The Nature of NLP: Analyzing Contributions in NLP Papers

Aniket Pramanick[1], Yufang Hou[2,3], Saif M. Mohammad[4], Iryna Gurevych[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technische Universität Darmstadt
[2]IT:U Interdisciplinary Transformation University Austria
[3]IBM Research Europe - Ireland
[4]National Research Council Canada
www.ukp.tu-darmstadt.de, yufang.hou@it-u.at, saif.mohammad@nrc-cnrc.gc.ca

Abstract

Natural Language Processing (NLP) is an established and dynamic field. Despite this, what constitutes NLP research remains debated. In this work, we address the question by quantitatively examining NLP research papers. We propose a taxonomy of research contributions and introduce `NLPContributions`, a dataset of nearly $2k$ NLP research paper abstracts, carefully annotated to identify scientific contributions and classify their types according to this taxonomy. We also introduce a novel task of automatically identifying contribution statements and classifying their types from research papers. We present experimental results for this task and apply our model to $\sim 29k$ NLP research papers to analyze their contributions, aiding in the understanding of the nature of NLP research. We show that NLP research has taken a winding path — with the focus on language and human-centric studies being prominent in the 1970s and 80s, tapering off in the 1990s and 2000s, and starting to rise again since the late 2010s. Alongside this revival, we observe a steady rise in dataset and methodological contributions since the 1990s, such that today, on average, individual NLP papers contribute in more ways than ever before. Our dataset and analyses offer a powerful lens for tracing research trends and offer potential for generating informed, data-driven literature surveys. [1]

1 Introduction

Categorizing research by scientific discipline has several benefits, including bringing together scientists to make progress in a cohesive area of interest. While there is often some broad description of what a scientific discipline constitutes, the nature of a discipline is dynamic and multifaceted, and it can change with time. NLP is a particularly interesting discipline in this regard, not only because of its interdisciplinary nature, drawing on ideas and techniques from computer science, linguistics, social science, etc., but also because fundamental questions such as '*what is NLP research?*' can be contentious. Is it the study and development of algorithms that give machines the ability to respond to and generate language? Is it the study of natural language using computational approaches? Does it cover all research at the intersection of computation and language? Or is it something more narrow?

A compelling way to answer '*what is NLP research?*' is to examine the papers published in NLP conferences and journals. After all, the body of current published research is the best indicator of what a field is and the nature of the field, regardless of how that field may have once been defined or understood.

A key window into the nature of a particular research project is how the authors articulate their contributions. *Contributions* are new scientific achievements attributed to the authors. Roughly speaking, scientific contributions are of two types: i) those that add to human knowledge, e.g., discovering the structure of DNA; and ii) those that create new and useful artifacts, e.g., a general-purpose chat system such as ChatGPT. When authors present their work in scientific papers, they describe their contributions to the research community. We define *contribution statements* as descriptions of these contributions.

In this paper, we propose that automatically extracting, categorizing, and quantitatively analyzing the contribution statements in the research papers of a field provides key insights into the nature of the field. Additionally, such an effort enables historical (longitudinal) analyses of the field (Shapere, 1964) and can help researchers identify emerging trends and stay current amid the rapid proliferation of scientific publications.

[1]Code and data are available at: https://github.com/UKPLab/acl25-nlp-contributions

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 25169–25191
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

We explore this idea concretely and empirically by examining $28,937$ NLP papers published between 1974 and 2024. Specifically, we:

1. Introduce a *taxonomy* of contribution types common in NLP papers (§ 3.1).
2. Create a *dataset* `NLPContributions` comprising of $1,995$ NLP research papers with manually annotated contribution statements and contribution types from their abstracts (§ 3.2).
3. Propose a *novel task* to automatically extract and classify contribution statements into contribution types from NLP papers (§ 4.1).
4. Finally, ask (and answer) some preliminary questions on the *nature of NLP research* and how it has changed over the years (§ 5).

2 Related Work

NLP Scientometrics. The study of trends in scientific research gained attention following the seminal work by Hall et al. (2008). This line of work, broadly known as "scientometrics", focuses on the quantitative analysis of scholarly literature. NLP scientometrics has gained interest in recent years, as researchers strive to understand the growing landscape of NLP research and its evolution (Mingers and Leydesdorff, 2015; Chen and Song, 2019). One prominent research direction in NLP scientometrics is the analysis of metadata (Mohammad, 2020b), employing bibliometric techniques (Wahle et al., 2022), co-authorship analysis (Mohammad, 2020a), and topic modeling (Jurgens et al., 2018a) to gain insights into the dynamics of the field, identifying research trajectories. Text mining and deep learning techniques have also been utilized in NLP scientometrics to extract information from research papers, create structured datasets, and enable detailed analyses of the interactions among topics and their evolution (Prabhakaran et al., 2016; Tan et al., 2017; Salloum et al., 2017; Yang and Li, 2018; Prabhakaran et al., 2016; Hou et al., 2019; Pramanick et al., 2023; Şahinuç et al., 2024). Our study delves deeper into NLP scientometrics by analyzing the content in research paper.

Citation Intent Analysis. A large body of research has focused on understanding the purpose behind citations and developing classification systems for them (Stevens and Giuliano, 1965; Oppenheim and Renn, 1978; Garzone and Mercer, 2000; Teufel et al., 2006; Dong and Schäfer, 2011; Jurgens et al., 2018b). While citation intents signal the purpose of a citation, such as providing background information or making comparisons, contributions differ as they present novel additions that a research paper introduces to its field. Citation intents may indirectly reflect a paper's contributions from the perspective of citing papers; our focus is on contributions as articulated by the authors themselves within their own work.

NLP Contribution Graph. D'Souza and Auer (2020) introduced an annotation scheme to identify *information units* in scientific documents related to contributions, focusing on artifacts like models, datasets, or baselines linked to a pre-defined set of NLP tasks. D'Souza et al. (2021) employed this annotation scheme to construct a knowledge graph that connects these artifact information units across NLP tasks. Deep learning methods have been applied to automate the extraction of this information units (Gupta et al., 2021, 2024). It is important to note that these units are not necessarily novel contributions from the papers they are extracted from. Unlike these efforts, our work broadens the scope by identifying and categorizing contribution statements from research papers without limiting them to specific NLP tasks. Additionally, our approach encompasses contributions that expand knowledge as well as introduce new artifacts.

Claim and Opinion Summarization. Researchers have explored automated methods to study diverse perspectives of claims (Chen et al., 2019). This includes the growing interest in key-point analysis (Bar-Haim et al., 2021; Friedman et al., 2021). Neural network and graph-based approaches have been proposed for claim summarization from newspaper reports and online discussions (Zhao et al., 2022; Inácio and Pardo, 2021). Some research has focused on extracting claims from scientific papers (Achakulvisut et al., 2019; Al Khatib et al., 2021; Sosa et al., 2023). While claims in research papers provide declarations to support hypotheses or research questions, contributions present new elements (knowledge and artifacts) that a paper introduces to its field. In this work, we explore methods to extract and analyze contribution statements from NLP research papers.

Type	Sub-type	Description	Example
Knowledge	k-dataset	Describes new knowledge about datasets, such as their new properties or characteristics.	"Furthermore, our thorough analysis demonstrates the average distance between aspect and opinion words are shortened by at least 19% on the standard SemEval Restaurant14 dataset." – Zhou et al. (2021)
	k-language	Presents new knowledge about language, such as a new property or characteristic of language.	"In modern Chinese articles or conversations, it is very popular to involve a few English words, especially in emails and Internet literature." – Zhao et al. (2012)
	k-method	Describes new knowledge or analysis about NLP models or methods (which predominantly draw from Machine Learning).	"Different generative processes identify specific failure modes of the underlying model." – Deng et al. (2022)
	k-people	Presents new knowledge about people, humankind, society, or human civilization.	"Combating the outcomes of this infodemic is not only a question of identifying false claims, but also reasoning about the decisions individuals make." – Pacheco et al. (2022)
	k-task	Describes new knowledge about NLP tasks.	"We show that these bilingual features outperform the monolingual features used in prior work for the task of classifying translation direction." – Eetemadi and Toutanova (2014)
Artifact	a-dataset	Introduces a new NLP dataset (i.e., textual resources such as corpora or lexicon).	"We present a new corpus of Weibo messages annotated for both name and nominal mentions." – Peng and Dredze (2015)
	a-method	Introduces or proposes a new or novel NLP method or model (primarily to solve NLP task(s)).	"The paper also describes a novel method, EXEMPLAR, which adapts ideas from SRL to less costly NLP machinery, resulting in substantial gains both in efficiency and effectiveness, over binary and n-ary relation extraction tasks." – Mesquita et al. (2013)
	a-task	Introduces or proposes a new or novel NLP task (i.e., well-defined NLP problem).	"We formulate a task that represents a hybrid of slot-filling information extraction and named entity recognition and annotate data from four different forums." – Durrett et al. (2017)

Table 1: Overview of the taxonomy for NLP research contributions with examples for each contribution type.

3 `NLPContributions`: A Corpus of Contribution Statements

We developed a taxonomy of various types of contributions found in NLP research papers. Using this taxonomy, we annotated contribution statements from the abstracts of NLP research papers. We chose the abstracts as our corpus for annotation because abstracts are uniquely positioned at the beginning of papers and typically contain contribution statements. Moreover, abstracts efficiently summarize the paper, providing the context for understanding contributions, making them particularly suitable text segments to focus on for contribution annotation. Annotating entire papers would substantially escalate annotation efforts. Thus working with abstracts was a more efficient and effective option (Teufel et al., 1999).

3.1 Taxonomy of Contributions

In NLP research, contributions can broadly be divided into *two main types*. We call the first type *artifacts*, which encompasses the development of new or novel resources. NLP research heavily utilizes tools from machine learning, which relies on resources such as new methods or models, datasets - and the novel tasks they enable - all of which are recognized as significant contributions. Consequently, we categorize artifact contributions into three sub-types: *new methods* (a-methods), *new datasets* (a-datasets), and *new tasks* (a-tasks), each distinguished by the specific resource it brings to the field.

