Internal and External Impacts of Natural Language Processing Papers

Yu Zhang
Texas A&M University
yuzhang@tamu.edu

Abstract

We investigate the impacts of NLP research published in top-tier conferences (*i.e.,* ACL, EMNLP, and NAACL) from 1979 to 2024. By analyzing citations from research articles and external sources such as patents, media, and policy documents, we examine how different NLP topics are consumed both within the academic community and by the broader public. Our findings reveal that language modeling has the widest internal and external influence, while linguistic foundations have lower impacts. We also observe that internal and external impacts generally align, but topics like ethics, bias, and fairness show significant attention in policy documents with much fewer academic citations. Additionally, external domains exhibit distinct preferences, with patents focusing on practical NLP applications and media and policy documents engaging more with the societal implications of NLP models.

1 Introduction

"*Is ACL an AI conference?*" Comments from opinion leaders[1][2] within the natural language processing (NLP) community definitely matter. Meanwhile, a complementary perspective is how the broader academic community and the public perceive and utilize papers from top-tier NLP conferences (*e.g.,* ACL, EMNLP, and NAACL). We can categorize the public use of NLP papers into two types: (1) *Internal Use*: NLP papers are cited by other research articles in not only NLP but also various other fields (Zhang et al., 2024), due to the ideas, methods, datasets, or findings they present (Jurgens et al., 2018; Cohan et al., 2019); and (2) *External Use*: Research can break out of the academic "ivory tower" (Cao et al., 2023) and diffuse into the technological, societal, and governmental

discussions (Yin et al., 2022). Specifically, NLP papers can be referenced by patents, media posts, policy documents, and text in other public channels. If we infer how public audiences define these NLP conferences based on which papers from the conferences are used by them, the question becomes: Papers of which topics (*e.g.,* cutting-edge AI models or linguistic foundations) are more extensively consumed internally and externally?

To answer this question, in this paper, we present a comprehensive scientometric study examining the internal and external impacts of NLP papers published in ACL, EMNLP, and NAACL between 1979 and 2024. We collect data from various sources (Priem et al., 2022; Marx and Fuegi, 2022; Adie and Roe, 2013; Szomszor and Adie, 2022) to obtain up-to-date references to NLP papers from internal (*i.e.,* research articles) and external (*i.e.,* patents, news/social media, and policy documents) domains. Based on the collected data, we calculate the average number of times that papers of a specific topic (*e.g.,* "*Language Modeling*" or "*Ethics, Bias, and Fairness*") are used in an internal/external domain, and employ it as an indicator to measure the impact of the topic within that domain.

Our analysis leads to the following key observations: First, among all research topics that NLP conferences are calling for, language modeling demonstrates the broadest internal and external impacts, whereas linguistic foundations, such as phonology, morphology, psycholinguistics, and pragmatics, exhibit relatively low influence. Second, the internal and external impacts of an NLP topic generally align, though there are outliers, such as "*Ethics, Bias, and Fairness*", which show notably high use in policy documents but are cited far less frequently in research articles. Third, different external domains favor different NLP topics, offering complementary insights into an NLP paper's internal impact. Patents tend to consume practical NLP applications, while media and policy documents are

---

more likely attracted by the behavior and societal influence of NLP models.

2 Data

We first describe how we collect data from multiple sources to measure the internal impact of NLP papers and their external uses across various domains, including patents, media, and policy documents.

NLP Papers: ACL Anthology. From the ACL Anthology[3], we extract all papers whose venue is marked as ACL, EMNLP, or NAACL (including main conference papers, findings papers, system demonstrations, industry track papers, *etc.*). We choose these three conferences because CSRankings designates them as "top conferences" for ranking purposes, and they are the top three in Google Scholar's "Computational Linguistics" category. In total, we obtain 24,821 papers published between 1979 and 2024 in these three conferences.

