Forgotten Knowledge: Examining the Citational Amnesia in NLP

Janvijay Singh[*♣], **Mukund Rungta**[*♣], **Diyi Yang**[◊], **Saif M. Mohammad**[Φ]
[♣]Georgia Institute of Technology, [◊]Stanford University, [Φ]National Research Council Canada
{iamjanvijay, mrungta8}@gatech.edu, diyiy@cs.stanford.edu,
saif.mohammad@nrc-cnrc.gc.ca

Abstract

Citing papers is the primary method through which modern scientific writing discusses and builds on past work. Collectively, citing a diverse set of papers (in time and area of study) is an indicator of how widely the community is reading. Yet there is little work looking at broad temporal patterns of citation. This work, systematically and empirically examines: *How far back in time do we tend to go to cite papers? How has that changed over time, and what factors correlate with this citational attention/amnesia?* We chose NLP as our domain of interest, and analyzed ~71.5K papers to show and quantify several key trends in citation. Notably, ~62% of cited papers are from the immediate five years prior to publication, whereas only ~17% are more than ten years old. Furthermore, we show that the median age and age diversity of cited papers was steadily increasing from 1990 to 2014, but since then the trend has reversed, and current NLP papers have an all-time low temporal citation diversity. Finally, we show that unlike the 1990s, the highly cited papers in the last decade were also papers with the least citation diversity; likely contributing to the intense (and arguably harmful) recency focus. Code, data, and a demo are available at the project homepage. [1] [2]

1 Introduction

Study the past if you would define the future.

— Confucius

The goal of scientific research is to create a better future for humanity. To do this we innovate on ideas and knowledge from the past. Thus, a central characteristic of the scientific method and modern scientific writing is to discuss other work: to build on ideas, to critique or reject earlier conclusions, to borrow ideas from other fields, and to situate the proposed work. Even when proposing something that others might consider dramatically novel, it is widely believed that these new ideas have been made possible because of a number of older ideas (Verstak et al., 2014). *Citation* (referring to another paper in a prescribed format) is the primary mechanism to point the reader to these prior pieces of work and also to assign credit for shaping current work (Mohammad, 2020a; Rungta et al., 2022). Thus, we argue that examining citation patterns across time can lead to crucial insights into what we value, what we have forgotten, and what we should do in the future.

Of particular interest is the extent to which good older work is being forgotten — *citational amnesia*. More specifically, for this paper, we define citational amnesia as shown below:

> *Citational Amnesia*: the tendency to not cite enough relevant good work from the past (more than a few years old).

We cannot directly measure citational amnesia empirically because determining "enough", "relevance", and "good" require expert researcher judgment. However, what we can measure is the collective tendency of a field to cite *older* work. Such an empirical finding enables reflection on citational amnesia. A dramatic drop in our tendency to cite older work should give us cause to ponder whether we are putting enough effort to read older papers (and stand on the proverbial shoulders of giants).

Note that we are not saying that old work should be cited simply because it exists. We are saying that we should consciously reflect on the diversity of the papers we explore when conducting research. Diversity can take many forms, including reading relevant papers from diverse fields, by authors from diverse regions, and relevant papers published from various time periods — the focus of this paper. Exploring a diverse set of papers allows us to ben-

[*]Equal contribution.

[1]Code, data: https://github.com/iamjanvijay/CitationalAmnesia/

[2]Online demo: https://huggingface.co/spaces/mrungta8/CitationalAmnesia/

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6192–6208

efit from important and diverse research perspectives. Looking at older literature makes us privy to broader trends, and informs us in ways that are beneficial well beyond the immediate work.

Historically, citational amnesia was impacted by various factors around access and invention. For example, the invention of the printing press in the year 1440 allowed a much larger number of people to access scientific writing (Eisenstein, 1985). The era of the internet and digitization of scientific literature that began in the 1990s also greatly increased the ease with which one could access past work (Verstak et al., 2014). However, other factors such as the birth of paradigm-changing technologies may also impact citation patterns; ushering in a trend of citing very new work or citing work from previously ignored fields of work. Such dramatic changes are largely seen as beneficial; however, strong tailwinds may also lead to a myopic focus on recent papers and those from only some areas, at the expense of benefiting from a wide array of work (Pan et al., 2018; Martín-Martín et al., 2016).

We choose as our domain of interest, papers on Natural Language Processing (NLP), specifically those in the ACL Anthology. This choice is motivated by the fact that NLP (and other related fields of Artificial Intelligence) are in a period of dramatic change: There are notable and frequent gains on benchmark datasets; NLP technology is becoming increasingly ubiquitous in society; and new sub-fields of NLP such as Computational Social Science, Ethics and NLP, and Sustainable NLP are emerging at an accelerated rate. The incredibly short research-to-production cycle and move-fast-and-break-things attitude in NLP (and Machine Learning more broadly) has also led to considerable adverse outcomes for various sections of society, especially those with the least power (Buolamwini and Gebru, 2018; ARTICLE19, 2021; Mohammad, 2021). Thus reading and citing more broadly is especially important now.

In this work, we compiled a temporal citation network of 71.5K NLP papers that were published between 1990 and 2021, along with their meta-information such as the number of citations they received in each of the years since they were published — the *Age of Citations (AoC) dataset*. We use AoC to answer a series of specific research questions on *what we value, what we have forgotten, what factors are associated with this citational attention/amnesia, what are the citation patterns of different types of papers,* and *how these citation patterns have changed over time*. Finally, we show that many of the highly cited papers from the past decade have very low temporal citation diversity; and because of their wide reach, may have contributed to the intense recency focus in NLP. All of the data and code associated with the project will be made freely available on the project homepage.

2 Related Work

In the broad area of Scientometrics (study of quantitative aspects of scientific literature), citations and their networks have been studied from several perspectives, including: paper quality (Buela-Casal and Zych, 2010), field of study (Costas et al., 2009), novelty, length of paper (Antoniou et al., 2015; Falagas et al., 2013), impact factor (Callaham et al., 2002), venue of publication (Callaham et al., 2002; Wahle et al., 2022), language of publication (Lira et al., 2013), and number of authors (Della Sala and Brooks, 2008; Bosquet and Combes, 2013), collaboration (Nomaler et al., 2013), self-citation (Costas et al., 2010), as well as author's reputation (Collet et al., 2014), affiliation (Sin, 2011; Lou and He, 2015), geographic location (Nielsen and Andersen, 2021; Lee et al., 2010; Pasterkamp et al., 2007; Paris et al., 1998), gender, race and age (Ayres and Vars, 2000; Leimu and Koricheva, 2005; Chatterjee and Werner, 2021; Llorens et al., 2021).

However, there has been relatively little work exploring the temporal patterns of citation. Verstak et al. (2014) analyzed scholarly articles published in 1990–2013 to show that the percentage of older papers being cited steadily increased from 1990 to 2013, for seven of the nine fields of study explored. (They treated papers that were published more than ten years before a particular citation as *old papers*.) For Computer Science papers published in 2013, on average, 28% of the cited papers were published more than ten years before. This represented an increase of 39% from 1990. They attributed this increasing trend in citing old papers to the ease of access of scientific literature on the world wide web, as well as the then relatively new scientific-literature-aggregating services such as Google Scholar.

Parolo et al. (2015) analyzed about 25 million papers from Clinical Medicine, Molecular Biology, Physics, and Chemistry published until 2014 to show that typically the number of citations a paper receives per year increases in the years after publication, reaches a peak, and then decays exponentially. Interestingly they showed that this rate of decay was increasing in the more recent papers of their study. They attribute this quicker decay (or more "forgetting" of recent papers) to the substantial increase in the number of publications; a lot more papers are being published, and due to the limited attention span of subsequent researchers, on average, papers are being forgotten faster.