We term the second category as *knowledge* contributions that enrich the field with new insights or knowledge. Depending on what these contributions add knowledge to, we further categorize them into five sub-types: *knowledge about method* (k-method), *knowledge about dataset* (k-dataset), *knowledge about task* (k-task), *knowledge about language* (k-language), and *knowledge about people* (k-people). This sub-categorization also mirrors the important elements in NLP research.

In Table 1, we provide a detailed description of each type and subtype, with examples of contribution statements from research papers. While we acknowledge that alternative taxonomies may also be possible, we note that our proposed taxonomy is in line with the ACL'23 call for papers[2], which seeks submissions either that conduct analysis (thereby adding *knowledge*) or introduce new resources (thereby adding *artifacts*).

3.2 Curation

Data Preparation. We compile a corpus of abstracts from 1,995 papers published under ACL Anthology using the S2ORC (Lo et al., 2020), a large collection of papers released to support research. We randomly selected these papers, guaranteeing a selection of at least five papers from each year between 1974 and February 2024. The selected papers were published in journals and conferences affiliated with "ACL Events", while those from workshops were excluded. Additionally, we retrieve the metadata (i.e., unique id, title, authors, publication venue, and date) for each selected paper from `anthology.bib`.[3]

Annotation. The main annotator is one of the authors of this paper, who has six years of experience in NLP research. Additionally, a PhD student with four years of research experience

[2] https://tinyurl.com/mpdkmzkj
[3] https://aclanthology.org/anthology.bib.gz

Typ.	Sub-typ.	κ
Knowledge	k-dataset	0.70
	k-language	0.69
	k-method	0.71
	k-people	0.67
	k-task	0.70
Artifact	a-dataset	0.76
	a-method	0.71
	a-task	0.73
Overall Agreement		0.71

Table 2: Inter-annotator Agreement

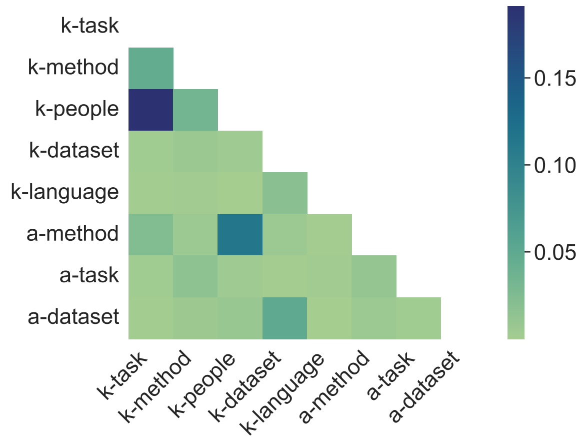


Figure 1: Pointwise mutual information (PMI) between contribution types shows the co-occurrence of multiple contribution types within the same contribution statements.

took part in annotation. We develop an annotation scheme to identify and classify the contribution statements in NLP research papers. We use ontology-oriented annotation guidelines (refer to Appendix C for details) following Liakata et al. (2010). Regular meetings were conducted between the annotators to refine the guidelines as necessary (Klie et al., 2024).

Both annotators annotated 100 papers, ensuring representation from each decade between 1980 and 2024. We assess annotator agreements on these 100 papers. Subsequently, the senior annotator proceeded to annotate the abstracts of the additional 1,895 papers, adhering to the guidelines. Finally, the senior authors of this paper reviewed the dataset, particularly focusing on samples with disagreements, as a final check to ensure its quality. We name the corpus of 1,995 annotated papers `NLPContributions`.

Agreement. We measure the inter-annotator agreement (IAA) by comparing the contribution statements from the 100 aforementioned papers annotated by the two annotators under the same contribution labels. All annotations were conducted in Label Studio (Tkachenko et al., 2020-2022). Table 2 shows an average Fleiss' κ of 0.71, comparable to similar works on scholarly documents (Yang and Li, 2018; Hou et al., 2021; Lauscher et al., 2022). Further, we observe the error bounds of κ between 0.60 (lower bound) and 0.82 (upper bound) with 95% confidence level ($p < 0.05$).

3.3 Data Statistics

We highlight three aspects of our dataset: first, it includes abstracts and metadata of $1,995$ papers from "ACL Events" in the ACL Anthology. On average, each abstract comprises 5.42 sentences, with 2.95 sentences annotated as contribution statements, resulting in a total of 5,890 annotated contribution statements. Second, we illustrate the distribution of labels across these statements in Table 3. Lastly, we note that 57.6% of the contribution statements received multiple labels. Figure 1 shows the co-occurrence of different contribution types within the same contribution statements, measured using pointwise mutual information (PMI) scores. Overall, the PMI values between any pair of contribution types are low, indicating low co-occurrence. However, *k-people* and *k-task* appear together more frequently than others, possibly because authors often explain how NLP tasks yield insights into humans or society. We divide our dataset into train-val-test (70-15-15) split at the paper level to maintain consistency in our experiments and prevent information leakage.

4 Automatically Identifying Contribution Statements and Contribution Types

We introduce the novel task of automatically detecting and categorizing contribution statements from NLP research papers. We use `NLPContributions` and benchmark multiple models to evaluate their performance on this task.

4.1 Task Definition

The task involves two steps: detecting contribution statements and subsequently categorizing them by type. We model it end-to-end as a multi-label extension of multi-class classification, where, given

Typ.	Sub-typ.	Prop. (%)
Knowledge	k-dataset	5.1
	k-language	4.0
	k-method	12.6
	k-people	9.2
	k-task	36.1
Artifact	a-dataset	2.2
	a-method	27.2
	a-task	3.6

Table 3: Occurrence percentages of different contribution types in contribution statements from paper abstracts in `NLPContributions`.

Setting	Model	P	R	F1
	Random	0.19	0.17	0.17
Finetuning	BERT	0.31	0.50	0.38
	BiomedBERT	0.64	0.59	0.60
	SciBERT	**0.81**	**0.80**	**0.80**
	Flan-T5	0.79	0.78	0.78
	RoBERTa	0.33	0.50	0.40
Prompting	GPT-3.5-Turbo	0.75	0.71	0.73
	GPT-4-Turbo	0.80	**0.80**	0.80
	LLaMA-3	0.60	0.56	0.53

Table 4: Performance of different models for `contribution` statement classification.

a statement, the objective is to assign it types and subtypes if it qualifies as a contribution; otherwise, assign Null. Formally, given a statement S, and a set of n labels $L = [l_1, l_2, ..., l_n]$, the task is to predict a subset of these labels $Y = [y_1, y_2, ..., y_n]$ associated with S, where $y_j = 1$, if l_j is associated with S, and 0 otherwise.

4.2 Methods

In our study, we explore two methods. The first method involves utilizing pre-trained language models (PLMs) that are further fine-tuned using the training split of `NLPContributions`. Second, we use large language models (LLMs) and utilize prompting techniques for our task (refer to Appendix D for the prompting details). We use the *binary relevance* (Read et al., 2009) for the task, treating each label as an independent binary classification problem. This avoids overfitting by not depending on previous label combinations and allows for flexible modifications to the label set without affecting other parts of the model.

PLMs. We start our study with BERT (Devlin et al., 2019) and RoBERTa (Zhuang et al., 2021), which are general-purpose pre-trained language models. Moving further, we use BiomedBERT (Gu et al., 2021) and SciBERT (Beltagy et al., 2019), which are pre-trained on scientific texts. Additionally, we experiment with Flan-T5 (Chung et al., 2024), which is pre-trained over a collection of 1,836 fine-tuning tasks. We also implement a random baseline that assigns labels to sentences with a uniform random probability.

LLMs. Addressing the task through prompting, we use GPT-3.5-Turbo and GPT-4-Turbo (OpenAI, 2023), which are instruction-following large language models fine-tuned with reinforcement learning from human feedback (RLHF). Additionally, we use the open-source LLaMA-3-8B model (Meta, 2023), which has been trained on over 15 trillion tokens gathered from publicly available domains.

4.3 Training and Evaluation

During fine-tuning the pre-trained language models, we use a grid search across various epochs $e \in \{1, 2, 3, 4, 5\}$ and learning rates $lr \in \{1 \cdot 10^{-4}, 5 \cdot 10^{-4}, 1 \cdot 10^{-5}\}$, using a batch size of 32. For prompting, we start with a zero-shot setting and gradually progress to a five-shot, respecting the context length limitations of the models. We repeat each experiment three times and observe the variance < 0.02 for all of the models.

Following Uma et al. (2022), for multi-label classification, we use label-based evaluation (macro-averaged precision, recall, and F1-score), which assesses performance on a per-label basis and then aggregates scores across all labels. We avoid label-set-based evaluation, also known as the exact match measure, because it does not effectively account for the sparsity characteristic of multi-labeling, often missing nuanced label variations.

4.4 Results and Discussion

Table 4 shows the results. We observe that SciBERT outperformed other fine-tuned pre-trained language models, likely due to its pre-training on a collection of scholarly documents. Additionally, we note that GPT-4-Turbo's performance is on par with fine-tuned SciBERT. Hence, for environmental sustainability and cost-efficiency, we have chosen to use SciBERT for subsequent analyses. Note that we tested the LLMs with five different prompt variants and recorded the most effective ones in Table 11 (Appendix D). We also found that LLM performance decreased when prompts included titles or entire abstracts, likely because titles may not accurately represent contributions, and LLMs are optimized for data with fixed context lengths. All reported results from experiments using pre-trained and large language models are statistically significant (McNemar's $p < 0.001$).

4.5 NLPContributions-Auto: A Corpus of Auto-Identified Contribution Statements

We applied the fine-tuned SciBERT model to the sentences from the abstracts of papers in the ACL Anthology and classified them according to the predefined taxonomy. We call this corpus NLPContributions-Auto. This corpus can be used for diverse research purposes on NLP papers, including efficient semantic searches and key point analysis, among others. In the following section, we explore various NLP research trends using this corpus.