Internal Impact (Citation): OpenAlex. Because Microsoft Academic (Sinha et al., 2015) terminated its service at the end of 2021, to get the up-to-date number of citations for each paper, we link our extracted NLP papers to OpenAlex (Priem et al., 2022)[4]. In all, we successfully map 21,104 papers to the OpenAlex database and get their numbers of citations. We use $\mathcal{P}$ to denote the set of these mapped papers. We should mention that each paper $p \in \mathcal{P}$ may have multiple versions (*e.g.,* a preprint version and an NLP conference version). When calculating the internal or external impact of p, we merge all versions into a single record by summing their citation counts.

External Impact (Patent): Reliance on Science. To study references to NLP papers in patents, we follow previous studies (Yin et al., 2022; Cao et al., 2023) and utilize the Reliance on Science dataset (Marx and Fuegi, 2020, 2022)[5], which contains references from USPTO patents to OpenAlex papers. In total, we identify 20,218 links from patents to papers in $\mathcal{P}$.

External Impact (Media): Altmetric. To quantify how NLP papers are mentioned in media, we rely on the data provided by Altmetric (Adie and Roe, 2013)[6], which archives references from media, in-

cluding both news and social media (*i.e.,* Twitter, Facebook, Reddit, and blogs), to academic papers. In all, we obtain 18,586 media-to-paper links for the NLP publications in $\mathcal{P}$.

External Impact (Policy Document): Overton. To examine references to NLP papers in policy documents, we follow Yin et al. (2021) and exploit the Overton database (Szomszor and Adie, 2022)[7]. Overton defines "policy documents" as "research, briefs, reviews, or reports written with the goal of influencing or changing policy" from "governments, public bodies, IGOs, NGOs and think tanks". We obtain 1,223 links from policy documents to papers in $\mathcal{P}$.

3 Analysis

3.1 Internal and External Impacts by Topic

We adopt the 25 submission topics (excluding the special theme) outlined in the ACL 2025 Call for Papers[8] as our topic space for NLP papers, denoted as $\mathcal{T}$. For each NLP paper $p \in \mathcal{P}$, we use GPT-4o (Hurst et al., 2024) to predict its most relevant topic $t_p \in \mathcal{T}$. Human evaluation is then conducted to assess the quality of GPT-4o annotations. We find that GPT-4o gives reliable annotations, and the human evaluators share "substantial agreement" according to the Fleiss' kappa (Fleiss, 1971). One can refer to Appendix A for more details about the annotation process. Based on the annotations, we use $\mathcal{P}_t = \{p : p \in \mathcal{P} \text{ and } t_p = t\}$ to denote the set of NLP papers labeled with the topic t ($\forall t \in \mathcal{T}$).

To quantify the impact of NLP papers with a certain topic $t \in \mathcal{T}$ in a certain domain $d \in \{\texttt{Citation}, \texttt{Patent}, \texttt{Media}, \texttt{PolicyDocument}\}$, we consider the following metric:

$$\text{Impact}(t|d) = \frac{\sum_{p \in \mathcal{P}_t} \#\text{citation}(p|d)/|\mathcal{P}_t|}{\sum_{p \in \mathcal{P}} \#\text{citation}(p|d)/|\mathcal{P}|}.$$

Here, $\#\text{citation}(p|d)$ represents the "citation count" (in a broader sense) of paper p given domain d. To be specific, when $d = \texttt{Citation}$, it is the number of times that p is cited by other academic papers (from the entire OpenAlex database rather than ACL Anthology only); when $d \in \{\texttt{Patent}, \texttt{Media}, \texttt{PolicyDocument}\}$, it is the number of times that p is mentioned by patents, media posts, and policy documents, respectively, in our

---

Figure 1: Impact of NLP papers with a certain topic in a certain domain.

collected data. $\text{Impact}(t|d)$ is the average "citation count" of papers with topic t given domain d, normalized by the average "citation count" of all papers given domain d. Our proposed metric is inspired by Yin et al. (2022), with one key difference: we consider the actual number of times a paper p is cited, whereas Yin et al. (2022) binarize this value, focusing only on whether p is cited or not (*i.e.,* $\mathbf{1}_{\#\text{citation}(p|d)>0}$).