Past work on NLP papers and their citations includes work on gender bias (Schluter, 2018; Vogel and Jurafsky, 2012; Mohammad, 2020b), author location diversity (Rungta et al., 2022), author institution diversity (Abdalla et al., 2023), and on broad general trends such as average number of citations over time and by type of paper (Mohammad, 2020a,c; Wahle et al., 2022).

Bollmann and Elliott (2020) were the first to explore the recency bias of citations in NLP papers. They showed that the ACL Anthology papers published between 2017 and 2019 cited more recent work than papers published between 2010 and 2014. Question 3 in Section 4 of our paper is of the same spirit that was explored in their work; however, our work examines a much larger spread of NLP papers (published between 1965 and 2021). This will shed light on the reproducibility of those findings and, more importantly, determine the broader trajectory of temporal citation patterns (from the start of ACL to present day). Additionally, our work introduces a new citation age diversity metric to quantify the degree of spread of citations over time, as well as an interactive online demo system to visualize the citation age diversity of any paper. Going beyond how overall citation patterns have changed over time, our work takes a deep dive into six other novel research questions, notably around temporal citation patterns in subareas of NLP, of cited topics, and across sparsely and highly cited papers.

3 Dataset

The ACL Anthology (AA) Citation Corpus (Rungta et al., 2022) contains meta data (paper title, year of publication, and venue, etc. for the 71,568 papers in the ACL Anthology repository (published until January 2022). We used the Semantic Scholar API[3] to gather the references for each paper in the AA Citation Corpus, using the paper's unique Semantic Scholar ID (SSID). This allowed us to obtain additional information about the *cited papers*, such as their title, year of publication, and venue of publication. Note that these cited papers may or may not be part of AA. To study the dynamics of citations over time, we constructed year-wise citation networks using the data collected. Specifically, we created the citation networks for every year from 1965 to 2001. This representation of citation data allows us to answer several interesting questions, such as the number of citations a paper receives in a particular year after its publication. We refer to this dataset as *Age of Citations (AoC) dataset*.

[3]https://www.semanticscholar.org/

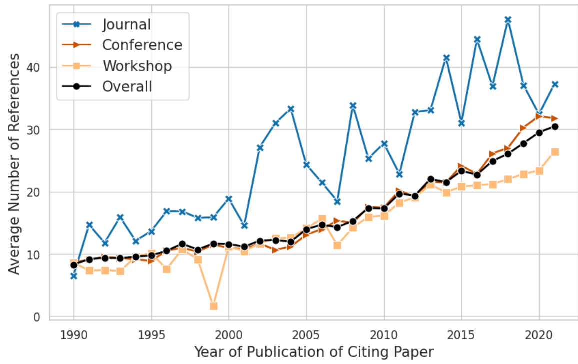


Figure 1: Average number of unique references in an AA paper published in different years.

	Mean	Median
Journal	23.24	15
Conference	21.11	19
Workshop	19.07	17
Overall	20.63	18

Table 1: Mean and median of the number of unique references in an AA paper.

4 Age of Citation

We used the *AoC dataset* to answer a series of questions on how research papers are cited and the trends across years.

Q1. What is the average number of unique references in the AA papers? How does this number vary by publication type, such as workshop, conference, and journal? Has this average stayed roughly the same or has it changed markedly over the years?

Ans. We calculated the average number of unique references for all papers in the *AoC dataset*, as well as for each publication type (workshops, conferences, and journals). We then binned all papers by publication year, computed the mean and median for each bin for each year.

Results The scores are shown in Table 1. Figure 1 shows how the mean has changed across the years.[4] The graph shows a general upward trend. The trend seems roughly linear until the mid 2000s, at which point we see that the slope of the trend line increases markedly. Even just considering the last 7 years, there has been a 41.74% increase in referenced papers in 2021 compared to 2014.

Similar overall trends can be observed when papers are grouped by publication type. Not surprisingly, the longer journal articles cite markedly more papers than conference and workshop papers. The plot for conferences and workshops is relatively smooth compared to journal articles. This is because the number of papers for each year in journals is far less. For example, in the year 2015, only 139 papers were published in journals, whereas 1709 and 983 papers were published in conferences and workshops respectively.

Discussion The steady increase in the number of unique references from 1965 is likely because of the increasing number of relevant papers as the field develops and grows. However, it is interesting that this growth has not plateaued even after 55 years. By the late-2000s, with the advent of widely accessible electronic proceedings, *ACL venues started experimenting with more generous page limits: relaxing it from a strict 8 pages to first allowing one or two additional pages for references to eventually allowing unlimited pages for references.[5] Other factors that may have contributed to more papers being referred to (cited) within a paper, include: an additional page for incorporating reviewer comments, allowing Appendices, and the inclusion of an increasing number of experiments.

Q2. On average, how far back in time do we go to cite papers? As in, what is the average age of cited papers? What is the distribution of this age across all citations? How do these vary by publication type?

Ans. If a paper x cites a paper y_i, then the age of the citation (*AoC*) is taken to be the difference between the year of publication (*YoP*) of x and y_i:

$$AoC(x, y_i) = YoP(x) - YoP(y_i) \tag{1}$$

We calculated the AoC for each of the citations in the AoC dataset. For each paper, we also calculated the mean AoC of all papers cited by it:

$$mAoC(x) = \frac{1}{N}\sum_{i=1}^{N} AoC(x, y_i) \tag{2}$$

here N refers to the number of papers cited by x.

Results The average mAoC for all the papers in the *AoC dataset* is 6.01. The scores were 7.16 for journal articles, 5.91 for conference papers, and 6.01 for workshop papers. Figure 2 shows the distribution of AoCs in the dataset across the years after the publication of the *cited* paper (overall, and across publication types). For example, the y-axis point for year 0 corresponds to the average of the percentage of citations papers received in the same year as it they were published. The y-axis point for year 1 corresponds to the average of percentage of citations the papers received in the year after they were published. And so on.

Observe that the majority of the citations are for papers published one year prior, (AoC = 1). This is true for conference and workshop subsets as well, but in journal papers, the most frequent citations are for papers published two years prior. Overall though all the arcs have a similar shape, rising sharply from the number in year 0 to the peak value and then dropping off at an exponential rate in the years after the peak is reached. For the full set of citations, this exponential decay from the peak has a half life of about 4 years. Roughly speaking, the line plot for journals is shifted to the right by a year compared to the line plots for conferences and workshops. It also has a lower peak value and its citations for the years after the peak are at a higher percentage than those for conferences and workshops. Additionally, citations in workshop papers have the highest percentage of current year citations (age 0), whereas citations in journal article have the lowest percentage of current year citations.

Analogous to Figure 2, Figure 3 presents the distribution of AoCs, albeit broken down by the total citations received by a paper. It is worth noting that the distribution leans more towards the right for papers with a higher number of citations. This shows that papers with a higher citation count continue to receive significant citations even far ahead in the future, which is intuitive.

Discussion Overall, we observe that papers are cited most in years immediately after publication, and their chances of citation fall exponentially after

[4]The numbers of AA papers published each year until 1990 were rather low, and so in Figure 1, we only show the trajectory from 1990. However, note that the numbers generally increase even from 1965 to 1990.

[5]In 2008, EMNLP became the first major NLP venue to allow an extra page for references. (ACL followed in 2009.)

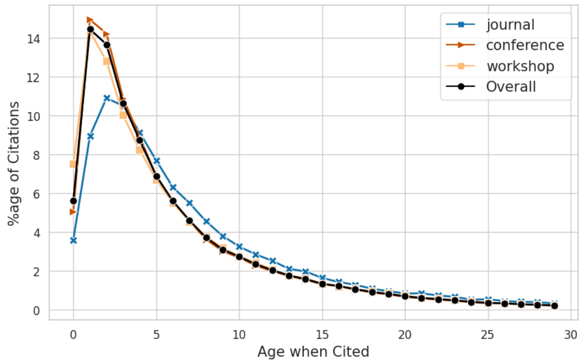


Figure 2: Distribution of AoC for papers in AA (overall and by publication type).