Specifically, we used S2ORC to gather the abstracts of $28,937$ papers published from conferences or journals falling within the "ACL Events" category between 1974 and February 2024 (details in Appendix B). We collected the metadata of these papers from the "anthology.bib". However, it is important to note here that while NLP papers are published outside of the ACL Anthology, it remains the largest single-source collection of NLP papers. Additionally, the Anthology's strict peer review process ensures high quality, making it a reliable source for our study.

5 Analyzing the Nature of NLP

We study the nature of NLP by examining the trends and evolution in research contributions (§ 5.1), the influence of publication venues (§ 5.2), and their impact on citation patterns (§ 5.3).

5.1 Evolving Contributions in NLP Research

Q1. How do the various types of contributions shape the landscape of NLP research?

To study the breadth of contributions in NLP research, we examine the percentage of contribution statements associated with each contribution type and subtype in NLPContributions-Auto.

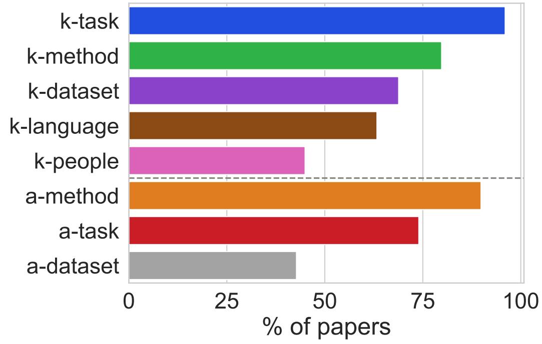


Figure 2: Occurrence percentages of different contribution types associated with contribution statements in paper abstracts from NLPContributions-Auto.

Results. Figure 2 shows the distribution of different types of contributions in the abstracts of papers. Overall, we observe that:

a. Contributions of type *Knowledge about people* (44.9%) and *Knowledge about language* (61.2%) are relatively few compared to contributions of type *k-task* (89.8%) and *k-method* (78.7%).

b. Within the *artifact* type contributions, $\sim$89% of the papers introduce new methods (the highest), followed by tasks ($\sim$75%) and, finally, datasets ($\sim$45%).

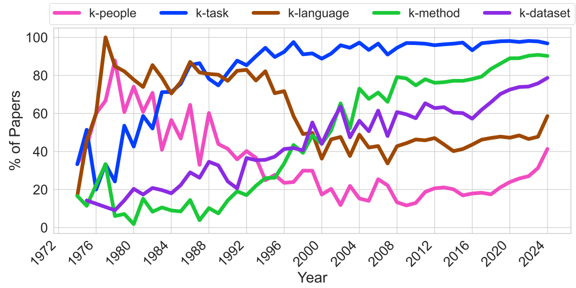
Discussion. While some researchers suggest that NLP research is more relevant to people or society (Clark and Schober, 1992), our observation reveals a significant focus of NLP research on knowledge about tasks and methods, particularly involving machine learning. Also, our findings resonate with those of Pramanick et al. (2023), who noted through causal entity analysis that new methods and tasks have been drivers of NLP research.

Q2. How has the nature of NLP evolved over the years?

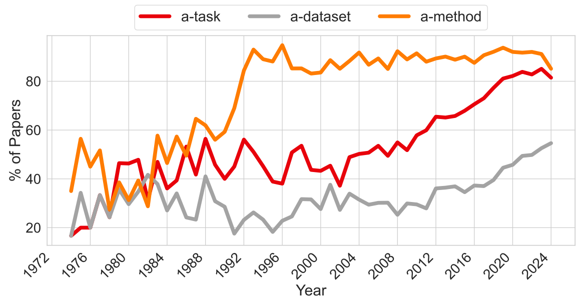
To study how contribution types have evolved in NLP research, we calculate, for each year, the percentage of papers that include at least one contribution statement corresponding to each contribution type.

Results. Figure 3a shows the following knowledge contribution trends:

a. Shift from language focus. In the seventies and eighties, NLP focused on language, evidenced by significant contributions toward knowledge of language. However, from the early nineties to 2020, there was a dramatic decline in k-language



(a) Knowledge contributions



(b) Artifact contributions

Figure 3: Percentage of papers in the `NLPContributions-Auto` that contain at least one contribution statement of various subtypes of (a) *knowledge contributions* and (b) *artifact contributions*.

contributions (drop to $\sim 40\%$). Yet, post-2020, we see a marked increase in contributions to language.

b. Evolution in human-centric studies. While early NLP research had a high percentage ($\sim 80\%$) of k-people contributions, the percentage has declined steadily over time, dropping to $\sim 20\%$ by the late 1990s. The percentage has stayed roughly steady since then until the late 2010s, when we see the beginnings of an increasing trend.

c. Consistent focus on NLP tasks. During the early eighties, we observed a sharp increase in contributions focused on knowledge about NLP tasks, which has remained consistently high over the decades.

d. Steadily rising method and dataset knowledge. Unlike other knowledge contributions, there has been a steady rise in contributions towards knowledge about datasets and methods over the years, with contributions to methods showing a more pronounced increase since the nineties.

Figure 3b shows artifact contribution trends as follows:

a. Sharp rise in method artifacts. We observe a sharp rise in contributions related to new methods beginning in the early nineties, which has sustained at high levels since then.

b. Steady rise in task and dataset artifacts. Similar to methods, contributions toward artifact tasks have increased since the nineties. Both artifact tasks and datasets have shown a steady rise over the years, with a positive correlation between their growth.

Discussion. In its early days (throughout the seventies and eighties), NLP research had a strong focus on language (Brachman, 1979; Lebowitz, 1979; Herskovits, 1980). The early nineties marked a shift in NLP's focus with the advent of statistical models (Brown et al., 1993), the release of the Penn dataset (Marcus et al., 1993), and later the establishment of the EMNLP conference. While recent discussions often highlight newer methods or models (such as transformers or LLMs), our findings indicate that the shift towards contributions in methods or models began in the early nineties. That era set the stage for the development of newer methods that continue to shape NLP research. The recent rise in contributions toward new knowledge about people and language is likely due to the rise of new NLP sub-fields such as computational social science, culturonomics and digital humanities, and ethics in NLP (refer to Appendix A.1 for an analysis of recent advancements in NLP research.). Finally, it is interesting to note that while different contribution types have ebbed and risen at different times, the percentages of *all types* are moderate to high in the past five years. This is evidence that a great number of NLP papers are now contributing in multiple ways — adding to knowledge and artifacts of many kinds (see supplementary analysis in Appendix A.2).

5.2 Contributions and Venues

Q3. How do venues influence the nature of NLP research?

Each publication venue maintains distinct expectations regarding the types of work it accepts, such as the focus on particular topics or the nature of the experiments conducted. We examine the distinct types and further sub-types of the contribution statements in the abstracts of the papers across different venues normalized by the number of papers published in that venue.

Results. We present detailed venue-specific statistics in Figure 7 (Appendix D), and summarize the key findings below.

a. Similar contributions across majority venues. The majority of conferences (such as

ACL, EMNLP, NAACL, etc.) display similar distributions regarding the types of contributions in their published papers: roughly 68% artifacts (task: 71%, method: 89%, dataset: 42%) and 69% knowledge (task: 94%, method: 77%, people: 44%, dataset: 65%, language: 61%).
b. Distinctiveness of EMNLP and CL. EMNLP is distinguished by a notably higher volume of artifact-method contributions, highlighting its emphasis on empirical methodologies. Conversely, the CL journal is unique among *CL venues for its greater focus on expanding knowledge about language and people and a comparatively lesser emphasis on machine learning.

Q4. How has the nature of NLP research papers changed across different venues over time?

We hypothesize that as a field matures, latent community norms develop, steering the research direction and leading to a more uniform distribution of contributions across different venues over time. To test this, we examine the change of specific types and sub-types of contributions in the abstracts of papers from each venue over time.

Results. We present the temporal distribution of contribution types across venues in Figure 6 (Appendix D) and summarize the key results below.
a. Spread of trends. We first observe a decline in contributions concerning knowledge about people and language at the ACL in the early 1990s, which then gradually appears in CL.
b. Increasing similarity of newer conferences. The trend towards similar contribution distributions is evident in newer conferences (such as EMNLP, NAACL, AACL, etc.).

Additionally, in Appendix A.2, we examine whether these venues increasingly mirror ACL's contribution distribution over time.

5.3 Contributions and Citation Impact

Q5. How do different contribution types influence citation dynamics?
All contribution types are important for a thriving and vibrant ecosystem of NLP research. Thus, marked disparities in citation counts could potentially disincentivize work on certain types of contributions. Therefore, through this question, we track the citational impact of different contribution types. We calculate the average and median citation counts for each type of contribution from the papers that have at least one contribution statement pertaining to that type. To ensure a meaningful assessment of citation trajectories, we focus on papers with at least five years of publication history (Anderson et al., 2012). For this purpose, we selected 352 papers from the ACL'18 to examine the citation impact of papers published simultaneously for this experiment.

Results. Below we summarize the results presented in Table 5.
a. Dataset artifacts attract higher citations. Regarding artifact contributions, papers that introduce new datasets tend to attract notably high citations. Additionally, those proposing new methods attract more citations compared to those introducing new tasks.
b. Greater interest in technical advancements. Papers that contribute knowledge about methods or datasets (primarily through analysis) tend to receive more citations than those focused on people or language. This suggests a greater community interest in technical advancements over sociolinguistic studies.
c. Lower citation impact for language contributions. Notably, even though more papers focus on expanding knowledge about language compared to those about people, language-focused contributions tend to receive fewer citations.

We additionally analyze 277 papers from ACL'17 (Table 10, Appendix D), and our analysis reveals similar trends.

Discussion. It is important to recognize that citations are influenced by various factors beyond just contribution types. Our objective is neither to identify all possible influences on citation counts nor to pinpoint the most influential factors. However, the high citations for papers that create new datasets perhaps reflect the importance of datasets in much of NLP research, particularly for training and evaluating models - a common practice in modern NLP. Figure 9 (Appendix D) shows the distribution of citation counts.