Figure 1 visualizes $\text{Impact}(t|d)$ across all topics and domains considered by us, from which we have the following key observations:

Observation 1: Papers on language modeling have a broad impact across all internal and external domains.

The topic "*Language Modeling*" enjoys the highest impact in the Citation, Patent, and Media domains, as well as the second highest impact in the PolicyDocument domain. It is the only over-represented topic (*i.e.,* $\text{Impact}(t|d) > 1$) across all domains, indicating the significant attention received from a broad range of public audiences on language models.

Observation 2: Papers on ethics, bias, and fairness are highly consumed in policy documents but have significantly lower usage in patents and academic papers.

The topic "*Ethics, Bias, and Fairness*" achieves the highest impact in PolicyDocument, exceeding the average in Media, while being under-represented in Citation and ranking the last in Patent. The topic "*Computational Social Science and Cultural Analytics*" exhibits a similar pattern of external impacts, but with a less pronounced contrast.

Observation 3: Linguistic foundations are relatively under-represented in all internal and external domains.

Topics that provide linguistic and theoretical underpinnings of NLP, such as "*Phonology, Morphology, and Word Segmentation*", "*Linguistic Theories, Cognitive Modeling and Psycholinguistics*", as well as "*Discourse and Pragmatics*", have $\text{Impact}(t|d) < 1$ across all domains.

Observation 4: Tasks and techniques benefiting real-world NLP applications have a larger impact in patents.

Besides "*Language Modeling*", topics strongly consumed in Patent include "*Machine Learning for NLP*", "*Information Retrieval and Text Mining*", "*Speech Processing and Spoken Language Understanding*", "*Syntax: Tagging, Chunking and Parsing*", as well as "*Machine Translation*". Many of these topics facilitate practical applications, such as search engines and speech recognition systems.

Observation 5: Analyses of the behavior and societal influence of NLP models see stronger adoption in media and policy discussions.

6 topics are over-represented in both Media and PolicyDocument. Apart from "*Language Model-*

490

Table 1: Pearson correlation coefficient between the impact of NLP papers in the internal domain (*i.e.,* `Citation`) and that in each external domain.

d	Patent	Media	PolicyDocument
$\text{Corr}(\mathbf{I}_{\texttt{Citation}}, \mathbf{I}_d)$	0.654	0.725	0.247 (0.599 if excluding *"Ethics, Bias, and Fairness"*)

Table 2: Hit rate of predicting the top-1% (internally) highly cited papers when using different external information.

External Domain(s) Considered	Hit Rate
$\emptyset$	1.00%
{Patent}	5.46%
{Media}	9.26%
{PolicyDocument}	18.29%
{Patent, Media}	26.72%
{Patent, PolicyDocument}	34.02%
{Media, PolicyDocument}	45.71%
{Patent, Media, PolicyDocument}	71.88%

ing", the other 5 topics—*"Resources and Evaluation"*, *"Interpretability and Analysis of Models for NLP"*, *"Ethics, Bias, and Fairness"*, *"Multilinguality and Language Diversity"*, as well as *"Computational Social Science and Cultural Analytics"*—all address the understanding and evaluation of NLP models.

3.2 Correlation between Internal and External Impacts

Our next study will draw the following conclusion:

Observation 6: NLP papers attracting attention from external domains are more likely to be internally impactful as well.