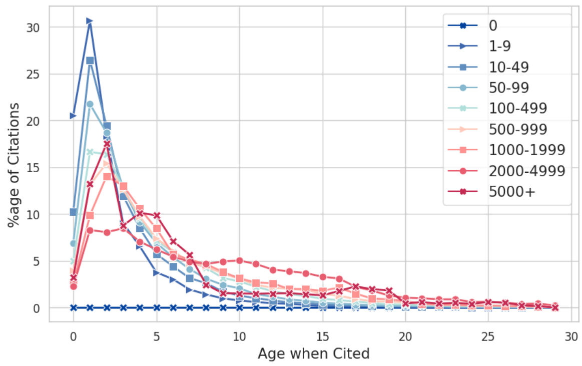


Figure 3: Distribution of AoC for AA papers with different citation counts (shown in legend).

that. The slight right-shift for the journal article citations is likely, at least in part, because journal submissions have a long turn-around time from the first submission to the date of publication (usually between 6 and 18 months). A list of the oldest papers cited by AA papers is available on the project's GitHub repository.

Q3. What is the trend in the variation of *AoC* over time and how does this variation differ across different publication venues in NLP?

Ans. To answer this question, we split the papers into bins corresponding to the year of publication, and then examined the distribution of $mAoC$ in each bin. We define a new metric called the *Citation Age Diversity (CAD) Index*, which measures the diversity in the $mAoC$ for a set of papers. In simpler terms, a higher *CAD Index* indicates that $mAoCs$ covers a broader range, implying that the cited papers span a wider time period of publication. This metric offers valuable insights into the temporal spread of scholarly influence and the long-term impact of research. Precisely, the *CAD Index* for a bin of papers b, is defined using the Gini

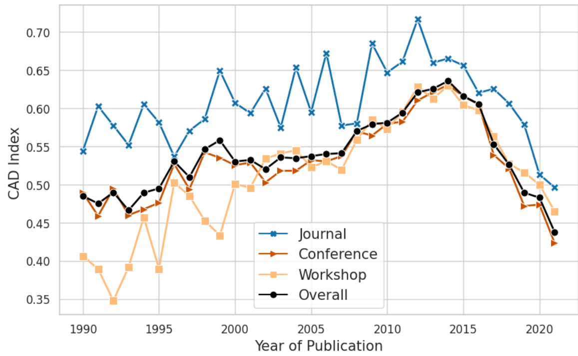


Figure 4: Citation Age Diversity Index across years.

Coefficient as follows:

$$CAD(b) = 1 - \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{|mAoC(b_i) - mAoC(b_j)|}{2N^2\bar{b}} \quad (3)$$

here, b_i corresponds to i^{th} paper within bin b, N denotes the total number of papers in bin b and $\bar{b}$ represents the mean of $mAoC$ of papers' associated with bin b. A *CAD Index* close to 0 indicates minimum temporal diversity in citations (citing papers from just one year), whereas a *CAD Index* of 1 indicates maximum temporal diversity in citations (citing papers uniformly from past years). In addition to *CAD Index*, we also compute median $mAoC$ of each such yearly bin. The results for both *CAD Index* and median $mAoC$ have roughly identical trends across the years. We discuss the *CAD Index* analysis below. (The discussion of the median $mAoC$ results is in the Appendix A.1.)

Results Figure 4 shows the *CAD Index* across years (higher *CAD Index* indicates high diversity), and across different publication types. The *CAD Index* plot of Figure 4 shows that the temporal diversity of citations had an increasing trend from 1990 to 2014, but the period from 1998 to 2004, and 2014 to 2021 (dramatically so) were periods of decline in temporal diversity (decreasing *CAD Index* scores). These intervals coincide with the year intervals in which we observed a decreasing trend in median $mAoC$ of published papers (discussed in the Appendix). This suggests that the increase or decrease in diversity is largely because of the decreased or increased focus on papers from recent years, respectively.

The *CAD Index* plots by publication type all have similar trends, with journal paper submissions consistently having markedly higher scores (indicating markedly higher temporal diversity) across the years studied. However, they also seem to be

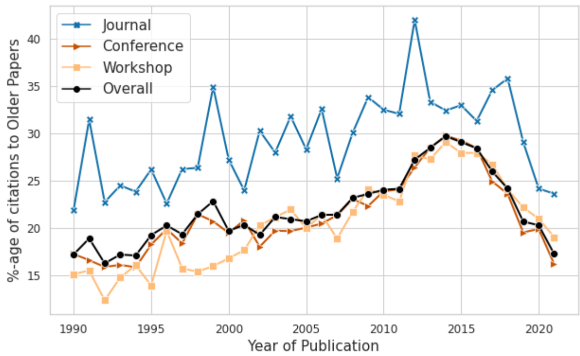


Figure 5: Percentage of citations in AA papers where the cited paper is at least 10 years old.

most impacted by the trend since 2014 to cite very recent papers. (*CAD Index* not only goes back to the 1990 level, but also undershoots beyond it.)

Discussion Overall, we find that all the gains in temporal diversity of citations from 1990 to 2014 (a period of 35 years), have been negated in the 7 years from 2014. This change is driven largely by the deep neural revolution in the early 2010's and strengthened further by the substantial impact of transformers on NLP and Machine Learning. Interestingly, our results until 2013 are in line with what Verstak et al. (2014) found for many fields of study, but since 2014 there has been a marked shift in trends in NLP. We hope future work will explore whether similar shifts in trends have occurred in other fields. Our results add to (and are consistent with) the mean-citation age results found by Bollmann and Elliott (2020), who examined mean citation age between 2010 and 2019. Our analysis of the broader period (from 1965 to 2021), situates those results in the overall trajectory of how temporal citation patterns have evolved since the beginning of the Association of Computational Linguistics to the present period. Additionally, the new CAD Index metric quantifies the degree temporal citation diversity as opposed to the recency focus of citations captured by mean citation age.

Q4. What percentage of cited papers are old papers? How has this varied across years and publication venues?

Ans. Just as Verstak et al. (2014), we define a cited paper as *older* if it was published at least ten years prior to the citing paper. We then divided all AA papers into groups based on the year in which they were published. For each AA paper, we determined the number of citations to older papers.

Results Figure 5 shows the percentage of older papers cited by papers published in different years. Observe that this percentage increased steadily from 1990 to 1999, before decreasing until 2002. After 2002, the trend of citing older papers picked up again; reaching an all time high of ~30% by 2014. However, since 2014, the percentage of citations to older papers has dropped dramatically, falling by 12.5% and reaching a historical low of ~17.5% in 2021. Similar patterns are observed for different publication types. However, we note that a greater (usually around 5% more) percentage of a journal paper's citations are to older papers, than in conference and workshop papers.

Discussion These results confirm that the trends in diversity discussed in Q2 are aligned with the trends in citing older papers. This dramatic drop in citing older papers since 2014 can largely be attributed to the explosion of paper count and the paradigm shift in the field of NLP brought on by deep learning and transformers.

Q5. What is the *mAoC* distribution for different areas within NLP? Relative to each other, which areas tend to cite more older papers and which areas have a strong bias towards recent papers?

Ans. The ACL Anthology does not include metadata for sub-areas within NLP. Further, a paper may be associated with more than one area and the distinction between areas can often be fuzzy. Thus, we follow a rather simple approach used earlier in Mohammad (2020b): using paper title word bigrams as indicators of topics relevant to the paper. A paper with *machine translation* in the title is likely to be relevant to the area of machine translation. Using title bigrams for this analysis also allows for a finer analysis within areas. For example, two bigrams pertaining to finer subareas within the same area can be examined separately. (Papers in different sub-areas of an area need not be similar in terms of the age of the papers they cite.)