6 Conclusions and Discussion

In this paper, we propose that automatically extracting, categorizing, and quantitatively analyzing contribution statements in research papers offers insights into the nature of the field. We introduce a taxonomy of contributions and develop a framework for automatically processing the contribution statements from NLP papers (§ 3).

Contribution		#papers	#citations	
Typ.	Sub-typ.		mean	median
Knowledge	k-dataset	219	121.1	56.0
	k-language	193	107.1	53.0
	k-method	280	127.8	56.0
	k-people	119	109.5	54.0
	k-task	328	115.7	55.0
Artifact	a-dataset	154	137.7	64.0
	a-method	310	122.2	58.0
	a-task	270	116.0	56.0

Table 5: Mean and median citation counts of papers for different contributions in ACL'18.

Our analysis reveals that although NLP is intrinsically linked to linguistics and society, its current research focus is dominant towards advancements in technical methods (§ 5.1). This shift toward newer methods, often discussed in the context of recent models like transformers and LLMs, actually began in the early nineties. However, an increased focus on methodology does not necessarily indicate a reduced emphasis on language or people. This is evident in post-2020 NLP research, where there is a growing interest in sociolinguistics and the use of NLP in social sciences alongside technical innovations like LLMs. Additionally, our analysis shows the field's growth and progression, as reflected by the growing complexity and diversity of contribution types within research papers (§ 5.2).

All contribution types play a vital role in sustaining a vibrant and dynamic NLP research ecosystem. Notably, we observe that artifact contributions – particularly papers introducing new datasets – tend to receive more citations than other types (§ 5.3). Although the growth of NLP is beneficial, we emphasize the importance of maintaining diversity in research contributions to ensure the field remains relevant to a broader community. As members of this community, we hold the strength to guide the future direction of these trends. However, we are not advocating for a specific stance on research practices but encourage an inclusive approach that embraces a variety of contribution types within NLP research. To foster future research in the area of contribution analysis and stimulate informed discussions within our community, we release our artifacts under 🅭🅯🄏🄎.

7 Applications and Future Work

The `NLPContributions` dataset, which includes contribution statements annotated with their respective types, makes it valuable for a wide range of research projects and applications. Identifying contribution statements helps researchers efficiently navigate the growing body of literature by capturing the core ideas of each paper (Fok et al., 2024). The dataset, along with the proposed taxonomy, also holds promise for advancing tasks such as automatic survey generation (Wang et al., 2024) and question answering within scientific literature (Dasigi et al., 2021). Categorizing contributions can further support researchers in locating studies with similar types of contributions or compiling structured literature review tables (Newman et al., 2024).

Beyond these applications, the dataset can be used to study how NLP research and its publication venues have evolved over time. For example, we are interested in studying the relationship between the diversity of contribution types and venue size, measured by the number of accepted papers and the range of research contributions represented. Additionally, we aim to explore and quantify the influence of different contributions of the same type, investigating how their impact evolves over time and what factors contribute to making a contribution influential.

Limitations

This study primarily examines NLP research papers from the ACL Anthology, specifically focusing on papers from conferences and journals under ACL Events, such as ACL, EMNLP, NAACL, EACL, and journals like TACL and CL. However, it is crucial to recognize that significant NLP research also appears outside the ACL Anthology, including in AI venues, regional conferences, and preprint servers. While papers published in the ACL Anthology are typically of high quality, research from other venues often contributes valuable insights to the field. We leave the effort to curate and include research papers from these alternative venues for future work.

Our study primarily analyzes the abstracts of research papers, which are typically concise, logically coherent paragraphs that hold a unique position within the paper. While abstracts are likely to contain the key contributions as highlighted by the authors, making them a focal point for initial analysis, it is important to acknowledge that unique contributions may also be found within the main body of the paper. However, annotating the full text of research papers requires significant time, effort, and substantial domain knowledge to accurately understand and contextualize the content. In future iterations of this work, we plan to extend our annotations to include the main body of the papers, providing a more inclusive dataset.

Finally, for our analysis, we first train a classifier on the high-quality, human-annotated dataset that we create and then deploy this trained model on the larger ACL Anthology dataset. We conduct our analysis based on the labels generated by this model. It is important to acknowledge that no model achieves perfect accuracy, which can impact the quality of such analysis. However, as demonstrated by Teodorescu and Mohammad (2023), when broader cumulative trends are derived from large datasets using such models, the results tend to be highly accurate and show a strong correlation with trends identified through gold-label analysis. This supports the reliability and accuracy of our analysis despite the inherent limitations of trained machine-learning models.

Ethics Statement

In this work, we utilize publicly accessible data from the ACL Anthology and do not involve any personal data. It is important to acknowledge that although our approach is data-driven, individual views on research are naturally subjective. Therefore, decisions in science should not only be based on data but also take into account ethical, social, and other qualitative considerations.

Acknowledgements

This work has been funded by the German Research Foundation (DFG) as part of the Research Training Group KRITIS No. GRK 2222. We also gratefully acknowledge Microsoft for providing access to OpenAI GPT models via the Azure cloud (Accelerate Foundation Model Academic Research).

We thank Aishik Mandal for his voluntary participation in the annotation study conducted for this research. We also appreciate the feedback on the initial draft of this manuscript provided by Hiba Arnaout, Sukannya Purkayastha, Fengyu Cai, Md Imbesat Hassan Rizvi, and Ilia Kuznetsov.

References

Titipat Achakulvisut, Chandra Bhagavatula, Daniel Acuna, and Konrad Kording. 2019. Claim extraction in biomedical publications using deep discourse model and transfer learning. *arXiv preprint arXiv:1907.00962*.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Khalid Al Khatib, Tirthankar Ghosal, Yufang Hou, Anita de Waard, and Dayne Freitag. 2021. Argument mining for scholarly document processing: Taking stock and looking ahead. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 56–65, Online. Association for Computational Linguistics.

Ashton Anderson, Dan Jurafsky, and Daniel A. McFarland. 2012. Towards a computational history of the ACL: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21, Jeju Island, Korea. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Yoav Kantor, Roni Friedman, and Noam Slonim. 2021. Every bite is an experience: Key Point Analysis of business reviews. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3376–3386, Online. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Ronald J. Brachman. 1979. Taxonomy, descriptions, and individuals in natural language understanding. In *17th Annual Meeting of the Association for Computational Linguistics*, pages 33–37, La Jolla, California, USA. Association for Computational Linguistics.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2024. Sparks of artificial general intelligence:

Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 10.

Chaomei Chen and Min Song. 2019. Visualizing a field of research: A methodology of systematic scientometric reviews. *PLOS ONE*, 14(10):1–25.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019. Seeing things from a different angle:discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Herbert H Clark and Michael F Schober. 1992. Asking questions and influencing answers. *Contemporary Sociology*, 22(1):15–48.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Yuntian Deng, Volodymyr Kuleshov, and Alexander Rush. 2022. Model criticism for long-form text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11887–11912, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cailing Dong and Ulrich Schäfer. 2011. Ensemble-style self-training on citation classification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 623–631, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Jennifer D'Souza and Sören Auer. 2020. NLPContributions: An annotation scheme for machine reading of scholarly contributions in natural language processing literature. In *Proceedings of the 3rd Workshop on Extraction and Evaluation of Knowledge Entities from Scientific Documents (EEKE 2022)*, Germany and Online. CEUR-WS.

Jennifer D'Souza, Sören Auer, and Ted Pedersen. 2021. SemEval-2021 task 11: NLPContributionGraph - structuring scholarly NLP contributions for a research knowledge graph. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 364–376, Online. Association for Computational Linguistics.

Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. 2017. Identifying products in online cybercrime marketplaces: A dataset for fine-grained domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2598–2607, Copenhagen, Denmark. Association for Computational Linguistics.

Sauleh Eetemadi and Kristina Toutanova. 2014. Asymmetric features of human generated translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 159–164, Doha, Qatar. Association for Computational Linguistics.

Raymond Fok, Luca Soldaini, Cassidy Trier, Erin Bransom, Kelsey MacMillan, Evie Cheng, Hita Kambhamettu, Jonathan Bragg, Kyle Lo, Marti A. Hearst, Andrew Head, and Daniel S. Weld. 2024. Accelerating scientific paper skimming with augmented intelligence through customizable faceted highlights. *ACM Trans. Interact. Intell. Syst.*, 14(4).

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mark Garzone and Robert E Mercer. 2000. Towards an automated citation classifier. In *Advances in Artificial Intelligence: 13th Biennial Conference of the Canadian Society for Computational Studies of Intelligence, AI 2000 Montéal, Quebec, Canada, May 14–17, 2000 Proceedings 13*, pages 337–346. Springer.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2021. ContriSci: a bert-based multitasking deep neural architecture to identify contribution statements from research papers. In *Proceedings of the 23rd International Conference on Asia-Pacific Digital Libraries: Towards Open and Trustworthy Digital Societies, ICADL 2021*, pages 436–452, Virtual Event. Springer.

Komal Gupta, Ammaar Ahmad, Tirthankar Ghosal, and Asif Ekbal. 2024. A BERT-based sequential deep neural architecture to identify contribution statements and extract phrases for triplets from scientific publications. In *International Journal on Digital Libraries*, volume 25, pages 1–28. Springer.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 363–371, Honolulu, Hawaii. Association for Computational Linguistics.

Annette Herskovits. 1980. On the spatial uses of prepositions. In *18th Annual Meeting of the Association for Computational Linguistics*, pages 1–5, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2019. Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5203–5213, Florence, Italy. Association for Computational Linguistics.

Yufang Hou, Charles Jochim, Martin Gleize, Francesca Bonin, and Debasis Ganguly. 2021. TDMSci: A specialized corpus for scientific literature entity tagging of tasks datasets and metrics. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 707–714, Online. Association for Computational Linguistics.

Marcio Inácio and Thiago Pardo. 2021. Semantic-based opinion summarization. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 619–628, Held Online. INCOMA Ltd.

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

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

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Anne Lauscher, Brandon Ko, Bailey Kuehl, Sophie Johnson, Arman Cohan, David Jurgens, and Kyle Lo. 2022. MultiCite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1875–1889, Seattle, United States. Association for Computational Linguistics.