In Section 3.1, we have raised a few examples whose internal and external impacts are positively correlated. Now, we quantitatively validate this. Intuitively, each domain d can be represented as a vector $\mathbf{I}_d$, where each entry corresponds to the impact of a specific topic within d:

$$\mathbf{I}_d = [\text{Impact}(t|d)]_{t \in \mathcal{T}}.$$

Table 1 presents the Pearson correlation coefficient between $\mathbf{I}_{\texttt{Citation}}$ and each external $\mathbf{I}_d$ ($d \in \{\texttt{Patent}, \texttt{Media}, \texttt{PolicyDocument}\}$). We observe strong positive correlation between `Citation` and `Patent`/`Media` (*i.e.,* with the Pearson correlation coefficient greater than 0.5). The correlation between `Citation` and `PolicyDocument` is weaker, although still positive. That being said, if we remove *"Ethics, Bias, and Fairness"* (which is an evident outlier according to Figure 1) from consideration, $\text{Corr}(\mathbf{I}_{\texttt{Citation}}, \mathbf{I}_{\texttt{PolicyDocument}})$ rises above 0.5 as well.

To present more evidence of the positive correlation between internal and external impacts, we follow Yin et al. (2022) and conduct an experiment that leverages external use of NLP research to predict internally most cited papers. Let us consider the top-1% (internally) highly cited NLP papers in $\mathcal{P}$. If we do not exploit any external signal and randomly pick one paper p from $\mathcal{P}$, the expected probability that p is among the top-1% cited papers

(*a.k.a.,* the hit rate) should be 1%. In comparison, Table 2 shows the hit rate if we consider papers referenced at least once in a specific external domain d (*i.e.,* $\#\text{citation}(p|d) \geq 1$). Papers consumed by patents, media, and policy documents exhibit hit rates of 5.46%, 9.26%, and 18.29%, respectively, which are all large multiples of the 1% baseline.

Combining the results from Tables 1 and 2, we find good alignment between what the public from external domains consume and what is regarded as impactful by researchers themselves.

3.3 Complementarity of Different External Impacts

Although we have revealed the positive correlation between `Citation` and each external domain, the alignment between two external domains does not always hold. Qualitatively, we have discussed the different focal points of `Patent` and `Media`/`PolicyDocument` in Observations 4 and 5 in Section 3.1. Quantitatively, we actually have $\text{Corr}(\mathbf{I}_{\texttt{Patent}}, \mathbf{I}_{\texttt{PolicyDocument}}) = -0.140$, implying the complementarity, rather than substitutability, of impacts in patents and policy documents. Table 2 echos this observation. Indeed, for papers referenced in two external domains, the hit rate increases to 26.72%-45.71%, significantly higher than those when we consider a single external domain. A paper referenced in all three external domains is a top-1% cited paper in $\mathcal{P}$ at an astonishing 71.88 times the baseline rate. To summarize, we draw the following conclusion:

Observation 7: Different external domains may favor different types of NLP papers. Papers attracting attention from multiple external domains are more likely to be internally impactful than those attracting one domain only.

4 Conclusions

In this paper, we conduct a large-scale sciento-metric analysis on the internal and external impacts of NLP papers. We find that the broader academic community and the public's attention to NLP research is primarily driven by the demand for language modeling studies, with the exception of policy-related discussions, which show a greater interest in ethics, bias, and fairness. Although different external domains have varying interests in the specific topics of NLP research, there is an overall positive correlation between the public interests and the impact of papers within the academic "ivory tower". By examining how the broader academic community and the public perceive NLP papers, our analysis offers complementary insights into whether ACL can be considered an AI conference.

Limitations

This work has the following limitations: First, although the three external domains we consider are all important public spaces, they do not encompass all channels that NLP research may impact. Even within these three domains, NLP papers may be consumed through channels not captured by our collected data. Second, we utilize GPT-4o to predict the most relevant topic to each NLP paper. On top of that, we conduct human evaluation for quality assessment. However, there are still chances where GPT-4o's annotations may not be accurate enough, potentially affecting the subsequent analysis. Third, we currently lack results (and data) to explain why there is a positive correlation between internal and external impacts. It might be a possible explanation that either researchers or the public are following the other's preference regarding NLP. Yin et al. (2022) propose another potential hypothesis that each external domain typically involves an "intermediary" to engage with science, such as journalists in the media, inventors in the patent field, and policy experts in the government. They leverage their expertise to select scientific results and introduce them into their respective domains. Finally, since our collected data from certain domains do not indicate when an NLP paper was cited, we are unable to perform a systematic temporal analysis. We believe that, with appropriate data support, all of the limitations mentioned above represent highly promising future directions.