We first compiled a list of the top 60 most frequent bigrams from the titles of AA papers. Next, for each of these bigrams, we created a bin containing all AA papers that had that bigram in their title.[6] For each paper included in any of these bins, we computed $mAoC$. Finally, we plotted the distribution of $mAoC$ values for the papers in each bin,

[6]A single paper may be included in multiple bins.

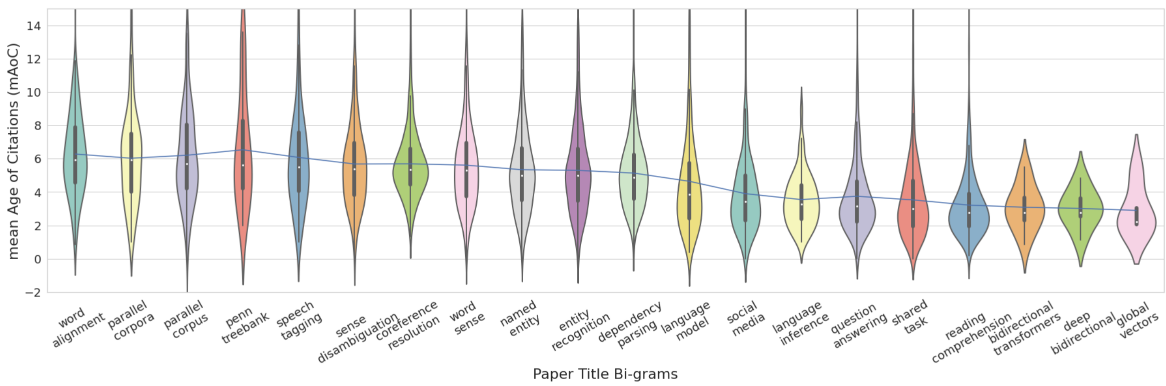


Figure 6: Distribution of $mAoC$ for frequent bigrams appearing in the titles of citing papers.

as shown in Figure 6. Note that, for the purpose of improving the visibility of the plot, only selected $mAoC$ distributions are depicted in the figure 6. We then examined the distribution of $mAoC$ for each of these bins.

Results Figure 6 shows the $mAoC$ violin plots for each of the bins pertaining to the title bigrams (in decreasing order of median $mAoC$). Observe that papers with the title bigrams *word alignment, parallel corpus/corpora, Penn Treebank, sense disambiguation* and *word sense* (common in the word sense disambiguation area), *speech tagging, coreference resolution, named entity* and *entity recognition* (common in the named entity recognition area), and *dependency parsing* have some of the highest median $mAoC$ (cite more older papers). In contrast, papers with the title bigrams *glove vector, BERT pre, deep bidirectional,* and *bidirectional transformers* (which correspond to new technologies) and papers with title bigrams *reading comprehension, shared task, question answering, language inference, language models,* and *social media* (which correspond to NLP subareas or domains) have some of the lowest median $mAoC$ (cite more recent papers).

Discussion The above results suggest that not all NLP subfields are equal in terms of the age of cited papers. In fact, some papers cited markedly more newer papers than others. This could be due to factors such as early adoption or greater applicability of the latest developments, the relative newness of the area itself (possibly enabled by new inventions such as social media), etc.

Q6. What topics are more pronounced in cited papers across different periods of time?

Ans. To address this question, we partitioned the research papers into those published between: 1990–1999, 2000–2009, 2010–2015, and 2016–2021.[7] For papers from each period: we first extracted all unigrams and bigrams from the titles of the cited papers. Next, for the top 100 most frequent unigrams and bigrams, we calculated the percentage of all citations that had the respective ngram in the cited paper's title — *the ngram citation percentage*.

Results Upon examining various bigram citation percentages, we found that bigrams pertaining to areas such as tree-adjoining grammars have been in decline since the 1990s (cited less as with every subsequent interval). Bigrams pertaining to areas such as conditional random fields and coreference resolution gained momentum in the middle periods (2000–2016) but have since lost popularity post-2016. On the other hand, techniques such as domain adaptation have consistently gained momentum since the 2010s. Post-2016 keywords related to deep learning technologies such as *convolutional neural nets, deep bi-directional, deep learning, deep neural, Global vectors,* and *jointly learning* experienced a substantial surge in popularity. Additionally, certain areas such as cross-lingual and entity recognition consistently gained momentum since the 1990s.

Upon examining various unigram citation percentages, we found that deep-learning-related terms such as *attention, bert, deep, neural, embeddings,* and *recurrent* saw a substantial increase in citation post-2016. Furthermore, we observed that since the 1990s, there has been a growing trend in NLP papers towards citing research on the social aspects of language processing, as evidenced by the increasing popularity of keywords such as *social* and *sentiment*.

[7] The 2010–2021 period was split into two because of the large number of papers published.

Figures 9 and 10 in the Appendix show a number of unigrams and bigrams with the most notable changes in the ngram citation percentage across the chosen time intervals.

Q7. Do well-cited papers cite more old papers and have more *AoC* diversity?

Ans. We introduce three hypotheses to explore the correlation between temporal citation patterns of target papers and the number of citations the target papers themselves get in the future.

H1. The degree of citation has no correlation with temporal citation patterns of papers.

H2. Highly cited papers have more temporal citation diversity than less cited papers.

H3. Highly cited papers have less temporal citation diversity than less cited papers.

Without an empirical experiment, it is difficult to know which hypothesis is true. H1 seemed likely, however, there were reasons to suspect H2 and H3 also. Perhaps cite more widely is correlated with other factors such the quality of work and thus correlates with higher citations (supporting H2). Or, perhaps, early work in a new area receives lots of subsequent citations and work in a new area often tends to have limited citation diversity as there is no long history of publications in the area (supporting H3).

On, Nov 30, 2022, we used the Semantic Scholar API to extract the number of citations for each of the papers in the AoC dataset. We divided the AoC papers into nine different bins as per the number of citations: 0, 1–9, 10–49, 50–99, 100–499, 500–999, 1000–1999, 2000–4999, or 5000+ citations. For each bin, we calculated the mean of $mAoC$ and *CAD Index*. We also computed the Spearman's Rank Correlation between the *CAD Index* of the citation bins and the mean of the citation range of each of these bins.

Results Figure 7 shows the $mAoC$ and *CAD Index* for each bin (a) for the full AoC dataset, and (b) for the subset of papers published between 1990 and 2000. (Figures 11a and 11b in the Appendix show plots for papers from two additional time periods.) On the full dataset (Figure 7a), we observe a clear pattern that the *CAD Index* decreases with increasing citation bin (with the exception of papers in the 1K–2K and 2K–5K bins). The mean $mAoC$ follows similar trend w.r.t. the *CAD Index*.

1990–99	2000-09	2010–15	1965–2021 (All)
0.16	-1.00*	-0.97*	-0.72*

Table 2: Correlation between the mean number of citations received and *CAD Index* for papers from various time periods. The * indicates statistically significant correlation (p-value < 0.05).

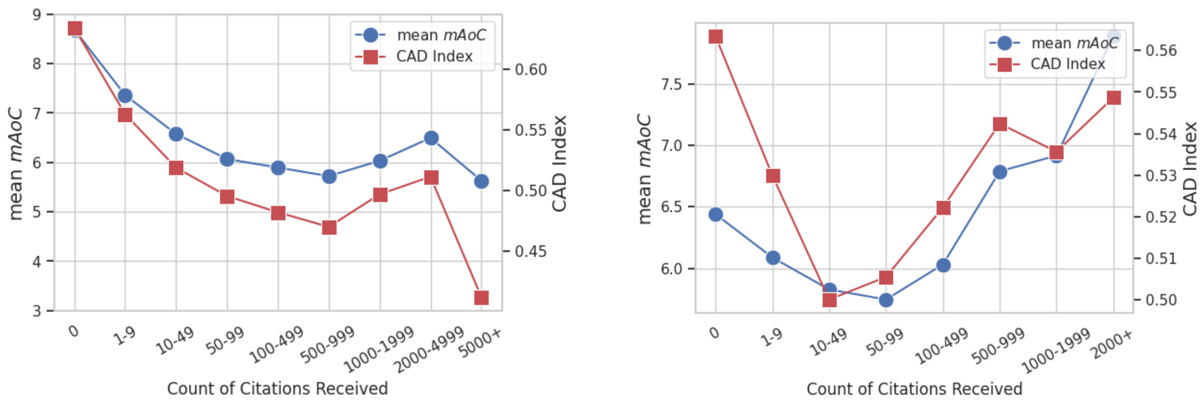
These results show that, for the full dataset, the higher citation count papers tend to have less temporal citation diversity than lower-citation count papers. However, on the 1990s subset (Figure 7b), the *CAD Index* decreased till the citation count < 50 and increased markedly after that. This shows that during the 1990s, the highly cited papers also cited papers more widely in time. Plots for the 2000s and 2010s (Figure 11) follow a similar trend as the overall plot (Figure 7a), indicating that trend of highly cited papers having less temporally diverse citations started around the year 2000.