Michael Lebowitz. 1979. Reading with a purpose. In *17th Annual Meeting of the Association for Computational Linguistics*, pages 59–63, La Jolla, California, USA. Association for Computational Linguistics.

Sha Li, Chi Han, Pengfei Yu, Carl Edwards, Manling Li, Xingyao Wang, Yi Fung, Charles Yu, Joel Tetreault, Eduard Hovy, and Heng Ji. 2023. Defining a new NLP playground. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11932–11951, Singapore. Association for Computational Linguistics.

Maria Liakata, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

ML Menéndez, JA Pardo, L Pardo, and MC Pardo. 1997. The jensen-shannon divergence. *Journal of the Franklin Institute*, 334(2):307–318.

Filipe Mesquita, Jordan Schmidek, and Denilson Barbosa. 2013. Effectiveness and efficiency of open relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 447–457, Seattle, Washington, USA. Association for Computational Linguistics.

Meta. 2023. Llama-3-8b: Open-source large language model trained on 15 trillion tokens. `https://ai.meta.com/blog/meta-llama-3/`.

John Mingers and Loet Leydesdorff. 2015. A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246(1):1–19.

Saif M. Mohammad. 2020a. Examining citations of natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209, Online. Association for Computational Linguistics.

Saif M. Mohammad. 2020b. NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 868–877, Marseille, France. European Language Resources Association.

Benjamin Newman, Yoonjoo Lee, Aakanksha Naik, Pao Siangliulue, Raymond Fok, Juho Kim, Daniel S Weld, Joseph Chee Chang, and Kyle Lo. 2024. ArxivDIGESTables: Synthesizing scientific literature into tables using language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9612–9631, Miami, Florida, USA. Association for Computational Linguistics.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2023. GPT-3.5-Turbo and GPT-4-Turbo: Instruction-following models fine-tuned with reinforcement learning from human feedback. `https://platform.openai.com/docs/models`.

Charles Oppenheim and Susan P Renn. 1978. Highly cited old papers and the reasons why they continue to be cited. *Journal of the American Society for Information Science*, 29(5):225–231.

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for Chinese social media with jointly trained embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 548–554, Lisbon, Portugal. Association for Computational Linguistics.

Vinodkumar Prabhakaran, William L. Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1170–1180, Berlin, Germany. Association for Computational Linguistics.

Aniket Pramanick, Yufang Hou, Saif Mohammad, and Iryna Gurevych. 2023. A diachronic analysis of paradigm shifts in NLP research: When, how, and why? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2312–2326, Singapore. Association for Computational Linguistics.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In *Machine Learning and Knowledge Discovery in Databases*, pages 254–269, Berlin, Heidelberg. Springer Berlin Heidelberg.

Furkan Şahinuç, Thy Thy Tran, Yulia Grishina, Yufang Hou, Bei Chen, and Iryna Gurevych. 2024. Efficient performance tracking: Leveraging large language models for automated construction of scientific leaderboards. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7963–7977, Miami, Florida, USA. Association for Computational Linguistics.

Said A Salloum, Mostafa Al-Emran, Azza Abdel Monem, and Khaled Shaalan. 2017. A survey of text mining in social media: Facebook and twitter perspectives. *Advances in Science, Technology and Engineering Systems*, 2(1):127–133.

Dudley Shapere. 1964. The structure of scientific revolutions. *The Philosophical Review*, 73(3):383–394.

Daniel Sosa, Malavika Suresh, Christopher Potts, and Russ Altman. 2023. Detecting contradictory COVID-19 drug efficacy claims from biomedical literature. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 694–713, Toronto, Canada. Association for Computational Linguistics.

Mary Elizabeth Stevens and Vincent Edward Giuliano. 1965. *Statistical association methods for mechanized documentation: Symposium proceedings*, volume 269. US Government Printing Office.

Chenhao Tan, Dallas Card, and Noah A. Smith. 2017. Friendships, rivalries, and trysts: Characterizing relations between ideas in texts. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 773–783, Vancouver, Canada. Association for Computational Linguistics.

Daniela Teodorescu and Saif Mohammad. 2023. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. In *Findings of the Association for Computational Linguistics:*

EMNLP 2023, pages 4124–4137, Singapore. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110, Sydney, Australia. Association for Computational Linguistics.

Simone Teufel et al. 1999. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, University of Edinburgh.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2022. Learning from disagreement: A survey. *J. Artif. Int. Res.*, 72:1385–1470.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Jan Philip Wahle, Terry Ruas, Saif Mohammad, and Bela Gipp. 2022. D3: A massive dataset of scholarly metadata for analyzing the state of computer science research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2642–2651, Marseille, France. European Language Resources Association.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. Autosurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems*, volume 37, pages 115119–115145. Curran Associates, Inc.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022b. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

An Yang and Sujian Li. 2018. SciDTB: Discourse dependency TreeBank for scientific abstracts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 444–449, Melbourne, Australia. Association for Computational Linguistics.

Chao Zhao, Tenghao Huang, Somnath Basu Roy Chowdhury, Muthu Kumar Chandrasekaran, Kathleen McKeown, and Snigdha Chaturvedi. 2022. Read top news first: A document reordering approach for multi-document news summarization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 613–621, Dublin, Ireland. Association for Computational Linguistics.

Jiayi Zhao, Xipeng Qiu, Shu Zhang, Feng Ji, and Xuanjing Huang. 2012. Part-of-speech tagging for Chinese-English mixed texts with dynamic features. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1379–1388, Jeju Island, Korea. Association for Computational Linguistics.

Yuxiang Zhou, Lejian Liao, Yang Gao, Zhanming Jie, and Wei Lu. 2021. To be closer: Learning to link up aspects with opinions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3899–3909, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

A Supplementary Analysis

In this section, we supplement the main analysis (§ 5) with additional insights to provide a comprehensive overview of the nature of NLP research.

A.1 Evolving Contributions in NLP Research

Q6. How has the nature of NLP research evolved in recent years?

NLP research is experiencing an exciting phase (Li et al., 2023), often referred to as the "deep learning era" (Pramanick et al., 2023), beginning in the late 2010s with the seminal work by Vaswani et al. (2017), followed by BERT (Devlin et al., 2019), and the rise of Large Language Models (LLMs). We showed in Section 5, that the transformative shift in NLP research began in the early 1990s, setting the stage for these recent advancements. This section, however, focuses on the evolution of NLP research contributions since the late 2010s.

Results. We refer to Figure 3a, and summarize the key findings regarding knowledge contributions below.

a. Beginning of new research trends. While contributions toward *k-language* and *k-people* declined in the early 1990s, the late 2010s (and especially early 2020s) have seen the beginning of a research trend marked by increased research contributions in these areas.

b. Broad contributions spectrum. In the last five years, there has been a moderate to high increase in the percentage of all types of contributions.

We refer to Figure 3b, to summarize the key findings regarding artifact contributions.

a. Steeper rise in dataset contributions. Since the late 2010s, there has been a marked increase in research papers contributing new dataset artifacts, with *a-dataset* showing a steeper rise compared to earlier periods.

b. Increasing new tasks. Alongside the increasing contributions of type *a-dataset*, there is also a rise in contributions toward *a-task*, indicating a growth in the introduction of new tasks within NLP.

Discussion. Newer models (such as the LLMs) excel at solving standard NLP tasks such as entity typing, sentiment analysis, and textual entailment (Wei et al., 2022a). Rather than setting benchmarks, researchers explore new capabilities of these models and propose novel tasks (Bubeck et al., 2024). These models are also adept at handling various complex tasks like chain-of-thought reasoning (Wei et al., 2022b). However, evaluating their capabilities often necessitates the collection of larger datasets, likely contributing to an increase in *a-dataset* contributions.

The increasing contributions to language knowledge may be tied to developments in newer models. Although Large Language Models (LLMs) are multilingual, i.e., trained on data from multiple languages, their performance is not uniformly effective across all languages for tasks such as classification or generation. To address this issue and improve model efficiency across various languages, researchers are increasingly focusing on studying the nuances of languages (Aguilar et al., 2020), thereby contributing to the knowledge of language. Similarly, efforts to address and mitigate social biases and stereotypes in LLM outputs (Omrani et al., 2023), which often reveal their inherent flaws, have led to an increase in contributions focused on human-centered NLP.

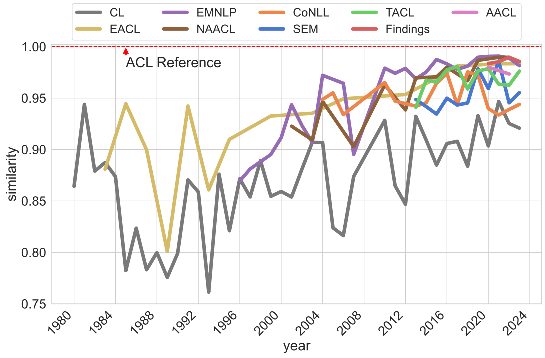


Figure 4: Comparison of venue similarity based on contribution types.

A.2 Contributions and Venues

Q7. Are other venues mirroring the ACL conference in shaping the nature of NLP research?

The ACL conference is the largest and arguably most prestigious among the ACL Events, hosting about 30.3% of the papers published in these venues. Given its prominent position, it is interesting to investigate whether other conferences have gradually begun to mirror the distribution of the types of contributions featured in ACL over time. We compare the distribution of types and sub-types of contribution statements in papers from these venues with that of those from the ACL conference in the same year, using the Jensen-Shannon divergence (Menéndez et al., 1997), where a value close to 1 indicates similar distributions.

Results. Figure 4 shows the following trend across the venues.

Convergence of NLP conferences with ACL. Over the years, conferences have become increasingly similar to the ACL conference in terms of the distribution of the types of contributions their papers present. For instance, the EMNLP conference, originally established to focus on empirical findings, has shown growing similarity to ACL. Similarly, newer venues like AACL and Findings closely align with ACL's contribution patterns.

