are older documents published from 1963-1984. Future work should consider automating an end-to-end pipeline, which would be of most benefit to the scientometrics and informetrics communities.

Replicability to Other Publications

We note that Paul et al. (1963) is a short paper, and its findings are presented as section headers, making it a relatively easy case for a citation context study that tracks the dissemination of information. We considered it for this study, since the Leng (2022) study could be used as a proxy for ground truth. While the approach may yield weaker results on a more complex paper, this study demonstrates the potential for (semi-)automated approaches. Future work could consider the construction of a larger dataset that can be used for evaluation and possibly for training or fine-tuning NLP models, including LLMs.

Conclusion

We examined the potential of NLP in tracking the propagation of scientific findings through citations by replicating a citation context study that relied on manual extraction and assessing the advantages and shortcomings of two approaches: an NER and RE pipeline, and an LLM. LLMs outperformed the traditional NLP methods in both extracting findings from the original study and mapping citation contexts to their referenced findings. Our results suggest that LLMs might be an effective tool for analyzing the propagation of information in science. In the future, we plan to evaluate additional NLP tools and LLMs (including open-weight models) and refine this approach to apply it to other similar citation context studies to better assess its generalizability.

References

Anderson, M. H., & Lemken, R. K. (2019). An empirical assessment of the influence of March and Simon’s Organizations: the realized contribution and unfulfilled promise of a masterpiece. *Journal of Management Studies, 56*(8), 1537–1569. https://doi.org/10.1111/joms.12527

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, *32*(suppl_1), D267-D270. https://doi.org/10.1093/nar/gkh061

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Cozzens, S. E. (1985). Comparing the sciences: citation context analysis of papers from neuropharmacology and the sociology of science. *Social Studies of Science, 15*(1), 127–153. https://doi.org/10.1177/030631285015001005

della Briotta Parolo, P., Kujala, R., Kaski, K., & Kivelä, M. (2020). Tracking the cumulative knowledge spreading in a comprehensive citation network. *Physical Review Research, 2*(1), 013181. https://doi.org/10.1103/PhysRevResearch.2.013181

Farooq, R. (2017). A framework for identifying research gap in social sciences: Evidence from the past. *IUP Journal of Management Research, 16*(4), 66-75.

French, E., & McInnes, B. T. (2023). An overview of biomedical entity linking throughout the years. *Journal of Biomedical Informatics, 137*, 104252. https://doi.org/10.1016/j.jbi.2022.104252

Google DeepMind. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens*. arXiv. https://arxiv.org/abs/2403.05530

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

Herrmannova, D., Patton, R. M., Knoth, P., & Stahl, C. G. (2018). Do citations and readership identify seminal publications?. *Scientometrics*, *115*(1), 239-262. https://doi.org/10.1007/s11192-018-2669-y

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., & Kan, M.-Y. (2016). Overview of the CL-SciSumm 2016 Shared Task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL)* (pp. 93–102).

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

Leng, R. I. (2022). Diversity in citations to a single study: A citation context network analysis of how evidence from a prospective cohort study was cited. *Quantitative Science Studies, 2*(4), 1216–1245. MIT Press. https://doi.org/10.1162/qss_a_00154

Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N., & Lu, Z. (2022). BioRED: A rich biomedical relation extraction dataset. *Briefings in Bioinformatics, 23*(5), bbac282. https://doi.org/10.1093/bib/bbac282

Min, C., Xu, J., Han, T., & Bu, Y. (2021). References of references: How far is the knowledge ancestry. In *Proceedings of the 2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 262-265). IEEE. https://doi.org/10.1109/JCDL52503.2021.00079

Neumann, M., King, D., Beltagy, I., & Ammar, W. (2019). SciSpaCy: Fast and robust models for biomedical natural language processing. *CoRR, abs/1902.07669*. https://doi.org/10.48550/arXiv.1902.07669

Ou, S., & Kim, H. (2019). Identification of citation and cited texts for fine- grained citation content analysis. *Proceedings of the Association for Information Science and Technology, 56*(1), 740-741. https://doi.org/10.1002/pra2.156

Paul, O., Lepper, M. H., Phelan, W. H., Dupertuis, G. W., Macmillan, A., McKean, H., & Park, H. (1963). A longitudinal study of coronary heart disease. *Circulation, 28*(1), 20–31. https://doi.org/10.1161/01.cir.28.1.20