The Spearman's rank Correlation Coefficients between the mean number of citations for a bin and the mean $mAoC$ of the citation bins are shown in Table 2.[8] Observe that for the 1990's papers there is essentially no correlation, but there are strong correlations for the 2000s, 2010s, and the full dataset papers.

Similar to Figure 7a, in Figure 12 (in the Appendix) we show how mean *mAoC* and *CAD Index* of AA papers published between 1965 and 2021 but when broken down by *research topics*. This examination across various research topics consistently shows a trend: the higher the citations, the lower the age diversity of citations. This may be because "mainstream" work in an area tends to cite lots of other very recent work and brings in proportionately fewer ideas from the past. In contrast, "non-mainstream" work tends to incorporate proportionally more ideas from outside, yet receives fewer citations as there may be less future work in that space to cite it.

Discussion Papers may receive high citations for a number of reasons; and those that receive high citations are not necessarily model research papers. While they may have some aspects that are appreciated by the community (leading to high citations), they also have flaws. High-citation papers (by definition) are more visible to the broader research community and are likely to influence early researchers more. Thus their strong recency focus in

[8] We did not consider 2016–2021 papers because they have had only a few years to accumulate citations.



(a) Papers published between 1965 and 2021.

(b) Papers published between 1990 and 2000.

Figure 7: Variation of mean $mAoC$ and Citation Age Diversity (CAD) Index (shown on y-axis) for papers with different citation counts (shown on x-axis).

citations is a cause of concern. Multiple anecdotal incidents in the community have suggested how early researchers often consider papers that were published more than two or three years back as "*old papers*". This goes hand-in-hand with a feeling that they should not cite old papers and therefore, do not need to read them. The lack of temporal citation diversity in recent highly cited papers may be perpetuating such harmful beliefs.

5 Demo: CAD Index of Your Paper

To encourage authors to be more cognizant of the age of papers they cite, we created an online demonstration page where one can provide the Semantic Scholar ID of any paper and the system returns the number of papers referenced, mean Age of Citation (mAoC), top-5 oldest cited papers, and their years of publication.[9] Notable, the demo also plots the distribution of mAoC for all the considered papers (all papers published till 2021) and compares it with mean Age of Citation of the input paper. Figure 13 in the Appendix shows a screenshot of the demo portal for an example input.

6 Conclusions and Discussion

This work looks at temporal patterns of citations by presenting a set of comprehensive analyses of the trend in the diversity of age of citations and the percentage of older papers cited in the field of NLP. To enable this analysis, we compiled a dataset of papers from the ACL Anthology and their meta-information; notably, the number of citations they received each year since they were published.

We showed that both the diversity of age of citations and the percentage of older papers cited increased from 1990 to 2014, but since then there has been a dramatic reversal of the trend. By the year 2021 (the final year of analysis), both the diversity of age of citations and the percentage of older papers cited have reached historical lows. We also studied the correlation between the number of citations a paper receives and the diversity of age of cited papers, and found that while there was roughly no correlation in the 1990s, the 2000s marked the beginning of a period where the higher citation levels correlated strongly with lower temporal citation diversity.

It is a common belief among researchers in the field that the advent of deep neural revolution in the early 2010's has led us to cite more recent papers than before. This analysis confirms and quantifies the extent to which temporal diversity is reduced in this recent period. In fact, it shows that the reduction in temporal diversity of citations is so dramatic that it has wiped out steady gains from 1990 to 2014. While some amount of increased focus on recent papers is expected (and perhaps beneficial) after large technological advances, an open question, now, is whether, as a community, we have gone too far, ignoring important older work. Our work calls for an urgent need for reflection on the intense recency focus in NLP: How are we contributing to this as researchers, advisors, reviewers, area chairs, and funding agencies?[10]

[9] Online demo: https://huggingface.co/spaces/mrungta8/CitationalAmnesia/

[10] This paper cites 16 papers published ten or more years back (35% of the cited papers).

7 Ethics Statement

This paper analyses scientific literature at an aggregate level. The ACL Anthology freely provides information about NLP papers, such as their title, authors, and year of publication. We do not make use of or redistribute any copyrighted information. All of the analyses in this work are at aggregate-level, and not about individual papers or authors. In fact, we desist from showing any breakdown of results involving 30 or fewer papers to avoid singling out a small group of papers.

8 Limitation

A limitation of this study is that it is based solely on papers published in the ACL Anthology, which primarily represents the international English-language NLP conference community. While the ACL Anthology is a reputable source of NLP research, it should be acknowledged that a significant amount of research is also published in other venues such as AAAI, ICLR, ICML, and WWW. Additionally, there are also vibrant local NLP communities and venues, often publishing in non-English languages, that are not represented in the ACL Anthology. As a result, the conclusions drawn from our experiments may not fully capture the global landscape of NLP research and further work is needed to explore the diversity of sub-communities and venues across the world.

This work focuses on the aggregate trends of citing older work in NLP, but does not investigate the reasons for lower citation of certain older papers. There may be various factors that contribute to this, such as the accessibility to these older papers, the large number of recent papers, the applicability of these old works, and the technical relevance of the older work. Determining the relative impact of each reason is a challenging task. Therefore, more research is needed to fully understand the underlying mechanisms that influence the citation of older NLP papers.

This study aims to investigate the factors that contribute to the citation of older works in the field of NLP. We have analyzed different factors such as the mean age of citation, diversity in the age of citations, venue of publication, and subfield of research. Our results indicate that these factors are associated with the citation of older works, but it should be noted that these associations do not establish any causal relationship between them.

Lastly, it is important to note that citations can be heterogeneous and can be categorized in different ways. For example, some classifications of citations include background, method, and result citations. However, certain citations may be more important than others, as shown by previous research such as "*Identifying Meaningful Citations*" by (Valenzuela-Escarcega et al., 2015).

Acknowledgments

Many thanks to Roland Kuhn, Rebecca Knowles, and Tara Small for thoughtful discussions.

References

Mohamed Abdalla, Jan Philip Wahle, Terry Ruas, Aurélie Névéol, Fanny Ducel, Saif M. Mohammad, and Karën Fort. 2023. The elephant in the room: Analyzing the presence of big tech in natural language processing research.

George A Antoniou, Stavros A Antoniou, Efstratios I Georgakarakos, George S Sfyroeras, and George S Georgiadis. 2015. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature. *Annals of vascular surgery*, 29(2):286–292.

ARTICLE19. 2021. Emotional entanglement: China's emotion recognition market and its implications for human rights.

Ian Ayres and Fredrick E Vars. 2000. Determinants of citations to articles in elite law reviews. *The Journal of Legal Studies*, 29(S1):427–450.

Marcel Bollmann and Desmond Elliott. 2020. On forgetting to cite older papers: An analysis of the ACL Anthology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7819–7827, Online. Association for Computational Linguistics.