Discussion. This trend tends to confirm our hypothesis that, over time, a common publication norm has emerged across conferences, leading to a more institutionalized standard in NLP research. On the other hand, it is also arguably a loss that the different venues do not have unique characteristics,

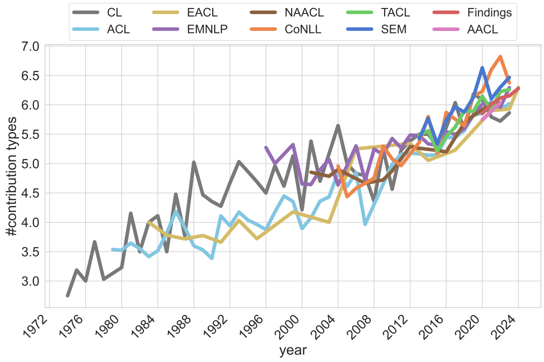


Figure 5: Average number of contribution types per paper across different venues.

championing and valuing different kinds of works.

Q8. Do journal papers exhibit a greater variety of contribution types than conference papers? Different publication venues have varying constraints on the number of pages they allow, with journals typically offering more space compared to the stricter page limits at NLP conferences. To investigate whether journal papers utilize the additional space to include a wider range of contribution types, we analyze the average number of unique contributions per paper across various venues on an annual basis.

Results. Figure 5 shows the following.
a. Rising diversity in contributions. The average number of unique contribution types in the abstracts of conferences and journals has been consistent, and this number has shown an upward trend over time.
b. Expansion in NLP applications. The consistent average length of abstracts in venues, yet diverse contribution types (Figure 8, Appendix D) indicates the growing sophistication and expansion of NLP applications.

B Corpus Details

To address the question *"What constitutes NLP Research?"*, first We used S2ORC to gather the abstracts of $29,010$ papers published from conferences or journals falling within the "ACL Events" category. Notably, in 1997, the ACL and EACL conferences were held jointly, resulting in 73 papers being listed under both events for that year. We treated these dual-listed papers as single entries, reducing our dataset to a total of $28,937$ unique research papers. Additionally, we collected the metadata associated with these papers from the `anthology.bib`.

Finally, we applied the SciBERT model, fine-tuned on `NLPContributions`, to the sentences from these abstracts to identify the contribution statements and classify them according to the predefined taxonomy.

C `NLPContributions` Annotation Guidelines

We propose a linguistic annotation scheme to study and analyze the types of contributions articulated in NLP research papers following the taxonomy we developed in § 3.1. The goal of this annotation scheme is to annotate contribution statements from the abstract section of NLP research papers into various types and further into sub-types as mentioned in Table 1, according to the specific aspect of field advancement they represent. In this section, we elaborate on both broad categories and more detailed sub-types of contributions within each category.

C.1 Contribution Types

C.1.1 Type: Artifact

This type of contribution includes the creation of new resources such as datasets, models, or algorithms. We further sub-categorize contributions into three distinct categories based on the type of artifact they introduce to the field.

- *a-dataset:* Researchers create new scientific corpus or language resources as artifacts to build models or analyze languages, such as SQuAD or Penn Treebank.
- *a-method:* Researchers often create new NLP methods such as algorithms or models (for example, BERT (pre-trained language model) or LLaMA (large language model)), as artifacts primarily to solve tasks and describe them in research papers.
- *a-task:* Researchers often identify or formulate new or previously unknown problems (such as linguistic problems like NER Tagging) and formally describe them in research papers as tasks.

C.1.2 Type: Knowledge

This type of contribution encompasses the addition of new insights or understandings to the field. These contributions often relate to linguistic studies due to the significant overlap between NLP and linguistics, or they may explore societal and human aspects because of NLP's focus on human language. Consequently, we further categorize it into five sub-types based on the specific area of knowledge it expands.

- *k-dataset:* Contribute new insights or analysis of an NLP dataset.
- *k-language:* Adds new knowledge about natural language.
- *k-method:* Enhances the understanding of algorithms, methods or models within NLP.
- *k-people:* Explores and adds knowledge about aspects of human behavior and social implications as revealed through natural language.
- *k-task:* Contribute new insights into specific NLP task(s).

C.2 Annotation Instructions

First, we present the following two definitions to the annotators along with the contribution types as described in § C.1.

Definition C.1 (Contribution)**.** A contribution is a scientific achievement attributed to the authors of a research paper, such as introducing a new model or dataset.

Definition C.2 (Contribution Statement)**.** A statement in a research paper that describes new scientific achievements attributed to its authors is called a contribution statement.

Next, we present the annotators with the title and abstract of a research paper and instruct them to annotate each statement of the abstract according to the following questions.

Q1. *Does the sentence qualify as a contribution statement according to the definitions presented earlier?*

In the second question, multiple options could be selected.

Q2. *If you answered yes to the previous question, which of the following options most accurately describes the type and sub-type of the contribution statement? Please select all that apply.*

- *Artifact-Task* (Introduces, proposes, or formulates a new or novel NLP task.)
- *Artifact-Method* (Introduces or creates a new or novel NLP method such as an algorithm or a new NLP model.)
- *Artifact-Dataset* (Creates a new corpus or language resource.)
- *Knowledge-Task* (Describes new knowledge about NLP task.)
- *Knowledge-Dataset* (Describes new knowledge about datasets, such as their new properties or characteristics.)
- *Knowledge-Method* (Describes or presents new knowledge or analysis about NLP models or methods, which are primarily drawn from Machine Learning.)
- *Knowledge-Language* (Presents new knowledge about language, such as a new property or characteristic of language.)
- *Knowledge-People* (Presents new knowledge about people, humankind, society, or human civilization.)
- Others (Any other type that does not fall under the categories mentioned above.)

Discussion: We observe that only a small number of NLP papers propose new *metrics for evaluation measures*. Since these metrics function as algorithms or methods, we categorize them under the class "artifact-method".

Also note that, in the initial pilot annotation study, we included an "Others" label (as mentioned above) to capture any types of contributions not already accounted for in our taxonomy. This allowed us to potentially expand our taxonomy based on the pilot results. Following the pilot study, however, we found that our existing taxonomy adequately covered all types of contributions identified by the annotators.

C.3 Annotation Statistics

We provide statistics for the abstracts of 100 papers annotated by two annotators post-adjudication. Of the 584 statements annotated, 359 were identified as contributions. Table 6 details the number of these statements across each type and subtype of contributions.

Typ.	Sub-typ.	#Contrib.
Knowledge	k-dataset	25
	k-language	23
	k-method	53
	k-people	41
	k-task	122
Artifact	a-dataset	20
	a-method	94
	a-task	22

Table 6: Number of annotated contribution statements pertaining to each contribution type, as annotated by two annotators.

D Supplementary Results

Typ.	Sub-typ.	macro-F1
Knowledge	k-dataset	0.80
	k-language	0.80
	k-method	0.80
	k-people	0.81
	k-task	0.81
Artifact	a-dataset	0.80
	a-method	0.81
	a-task	0.80

Table 7: SciBERT's performance in identifying statements for each contribution type.

	venue	avg. #sent.
conference	ACL	5.69
	EMNLP	6.23
	NAACL	5.74
	EACL	5.67
	AACL	6.43
	Findings	6.94
	SEM	5.48
	CoNLL	5.73
journal	TACL	6.10
	CL	9.01

Table 8: Conference-wise average abstract sentence count.

Model	0-shot	1-shot	3-shot	5-shot
GPT-3.5-Turbo	0.52	0.55	0.64	0.73
GPT-4-Turbo	**0.61**	**0.66**	**0.72**	**0.80**
LLaMA-3	0.50	0.50	0.51	0.53

Table 9: LLM performance (`macro-F1`) with different number of training examples.

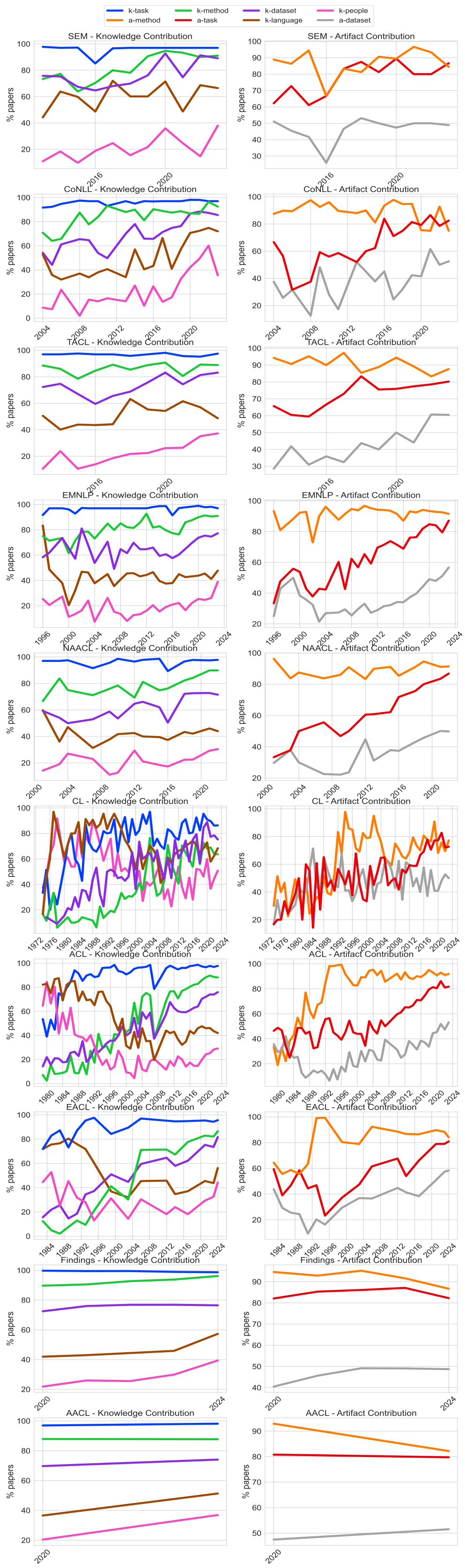


Figure 6: Evolution of NLP conferences (and journals) based on the percentage of papers containing at least one contribution of each type (Abbr.: knowledge (k), artifact (a)). Refer to Figure 10, 11, 12 for larger figures.