Sarol, M. J., Hong, G., Guerra, E., & Kilicoglu, H. (2024). Integrating deep learning architectures for enhanced biomedical relation extraction: a pipeline approach. *Database, 2024*, baae079. https://doi.org/10.1093/database/baae079

Schneider, J. W., & Costas, R. (2017). Identifying potential “breakthrough” publications using refined citation analyses: Three related explorative approaches. *Journal of the Association for Information Science and Technology*, *68*(3), 709-723. https://doi.org/10.1002/asi.23695

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110). Association for Computational Linguistics. https://aclanthology.org/W06-1613

Yang, A. J., & Deng, S. (2024). Dynamic patterns of the disruptive and consolidating knowledge flows in Nobel-winning scientific breakthroughs. *Quantitative Science Studies*, *5*(4), 1070–1086. https://doi.org/10.1162/qss_a_00323

Yang, J., Jin, H., Tang, R., Han, X., Feng, Q., Jiang, H., Zhong, S., Yin, B., & Hu, X. (2024). Harnessing the power of LLMs in practice: a survey on ChatGPT and beyond. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1-32. https://doi.org/10.1145/3649506

Yu, B. (2013). Automated citation sentiment analysis: What can we learn from biomedical researchers. *Proceedings of the American Society for Information Science and Technology*, *50*(1), 1-9. https://doi.org/10.1002/meet.14505001084

Appendix

Table 1. Named Entity Recognition Results and the Mapping of Concepts to UMLS Identifiers.

Concept	*UMLS Concepts*	*Citations*	*Recall*
Coronary Heart Disease	Coronary heart disease (C0010068) Coronary Arteriosclerosis (C0010054) Angina Pectoris (C0002962) Myocardial Infarction (C0027051)	264	62%
Cholesterol	Blood cholesterol (C0518017) Cholesterol measurement test (C0201950) Hypercholesterolemia (C0020443)	90	21%
Blood Pressure	Blood Pressure (C0005823) Systemic arterial pressure (C1272641) Diastolic blood pressure (C0428883) Hypertensive disease (C0020538)	57	86%
Coffee	Coffee (C0009237)	54	96%
Diet	Diet (C0012155) Eating (C0013470) fat intake (C0489488) salt intake (C0489767)	48	67%
Smoking	Smoking Habit (C4505437) Tobacco (C0040329) Cigar smoker (C0337666) Pipe Smoking (C4316784) Cigarette (C0677453) Cigarette smoke (substance) (C0239059)	43	72%
Body Fatness	Skinfold Thickness (C0037302) Skin-fold thickness (finding) (C0424680) Triceps skin fold thickness (observable entity) (C0518022)	29	24%
Physical Activity	Physically active (C0556453) Exercise (C0015259)	28	68%
Alcohol	Alcoholic Beverages (C0001967)	22	5%
Body Weight	Body Weight (C0005910) Weight Gain (C0043094)	13	54%
Electrocardiogram	Electrocardiogram (C0013798) Electrocardiogram finding (C0438154) Electrocardiographic changes (C0855329) Anatomical segmentation (C0441635) Abnormal T-wave (C1839341)	11	36%
Job Role	Occupations (C0028811)	7	71%
Blood Glucose	Blood Glucose (C0005802) Blood glucose measurement (C0392201)	6	67%
Somatotype	Somatotype (C0037669)	4	75%
Height	Height (C0489786)	3	100%
Heart Rate	Pulse Rate (C0232117)	3	33%
Peptic Ulcer	Peptic Ulcer (C0030920)	2	100%

Hemoglobin	Hemoglobin A measurement (C1281911) Chrysarobin (C0008721)*	2	100%
Age	Age (C0001779)	2	100%
Chest Discomfort	Chest discomfort (C0235710) Non-cardiac chest pain (C0476281)	2	50%
Chronic Cough	Chronic cough (C0010201)	1	100%
Gallbladder Disease	Gall Bladder Diseases (C0016977)	1	100%
Lipoprotein Lipase	LIPOPROTEIN LIPASE (C0023816)	1	100%
Shortness of Breath	Dyspnea (C0013404) Resting Dyspnea (C0743330)	1	100%
Early Death of Father		1	0%
Arteriovenous nicking	Retinal arteriovenous nicking (C1142247)	0	NA
Arcus Senilis	Arcus Senilis (C0003742)	0	NA
Family History	Family history (finding) (C0241889)	0	NA
TOTAL		695	59%

Table 2. Full. Pipeline Results: Mapping of Citing Papers to Findings ***(L refers to lack of association).***