Clément Bosquet and Pierre-Philippe Combes. 2013. Are academics who publish more also more cited? individual determinants of publication and citation records. *Scientometrics*, 97(3):831–857.

Gualberto Buela-Casal and Izabela Zych. 2010. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. *Psicothema*, pages 270–276.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Michael Callaham, Robert L Wears, and Ellen Weber. 2002. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama*, 287(21):2847–2850.

Paula Chatterjee and Rachel M Werner. 2021. Gender disparity in citations in high-impact journal articles. *JAMA Network Open*, 4(7):e2114509–e2114509.

François Collet, Duncan A Robertson, and Daniela Lup. 2014. When does brokerage matter? Citation impact of research teams in an emerging academic field. *Strategic Organization*, 12(3):157–179.

Rodrigo Costas, Maria Bordons, Thed N Van Leeuwen, and Anthony FJ Van Raan. 2009. Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers. *Journal of the American Society for Information Science and Technology*, 60(4):740–753.

Rodrigo Costas, Thed van Leeuwen, and María Bordons. 2010. Self-citations at the meso and individual levels: effects of different calculation methods. *Scientometrics*, 82(3):517–537.

Sergio Della Sala and Joanna Brooks. 2008. Multi-authors' self-citation: A further impact factor bias? *Cortex; a journal devoted to the study of the nervous system and behavior*, 44(9):1139–1145.

Elizabeth L Eisenstein. 1985. The printing press as an agent of change. *Cambridge: Cambridge*.

Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, and Michael N Mavros. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. *PLoS One*, 8(2):e49476.

Shi Young Lee, Sanghack Lee, and Sung Hee Jun. 2010. Author and article characteristics, journal quality and citation in economic research. *Applied Economics Letters*, 17(17):1697–1701.

Roosa Leimu and Julia Koricheva. 2005. What determines the citation frequency of ecological papers? *Trends in ecology & evolution*, 20(1):28–32.

Rodrigo Pessoa Cavalcanti Lira, Rafael Marsicano Cezar Vieira, Fauze Abdulmassih Gonçalves, Maria Carolina Alves Ferreira, Diana Maziero, Thais Helena Moreira Passos, and Carlos Eduardo Leite Arieta. 2013. Influence of English language in the number of citations of articles published in brazilian journals of ophthalmology. *Arquivos Brasileiros de Oftalmologia*, 76:26–28.

Anaïs Llorens, Athina Tzovara, Ludovic Bellier, Ilina Bhaya-Grossman, Aurélie Bidet-Caulet, William K Chang, Zachariah R Cross, Rosa Dominguez-Faus, Adeen Flinker, Yvonne Fonken, et al. 2021. Gender bias in academia: A lifetime problem that needs solutions. *Neuron*, 109(13):2047–2074.

Wen Lou and Jiangen He. 2015. Does author affiliation reputation affect uncitedness? *Proceedings of the AIST*, 52(1):1–4.

Alberto Martín-Martín, Enrique Orduña-Malea, Juan Ayllón, and Emilio Delgado. 2016. Back to the past: On the shoulders of an academic search engine giant. *Scientometrics*, 107(3):1477–1487.

Saif M. Mohammad. 2020a. Examining Citations of Natural Language Processing Literature. In *Proceedings of the 58th ACL*, pages 5199–5209, Online.

Saif M. Mohammad. 2020b. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.

Saif M. Mohammad. 2020c. NLP Scholar: A dataset for examining the state of NLP research. In *Proceedings of the 12th LREC*, pages 868–877, Marseille, France.

Saif M. Mohammad. 2021. Ethics sheets for AI tasks. In *Proceedings of the 60th ACL*, Dublin, Ireland.

Mathias Wullum Nielsen and Jens Peter Andersen. 2021. Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, 118(7):e2012208118.

Önder Nomaler, Koen Frenken, and Gaston Heimeriks. 2013. Do more distant collaborations have more citation impact? *Journal of Informetrics*, 7(4):966–971.

Raj K Pan, Alexander M Petersen, Fabio Pammolli, and Santo Fortunato. 2018. The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, 12(3):656–678.

Gianmarco Paris, Giulio De Leo, Paolo Menozzi, and Marino Gatto. 1998. Region-based citation bias in science. *Nature*, 396(6708):210–210.

Pietro Della Briotta Parolo, Raj Kumar Pan, Rumi Ghosh, Bernardo A. Huberman, Kimmo Kaski, and Santo Fortunato. 2015. Attention decay in science. *Journal of Informetrics*, 9(4):734–745.

Gerard Pasterkamp, Joris Rotmans, Dominique de Kleijn, and Cornelius Borst. 2007. Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics*, 70(1):153–165.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 EMNLP*, page 1371–1383, Abu Dhabi.

Natalie Schluter. 2018. The glass ceiling in nlp. In *Proceedings of the 2018 EMNLP*, pages 2793–2798.

Sei-Ching Joanna Sin. 2011. International coauthorship and citation impact: A bibliometric study of six LIS journals, 1980–2008. *Journal of the American Society for Information Science and Technology*, 62(9):1770–1783.

Marco Antonio Valenzuela-Escarcega, Vu A. Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

Alex Verstak, Anurag Acharya, Helder Suzuki, Sean Henderson, Mikhail Iakhiaev, Cliff Chiung Yu Lin, and Namit Shetty. 2014. On the shoulders of giants: The growing impact of older articles. *arXiv preprint arXiv:1411.0275*.

Adam Vogel and Dan Jurafsky. 2012. He said, she said: Gender in the ACL Anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, and Bela Gipp. 2022. D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research. *arXiv:2204.13384 [cs]*.

A Supplementary Statistics and Plots

In addition to the primary results presented in the main body of the paper, here, we describe included supplementary material in the form of additional statistics and plots.

A.1 Q3 Results Supplement: Distribution of $mAoC$ Over Years

Figure 8 shows the violin plots for distributions of $mAoC$ across various years. If a paper x was published in year t, then $mAoC(x)$ will be a data point for plotting the distribution for year t. The median $mAoC$ for a given year (marked with a white dot within the grey rectangle) reflects the recency of citations, with a lower median $mAoC$ indicating that papers published in that year have cited relatively recent papers.

The two halves of the grey rectangle on either side of the median correspond to the second and third quartiles. Observe that the third quartile is always longer (spread across more years than the second quartile. This shows that the rate at which papers are cited is higher in years before the median than in the years after the median. The violin plots indicate that the distributions have a single peak in each of the years considered.

Observe that the median $mAoC$ has an increasing trend from 1990 to 2014 (a trend towards citing more older papers) with the exception of a period between 1998 and 2004 when the median decreased. However, most notably, from 2014 onward the median $mAoC$ decreased markedly with every year. (The median $mAoC$ in 2021 is nearly 2.5 years less than that of 2014.)

Citation Bin	Full AoC 1965–2021	1990–1999	2000–09	2010–15
0	5559	457	1062	1453
1–9	26794	1813	5354	7090
10–49	21926	1714	5804	6272
50–99	4843	515	1517	1275
100–499	3860	496	1296	954
500–999	332	45	105	94
1000–1999	123	26	26	49
2000+	106	21	34	27

Table 3: Number of papers belonging to each citation bin on full AoC dataset, subset of papers published between 1990 to 2000, 2001 to 2010 and 2011 to 2016

The blue line in Figure 8 is the mean $mAoC$. The mean follows a similar trend as the median, with slight variations. In particular, it is consistently higher than the median, indicating that the data is skewed to the right, with a few papers having large $mAoC$ that significantly affect the mean.

A.2 Q6 Results Supplement: Pronounced Topics in the Cited Papers Across Year Intervals

We investigated the distribution of the most frequent unigrams and bigrams (ngrams) found in the title of cited papers, grouped by the publication years of the citing paper. Figures 9 and 10 show the unigrams and bigrams with notable changes in citation percentages across the chosen time intervals. A single star (*) indicates that the change in the ngram's percentage from the minimum interval value to maximum interval value is more than 1500% for unigrams and 3000% for bigrams. A double star (**) denotes that the ngram was not cited at all in at least one of the intervals.