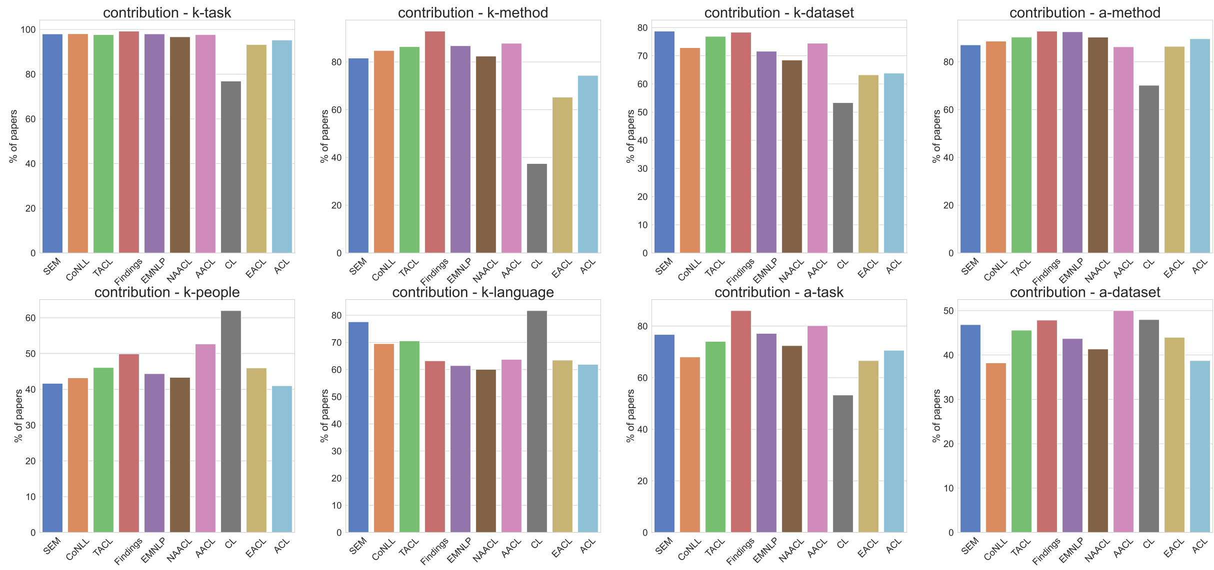


Figure 7: Distribution of contribution types across research papers by conference (Abbr.: knowledge (k), artifact (a)).

Contribution		#papers	avg. citation (↑)
Typ.	Sub-typ.		
Knowledge	k-dataset	180	120.0
	k-language	160	110.3
	k-method	230	135.3
	k-people	101	104.7
	k-task	254	121.5
Artifact	a-dataset	96	140.7
	a-method	250	131.5
	a-task	202	120.8

Table 10: Average citation counts by contributions for ACL 2017 Papers.

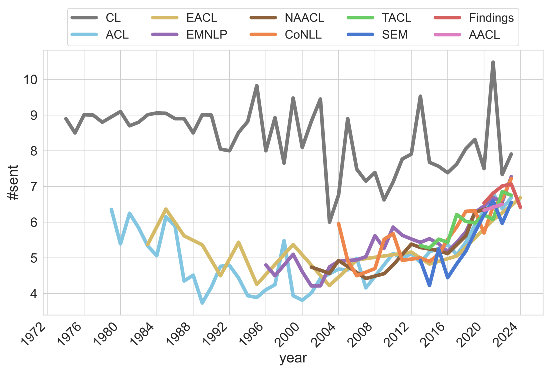
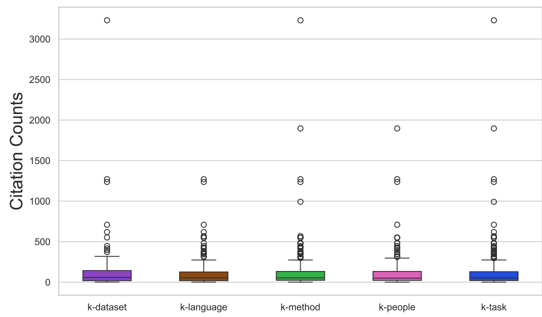
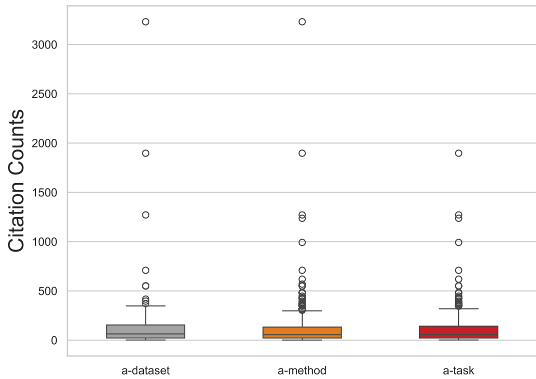


Figure 8: Average abstract length in papers from different venues.



(a) Knowledge contributions



(b) Artifact contributions

Figure 9: Variability and asymmetry in citation counts for each contribution type.

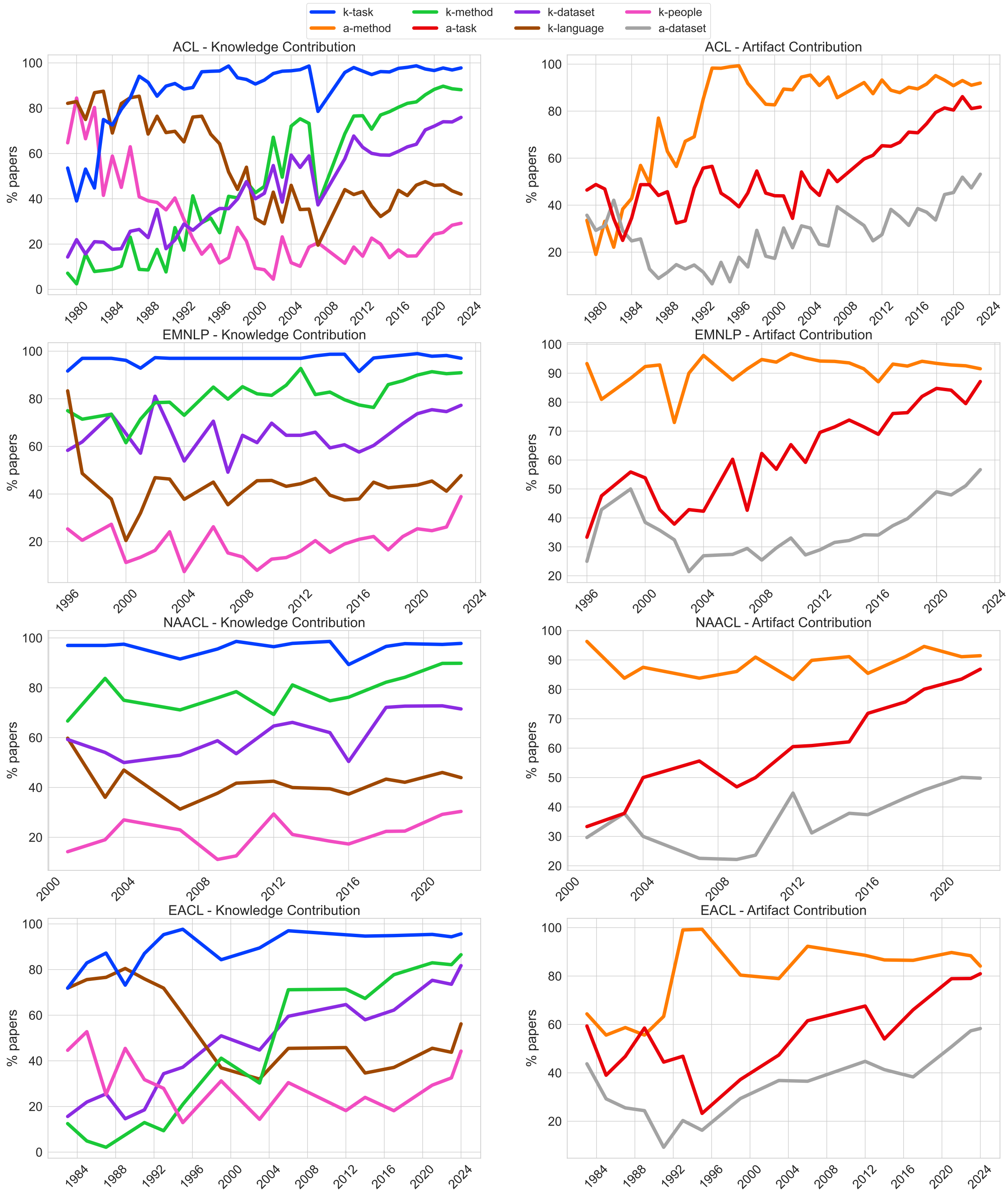


Figure 10: Evolution of the four venues (ACL, EMNLP, NAACL, and EACL) based on the percentage of papers containing at least one contribution of each type (Abbr.: knowledge (k), artifact (a)).