Relation	*Citations*	*Recall*			
		NER+RE		LLM	
		Step 2 Only	Full Pipeline	Step 2 Only	Full Pipeline
CHD/cholesterol	85	5%	5%	93%	96%
CHD/blood pressure	57	19%	19%	95%	96%
CHD/coffee	54	61%	61%	100%	93%
CHD/smoking	43	23%	23%	93%	98%
CHD/diet (L)	41	37%	37%	73%	49%
CHD/body fatness	29	0%	0%	59%	79%
CHD/physical activity (L)	28	32%	32%	71%	36%
CHD/alcohol (L)	22	5%	5%	91%	82%
CHD/body weight (L)	13	38%	38%	69%	38%
CHD/electrocardiogram	11	0%	0%	73%	73%
diet/cholesterol	7	0%	0%	71%	0%
CHD/job role (L)	7	0%	0%	86%	86%
CHD/blood glucose (L)	6	67%	67%	50%	33%
CHD/somatotype	4	0%	0%	100%	75%
CHD/height (L)	3	67%	67%	100%	100%
CHD/heart rate	3	0%	0%	67%	100%

CHD/age	2	50%	0%	100%	0%
CHD/chest discomfort	2	0%	0%	100%	100%
CHD/peptic ulcer	2	0%	0%	100%	100%
CHD/hemoglobin (L)	2	50%	50%	100%	100%
CHD/early death of father	1	0%	0%	100%	100%
CHD/chronic cough	1	0%	0%	100%	100%
CHD/shortness of breath	1	0%	0%	100%	100%
CHD/gallbladder disease (L)	1	100%	100%	100%	100%
CHD/lipoprotein lipase (L)	1	100%	100%	0%	0%
TOTAL	426	23%	23%	86%	80%

Table 3. Original Study Findings Identified using an LLM (Google Gemini 1.5 Pro).

Factor	*Finding*
New Coronary Cases	88 cases of coronary heart disease developed (47 angina pectoris, 28 myocardial infarction, 13 deaths). Approximately one new case per 100 men per year.
Family History (Parental Longevity)	No significant difference between coronary and non-coronary groups regarding parental age at death.
Family History (Paternal Age)	Fathers of non-coronary group lived 3.4 years longer on average than fathers of coronary group.
Prior Chest Discomfort	Significantly higher development of coronary disease in men reporting prior chest discomfort ($p < 0.001$).
Chronic Cough	Significantly more frequent in coronary group (25% vs 12%, $p < 0.001$).
Shortness of Breath	Significantly more frequent in coronary group (18% vs 11%, $p < 0.025$).
Peptic Ulcer History	Significantly more frequent in coronary group (18% vs 10%, $p < 0.025$).
Gallbladder Disease History	No significant relationship with coronary disease.
Height and Weight	No significant difference between groups.
Body Fatness (Skinfold Thickness)	Significantly greater in coronary group ($p < 0.025$ for triceps, $p < 0.01$ for scapular).
Somatotype	Suggests endomorphic dominance may be a factor, but requires larger sample size to confirm.
Resting Pulse Rate	While mean pulse rate was not significantly different, distribution differed significantly ($p = 0.025$), with coronary group having more extreme cases.
Blood Pressure	Positive relationship between elevated systolic ($p < 0.001$) and diastolic ($p < 0.05$) blood pressure and coronary disease.
Arcus Senilis	No relationship with coronary disease.

Arteriovenous Nicking (Fundi)	Significantly associated with coronary disease (p < 0.005).
Perceived Tension	No significant difference between groups.
Hemoglobin Levels	No significant difference between groups.
Cholesterol Levels	Significantly higher in coronary group (p < 0.01), with stepwise increases across angina, infarction, and death subgroups.
Lipoprotein Lipase	No relationship with coronary disease.
Blood Glucose (2-hour post 100g glucose)	No significant difference in mean levels, but significant heterogeneity of variance within groups, particularly angina.
Electrocardiogram (ST-segment/T-wave abnormalities)	Significantly associated with coronary disease (p < 0.005).
Job Type	No association with coronary disease.
Off-Job Physical Activity	No striking differences, though coronary group tended to report less sports participation.
Cigarette Smoking	Significant association with coronary disease (p < 0.005), with a stepwise increase in risk across angina, infarction, and death subgroups.
Diet (excluding coffee)	No significant association with coronary disease within the observed range of fat intake.
Coffee Consumption	Significant association with coronary disease (p < 0.025).
Alcohol Consumption	No association with coronary disease.