A.3 Q7 Results Supplement: Variation of $mAoC$ and *CAD Index* Across Citation Count Bins

Table 3 shows the number of papers in each citation bin for different segments of papers. We can see that for all the time periods most of the papers have a citation count < 50.

Figures 11a and 11b show the variation of mean *mAoC* and *CAD Index* for subsets of papers published between 2001 to 2010 and 2011 to 2016, respectively. These two plots follow a similar pattern to Figure 7a on the full *AoC dataset*. The *CAD Index* decreases with increasing the citation bin and the mean $mAoC$ also varies inversely with the citation bin.

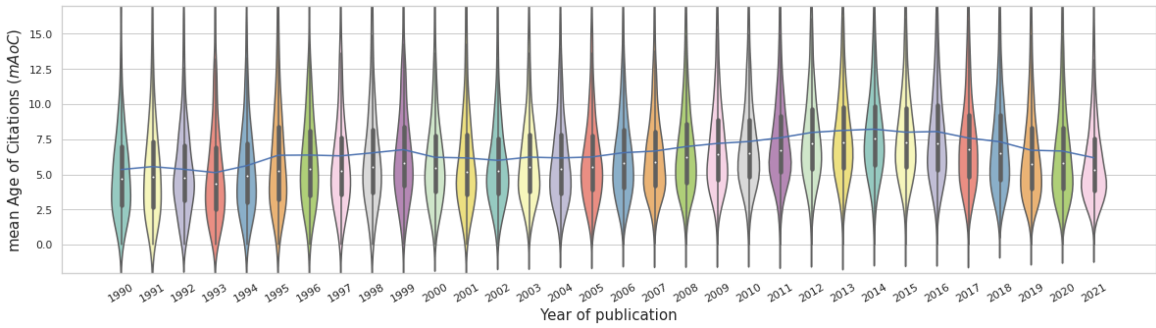


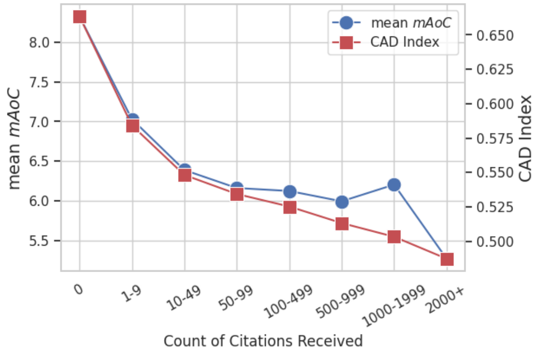
Figure 8: Distribution of $mAoC$ for papers published between 1990 and 2021.

	90-99	00-10	10-15	16-21
Annotation*	0.09	1.56	1.94	1.13
Answering*	0.14	1.41	0.66	1.76
Attention*	0.50	0.13	0.07	2.79
Bert**	0.00	0.00	0.00	2.31
Deep*	0.05	0.28	0.58	4.32
Embeddings**	0.00	0.00	0.15	2.23
Entity*	0.07	1.23	1.50	1.92
Mining*	0.02	0.95	1.65	0.96
Neural*	0.45	0.21	1.02	11.41
Pre*	0.04	0.05	0.09	1.92
Recurrent*	0.04	0.01	0.23	1.56
Sentiment*	0.00	0.37	2.26	2.45
Sequence*	0.08	0.44	0.54	2.09
Social*	0.12	0.14	0.80	1.60
Unification*	2.10	0.37	0.12	0.04
Web*	0.12	1.82	2.30	1.01

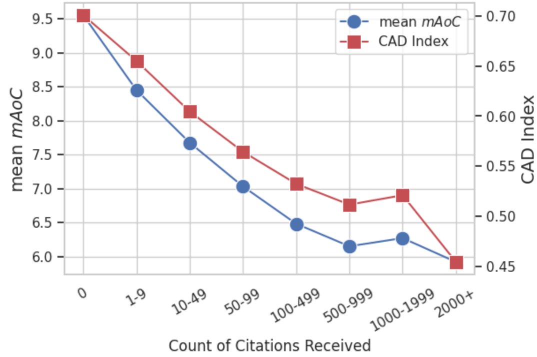
Figure 9: Unigram citation percentages of some notable terms found in the titles of cited papers across different time intervals. For example, "Neural" occurred in 11.41% of the titles of cited papers in the 2016–2021 interval.

	90-99	00-09	10-15	16-21
Adjoining Grammars*	0.635	0.274	0.084	0.020
Alignment Models**	0.000	0.513	0.382	0.061
Bert Pre**	0.000	0.000	0.000	0.912
Bidirectional Transformers**	0.000	0.000	0.000	0.917
Conditional Random**	0.000	0.619	0.684	0.284
Convolutional Neural**	0.000	0.000	0.084	0.817
Coreference Resolution*	0.022	0.321	0.765	0.428
Cross Lingual*	0.019	0.133	0.408	1.165
Deep Bidirectional**	0.000	0.000	0.002	0.916
Deep Learning**	0.000	0.002	0.064	0.679
Deep Neural**	0.000	0.002	0.119	0.467
Distributed Representations*	0.009	0.005	0.158	0.594
Domain Adaptation*	0.002	0.124	0.612	0.506
Entity Recognition*	0.007	0.654	0.652	0.954
Error Rate*	0.007	0.343	0.523	0.078
Global Vectors**	0.000	0.000	0.038	0.527
Jointly Learning**	0.009	0.000	0.031	0.461
Language Inference*	0.034	0.011	0.030	0.520
Long Short**	0.000	0.005	0.035	0.656
Multi Task*	0.004	0.004	0.046	0.621
Named Entity*	0.028	1.052	0.965	1.052
Neural Machine**	0.019	0.000	0.031	3.313
Open Source*	0.002	0.343	0.829	0.420
Parts Program*	0.464	0.042	0.005	0.001
Phrase Parser*	0.464	0.042	0.005	0.001
Pre Training**	0.000	0.000	0.006	1.551
Reading Comprehension*	0.009	0.058	0.050	0.637
Recurrent Neural*	0.015	0.006	0.177	1.144
Reinforcement Learning*	0.007	0.099	0.155	0.593
Relation Extraction**	0.000	0.199	0.463	0.736
Semantic Role**	0.000	0.462	0.445	0.352
Semi Supervised**	0.000	0.209	0.713	0.522
Sentiment Analysis**	0.000	0.136	1.136	1.289
Sentiment Classification**	0.000	0.109	0.421	0.454
Sequence Learning**	0.006	0.000	0.020	0.513
Shared Task**	0.000	0.373	0.869	1.159
Short Term*	0.013	0.012	0.045	0.669
Social Media**	0.000	0.005	0.258	0.970
Source Toolkit*	0.002	0.151	0.487	0.224
Stochastic Optimization*	0.007	0.006	0.048	0.652
Stochastic Parts*	0.464	0.042	0.005	0.001
Support Vector*	0.004	0.825	0.652	0.188
Term Memory*	0.011	0.009	0.042	0.662
Transfer Learning**	0.000	0.010	0.045	0.463
Tree Adjoining*	1.158	0.582	0.155	0.037
Unrestricted Text*	0.579	0.114	0.024	0.006
Vector Machines*	0.004	0.593	0.473	0.122
Word Embeddings**	0.000	0.000	0.076	1.070
Word Representations**	0.000	0.000	0.367	1.065

Figure 10: Bigram citation percentages of some notable terms found in the titles of cited papers across different time intervals. For example, "Neural Machine" occurred in 3.313% of the titles of cited papers in the 2016–2021 interval.