Acknowledgments

References

Euan Adie and William Roe. 2013. Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1):11–17.

Hancheng Cao, Yujie Lu, Yuting Deng, Daniel Mcfarland, and Michael S Bernstein. 2023. Breaking out of the ivory tower: A large-scale analysis of patent citations to hci research. In *CHI'23*, pages 760:1–760:24.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL-HLT'19*, pages 3586–3596.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *ACL'20*, pages 2270–2282.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *TACL*, 6:391–406.

Matt Marx and Aaron Fuegi. 2020. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594.

Matt Marx and Aaron Fuegi. 2022. Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations. *Journal of Economics & Management Strategy*, 31(2):369–392.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW'15*, pages 243–246.

Martin Szomszor and Euan Adie. 2022. Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, 3(3):624–650.

Yian Yin, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin F Jones. 2022. Public use and public funding of science. *Nature Human Behaviour*, 6(10):1344–1350.

Yian Yin, Jian Gao, Benjamin F Jones, and Dashun Wang. 2021. Coevolution of policy and science during the pandemic. *Science*, 371(6525):128–130.

Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. In *EMNLP'24*, pages 8783–8817.

A Topic Prediction of NLP Papers

A.1 Topic Annotation using GPT-4o

We adopt the following instruction prompt to guide GPT-4o in performing topic annotation.

```
Instruction:  You will be given the title
and abstract of a natural language processing
paper, as well as a list of candidate topics.
Select the most relevant topic to the given
paper from the list.

Paper title: [title; e.g., BERT: Pre-training
of  Deep  Bidirectional  Transformers  for
Language Understanding]

Paper abstract: [abstract; e.g., We introduce
a new language representation model called
BERT, which stands for ...]

Candidate labels:  "Computational  Social
Science and Cultural Analytics", "Dialogue
and  Interactive  Systems", "Discourse  and
Pragmatics", "Ethics, Bias, and Fairness", ...
(25 topics)
```

We set the temperature as 1.0 (the default setting for GPT-4o in OpenAI's API). In rare cases, the output of GPT-4o does not exactly match any of the candidate labels. To tackle this, we use a scientific pretrained language model, SPECTER (Cohan et al., 2020), to encode the output and each candidate label's name; then, we consider the nearest neighbor of the output in the SPECTER embedding space as the predicted label.

A.2 Human Evaluation

To assess the quality of annotations produced by GPT-4o, human evaluation is then conducted. We recruit 3 evaluators, all of whom have rich NLP research experiences. We randomly selected 100 samples for evaluation. For each submission, we ask each annotator to judge the correctness of the label predicted by GPT-4o. The scoring scale is 2 (the predicted label is the most relevant label to the paper), 1 (the predicted label is relevant to the paper, although there is a better option), and 0 (the predicted label is irrelevant to the paper).

According to human evaluation, GPT-4o gets an average score of 1.53 across the three evaluators. Notably, none of the 100 predictions gets 0

(*i.e.,* irrelevant), indicating the reliability of GPT-4o annotations. We also calculate the Fleiss' kappa (Fleiss, 1971), which is 0.62 and implies "substantial agreement" among the evaluators.

B Impact of NLP Papers in GitHub

As noted in the Limitations section, the impact of an NLP paper may not be limited to textual citations alone. As a complementary analysis, in this section, we examine the influence of NLP papers in the GitHub domain. Specifically, we analyze the average number of forks for GitHub repositories associated with each NLP topic. It is important to note that we consider this measure of impact to be a mixture of both internal and external influences, as forking a code repository may be done by other researchers for follow-up research, or by practitioners aiming to deploy the work in real-world applications.