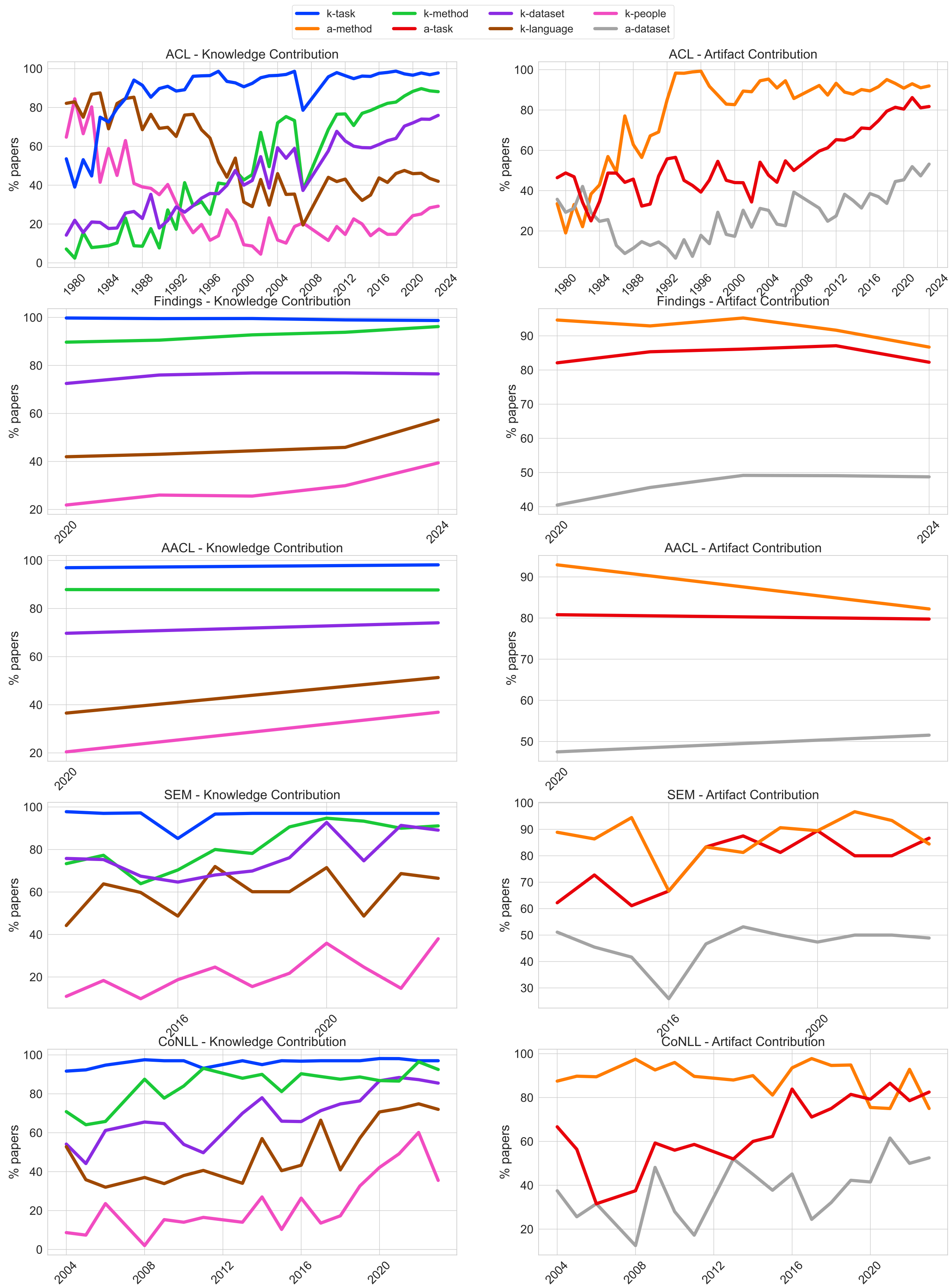


Figure 11: Evolution of the five venues (ACL, Findings, AACL, *SEM, and CoNLL) based on the percentage of papers containing at least one contribution of each type (Abbr.: knowledge (k), artifact (a)).

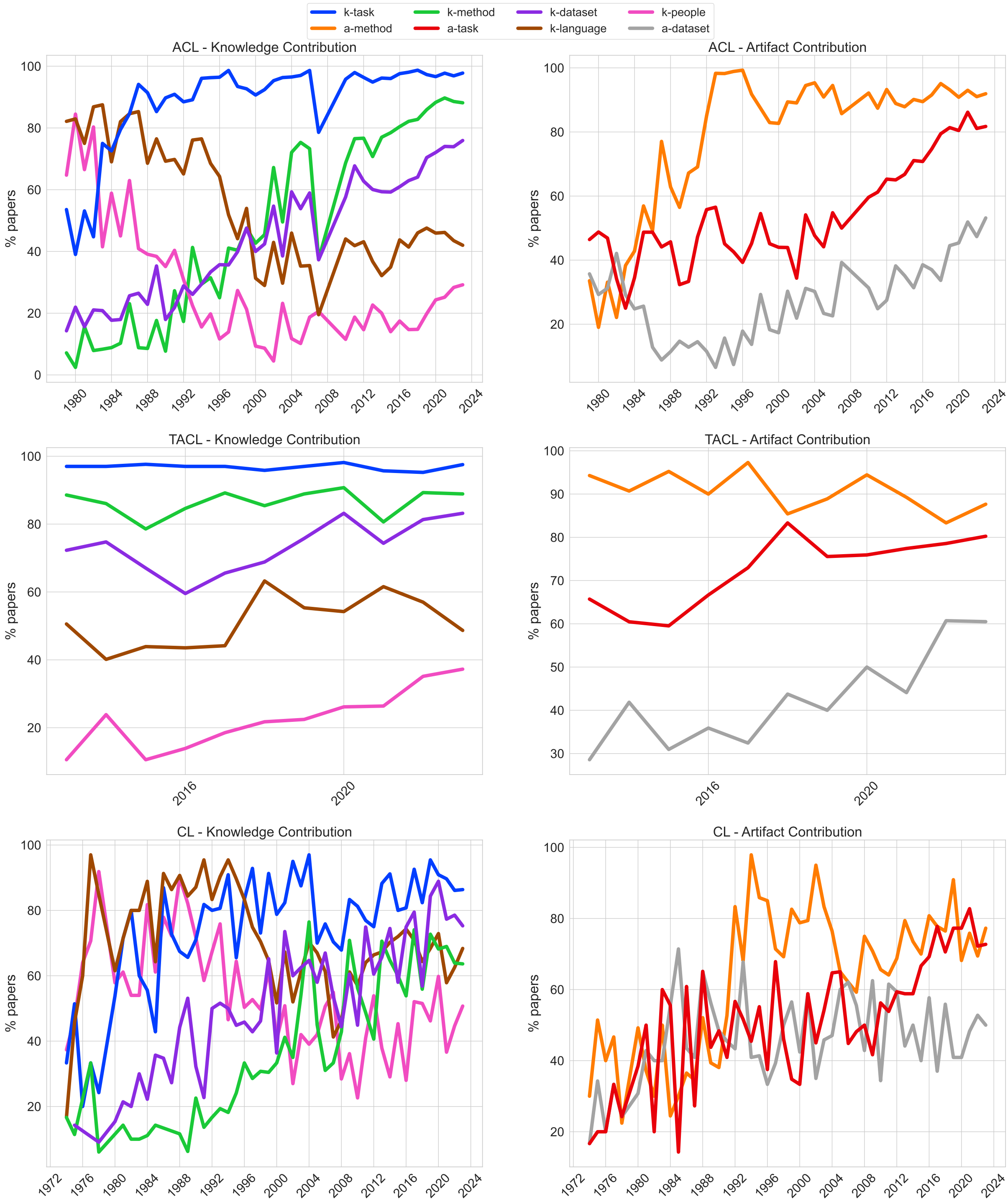


Figure 12: Evolution of the three venues (ACL, TACL, and CL) based on the percentage of papers containing at least one contribution of each type (Abbr.: knowledge (k), artifact (a)).

Type	Sub-type	Prompt
Knowledge	k-task	Central to NLP research are tasks such as Machine Translation, Named Entity Recognition, Language Modeling, etc. Your task is to assess whether the provided sentence from an NLP research paper describes new knowledge about any of such existing NLP tasks, including new knowledge about their properties or characteristics. However, the sentence should not propose a new NLP task. Respond with "yes" if the sentence presents new knowledge about one or more of these NLP tasks; otherwise, respond with "no".
	k-method	NLP Models such as RNNs, LSTMs or LLMs are indispensable for NLP Research. Your task is to determine if the provided sentence from an NLP research paper describes new knowledge or analysis about such existing NLP models or methods like RNNs, LSTMs, or LLMs. However, the sentence should not propose new models or methods. Respond with "yes" if the sentence presents new knowledge about NLP models; otherwise, respond with "no."
	k-people	In NLP research, every paper plays a role in advancing the field. Your task is to assess whether the given sentence from an NLP research paper presents new knowledge about people, humankind, society or human civilization. Respond with "yes" if the sentence describes novel knowledge about people, humankind, society or human civilization; otherwise, respond with "no." Use only a yes or no format for your answers.
	k-dataset	Datasets constitute a crucial aspect of NLP and machine learning research. Examining datasets can yield valuable insights into their properties and features. Your task is to assess whether the given sentence from an NLP research paper describes new knowledge about a dataset, such as its new properties or characteristics or describes new knowledge concerning properties or characteristics of datasets in general. Respond with "yes" if the sentence presents novel knowledge about the datasets; otherwise, respond with "no." Use only a yes or no format for your answers.
	k-language	In NLP research, every paper plays a role in advancing the field. Your task is to assess whether the given sentence from an NLP research paper presents new knowledge about language, such as a new property or characteristic of language. Respond with "yes" if the sentence describes novel knowledge about language; otherwise, respond with "no." Use only a yes or no format for your answers.
Artifact	a-task	Central to NLP research are tasks such as machine translation, named entity recognition, sentiment classification, and more. Your task is to assess if the given sentence from an NLP research paper introduces, or proposes a new or novel NLP task. This new task could either build upon existing NLP tasks or could be entirely novel. Respond with "yes" if the sentence introduces, or proposes a new or novel NLP task; otherwise, respond with "no."
	a-method	Algorithms and NLP models such as RNNs, LSTMs or LLMs are indispensable for NLP Research. Your task is to assess if the provided sentence from an NLP research paper introducing, or proposing a new or novel such NLP model, algorithm, or technique. This new model could have been built on top of existing models or methods or could be a completely new model. Respond with "yes" if the sentence introduces or proposes a new or novel NLP model; otherwise, respond with "no."
	a-dataset	Datasets constitute a crucial aspect of NLP research. Your task is to assess whether the given sentence from an NLP research paper introduces or discusses a new or novel NLP dataset. Respond with "yes" if it does; otherwise, respond with "no."

Table 11: Prompts to identify different types of contributions from NLP Research papers using LLMs.