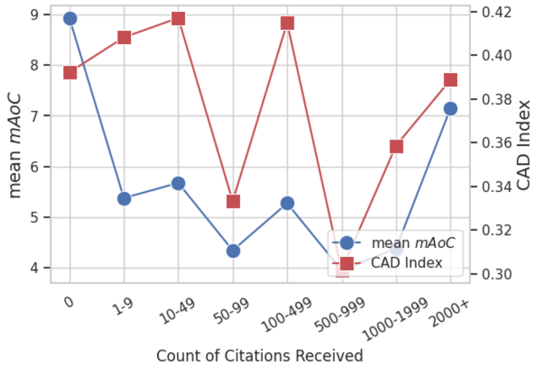


(a) Papers published between 2000 and 2010.

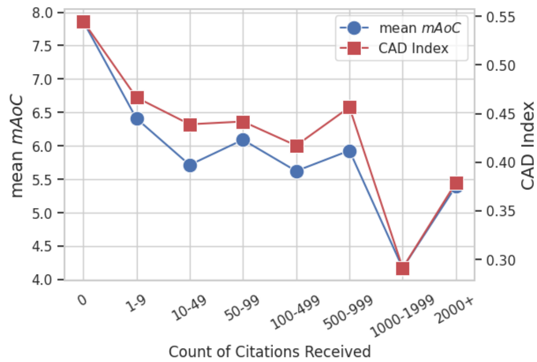


(b) Papers published between 2010 and 2016.

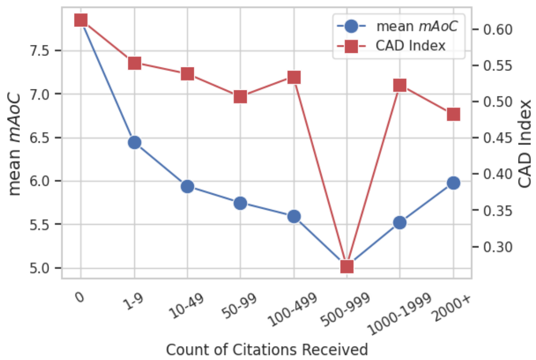
Figure 11: Variation of mean $mAoC$ and Citation Age Diversity (CAD) (shown on y-axis) for papers with different citation counts (shown on x-axis).



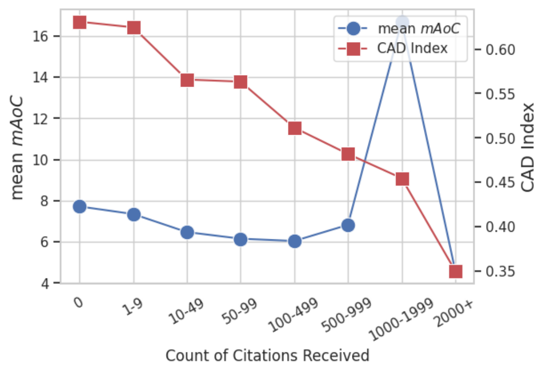
(a) Language inference.



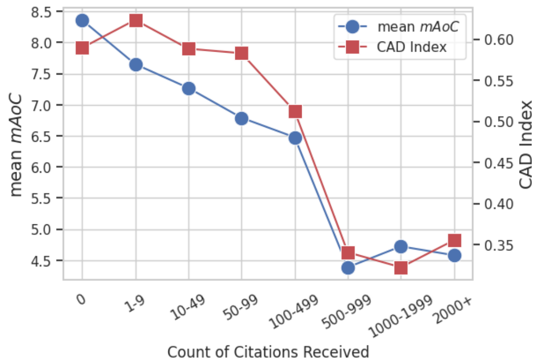
(b) Language model.



(c) Entity recognition.



(d) Sense disambiguation.



(e) Speech tagging.

Figure 12: Variation of mean $mAoC$ and *CAD Index* (shown on y-axis) for papers with different citation counts (shown on x-axis) for papers published between 1965 and 2021 across various *research topics*.

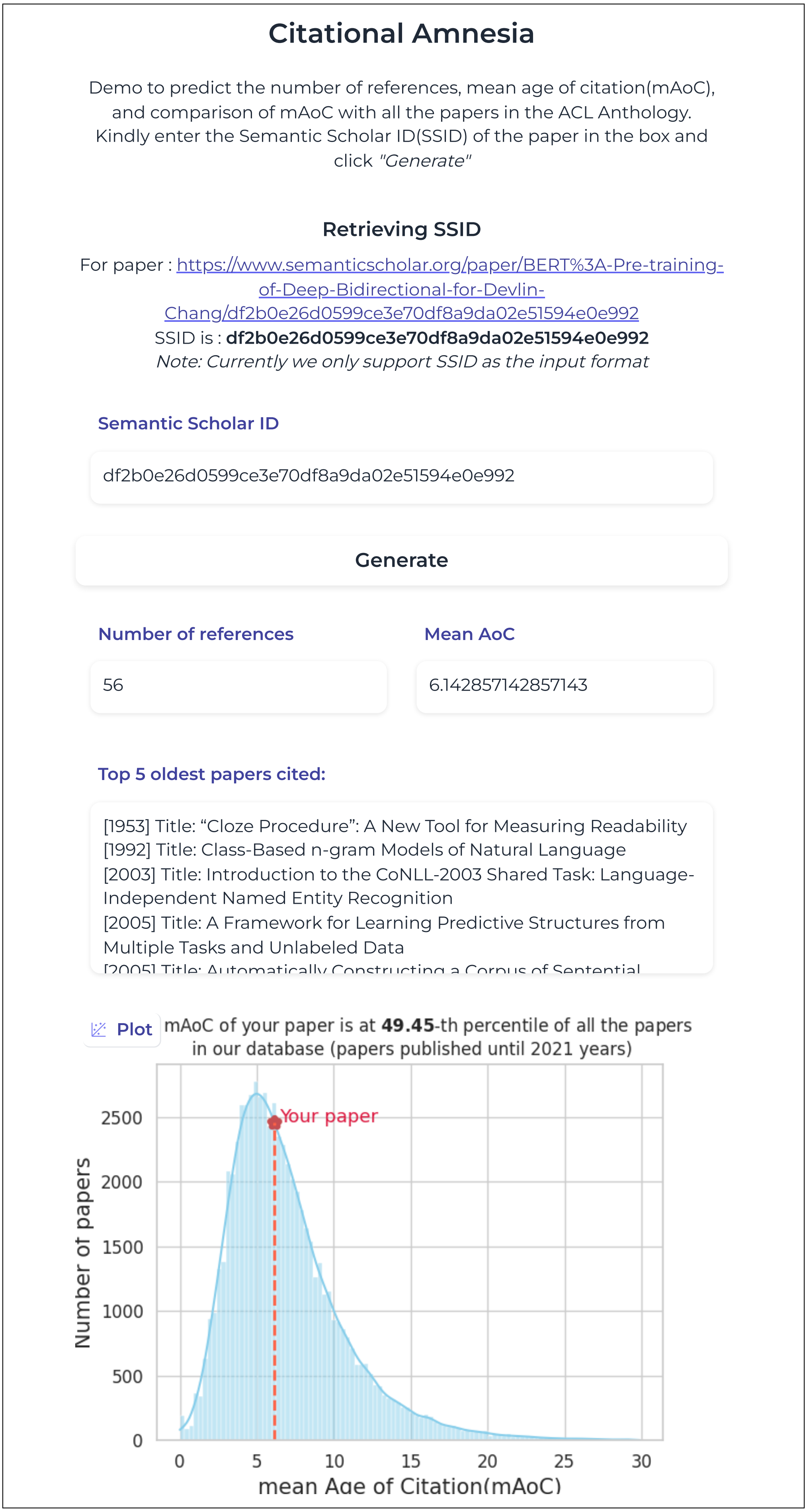


Figure 13: Screenshot of the online demo to compute Age-of-Citation metrics for a paper.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☐ A2. Did you discuss any potential risks of your work?
Not applicable. Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Not applicable. Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Not applicable. Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Not applicable. Left blank.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.