# Exploring the Landscape of Natural Language Processing Research

**Tim Schopf, Karim Arabi, and Florian Matthes**
Technical University of Munich, Department of Computer Science, Germany
{tim.schopf,karim.arabi,matthes}@tum.de

## Abstract

As an efficient approach to understand, generate, and process natural language texts, research in natural language processing (NLP) has exhibited a rapid spread and wide adoption in recent years. Given the increasing research work in this area, several NLP-related approaches have been surveyed in the research community. However, a comprehensive study that categorizes established topics, identifies trends, and outlines areas for future research remains absent. Contributing to closing this gap, we have systematically classified and analyzed research papers in the ACL Anthology. As a result, we present a structured overview of the research landscape, provide a taxonomy of fields of study in NLP, analyze recent developments in NLP, summarize our findings, and highlight directions for future work. [1]

## 1 Introduction

Natural language is a fundamental aspect of human communication and inherent to human utterances and information sharing. Accordingly, most human-generated digital data are composed in natural language. Given the ever-increasing amount and importance of digital data, it is not surprising that computational linguists have started developing ideas on enabling machines to understand, generate, and process natural language since the 1950s (Hutchins, 1999).

More recently, the introduction of the transformer model (Vaswani et al., 2017) and pretrained language models (Radford and Narasimhan, 2018; Devlin et al., 2019) have sparked increasing interest in natural language processing (NLP). Submissions on various NLP topics and applications are being published in a growing number of journals and conferences, such as TACL, ACL, and EMNLP, as well as in several smaller workshops that focus on specific areas. Thereby, the ACL Anthology[2] as a repository for publications from many major NLP journals, conferences, and workshops emerges as an important tool for researchers. As of January 2023, it provides access to over 80,000 articles published since 1952. Figure 1 shows the distribution of publications in the ACL Anthology over the 50-year observation period.

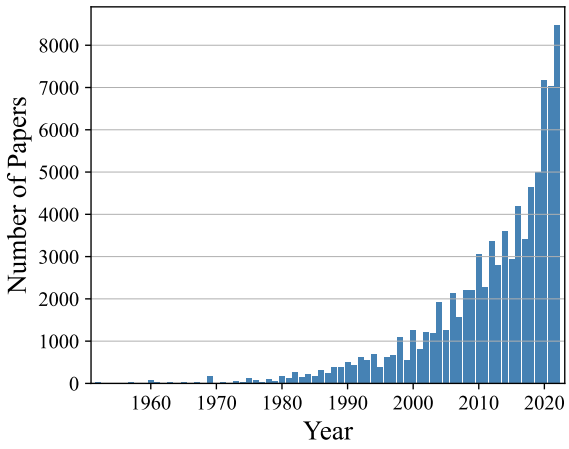


Figure 1: Distribution of the number of papers per year in the ACL Anthology from 1952 to 2022.

Accompanying the increase in publications, there has also been a growth in the number of different fields of study (FoS) that have been researched within the NLP domain. FoS are academic disciplines and concepts that usually consist of (but are not limited to) tasks or techniques (Shen et al., 2018). Given the rapid developments in NLP research, obtaining an overview of the domain and maintaining it is difficult. As such, collecting insights, consolidating existing results, and presenting a structured overview of the field is important. However, to the best of our knowledge, no stud-

[1] Code available: https://github.com/sebischair/Exploring-NLP-Research

[2] https://aclanthology.org



*Proceedings of Recent Advances in Natural Language Processing*, pages 1034–1045
Varna, Sep 4–6, 2023
https://doi.org/10.26615/978-954-452-092-2_111