B.1 Data and Metric

We refer to Papers With Code[9] to establish the correspondence between our collected NLP papers and their associated GitHub repositories. 7,113 (out of 21,104) NLP papers are mapped to at least one repository in this way. Then, we used the GitHub API[10] to retrieve the number of forks for each corresponding repository. If a paper p is associated with multiple repositories, its fork count $\#\mathrm{fork}(p)$ is computed as the sum of forks across all associated repositories. Analogous to defining impact through the citation count, we define the impact of NLP papers in GitHub based on the fork count as follows:

$$\mathrm{Impact}(t|\texttt{GitHub}) = \frac{\sum_{p\in\mathcal{P}_t} \#\mathrm{fork}(p)/|\mathcal{P}_t|}{\sum_{p\in\mathcal{P}} \#\mathrm{fork}(p)/|\mathcal{P}|}.$$

Here, we only include NLP papers in $\mathcal{P}$ and $\mathcal{P}_t$ that can be matched to at least one GitHub repository.

B.2 Analysis

Figure A1 shows $\mathrm{Impact}(t|\texttt{GitHub})$ across all topics. Once again, we observe that "*Language Modeling*" exhibits the highest impact by a significant margin. Other topics with relatively high impact include "*Speech Processing and Spoken Language Understanding*", "*Machine Learning for NLP*",

---

[9]https://production-media.paperswithcode.com/about/links-between-papers-and-code.json.gz, accessed on March 5, 2025.

[10]https://api.github.com/repos, accessed on March 5, 2025.
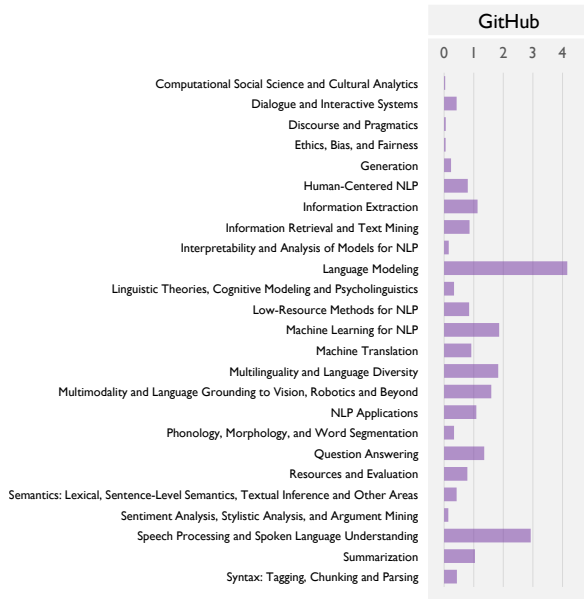
Figure A1: Impact of NLP papers with a certain topic in the GitHub domain.

Table A1: Pearson correlation coefficient between the impact of NLP papers in GitHub and that in other domains.

d	Citation	Patent	Media	Policy Document
$\mathrm{Corr}(\mathbf{I}_{\texttt{GitHub}}, \mathbf{I}_d)$	0.586	0.633	0.531	0.009

"*Multilinguality and Language Diversity*", as well as "*Multimodality and Language Grounding to Vision, Robotics and Beyond*". It is evident that topics related to practical NLP applications tend to achieve higher impact in GitHub. In this regard, GitHub is similar to Patent, which also aligns with our intuition. By contrast, linguistic foundations and "*Ethics, Bias, and Fairness*" exhibit significantly lower impact.

Table A1 demonstrates the Pearson correlation coefficient between the impact of NLP papers in GitHub and that in other domains. We observe a clear positive correlation between GitHub and Citation, Patent, and Media, with the strongest correlation observed with Patent, echoing our qualitative discussion above. Meanwhile, the near-zero correlation between GitHub and PolicyDocument reflects their distinct emphases.