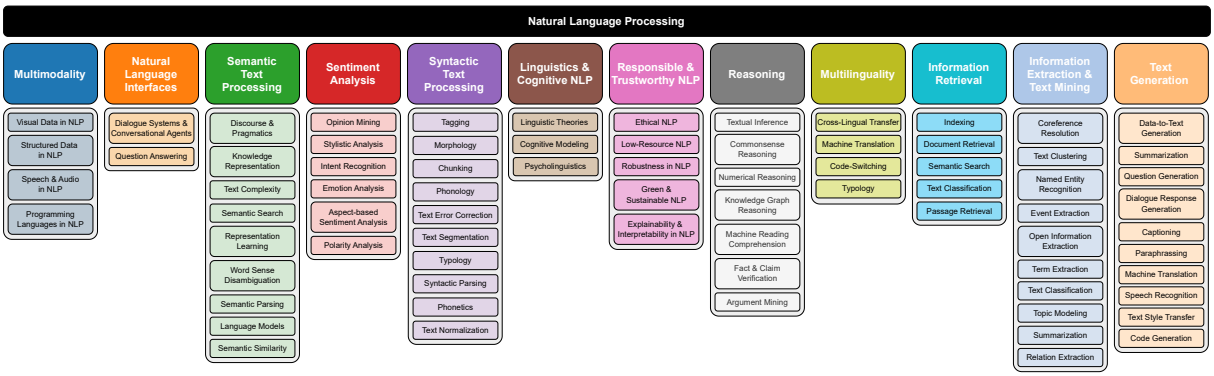


Figure 2: Taxonomy of fields of study in NLP.

ies exist yet that offer an overview of the entire landscape of NLP research. To bridge this gap, we performed a comprehensive study to analyze all research performed in this area by classifying established topics, identifying trends, and outlining areas for future research. Our three main contributions are as follows:

- We provide an extensive taxonomy of FoS in NLP research shown in Figure 2.
- We systematically classify research papers included in the ACL Anthology and report findings on the development of FoS in NLP.
- We identify trends in NLP research and highlight directions for future work.

Our study highlights the development and current state of NLP research. Although we cannot fully cover all relevant work on this topic, we aim to provide a representative overview that can serve as a starting point for both NLP scholars and practitioners. In addition, our analysis can assist the research community in bridging existing gaps and exploring various FoS in NLP.

## 2 Related Work

Related literature that considers various different FoS in NLP is relatively scarce. Most studies focus only on a particular FoS or sub-field of NLP research.

For example, related studies focus on knowledge graphs in NLP (Schneider et al., 2022), explainability in NLP (Danilevsky et al., 2020), ethics and biases in NLP (Šuster et al., 2017; Blodgett et al., 2020), question answering (Liu et al., 2022b), or knowledge representations in language models (Safavi and Koutra, 2021).

Studies that analyze NLP research based on the entire ACL Anthology focus on citation analyses (Mohammad, 2020a; Rungta et al., 2022) or visualizations of venues, authors, and n-grams and keywords extracted from publications (Mohammad, 2020b; Parmar et al., 2020).

Anderson et al. (2012) apply topic modeling to identify different epochs in the ACL's history.

Various books categorize different FoS in NLP, focusing on detailed explanations for each of these categories (Allen, 1995; Manning and Schütze, 1999; Jurafsky and Martin, 2009; Eisenstein, 2019; Tunstall et al., 2022).

## 3 Research Questions

The goal of our study is an extensive analysis of research performed in NLP by classifying established topics, identifying trends, and outlining areas for future research. These objectives are reflected in our research questions (RQs) presented as follows:

**RQ1:** *What are the different FoS investigated in NLP research?*

Although most FoS in NLP are well-known and defined, there currently exists no commonly used taxonomy or categorization scheme that attempts to collect and structure these FoS in a consistent and understandable format. Therefore, getting an overview of the entire field of NLP research is difficult, especially for students and early career researchers. While there are lists of NLP topics in conferences and textbooks, they tend to vary considerably and are often either too broad or too specialized. To classify and analyze developments in NLP, we need a taxonomy that encompasses a wide range of different FoS in NLP. Although this taxonomy may not include all possible NLP concepts, it needs to cover a wide range of the most popu-

lar FoS, whereby missing FoS may be considered as subtopics of the included FoS. This taxonomy serves as an overarching classification scheme in which NLP publications can be classified according to at least one of the included FoS, even if they do not directly address one of the FoS, but only subtopics thereof.

**RQ2:** *How to classify research publications according to the identified FoS in NLP?*

Classifying publications according to the identified FoS in NLP is very tedious and time-consuming. Especially with a large number of FoS and publications, a manual approach is very costly. Therefore, we need an approach that can automatically classify publications according to the different FoS in NLP.

**RQ3:** *What are the characteristics and developments over time of the research literature in NLP?*

To understand past developments in NLP research, we examine the evolution of popular FoS over time. This will allow a better understanding of current developments and help contextualize them.

**RQ4:** *What are the current trends and directions of future work in NLP research?*

Analyzing the classified research publications allows us to identify current research trends and gaps and predict possible future developments in NLP research.

## 4 Classification & Analysis

In this section, we report the approaches and results of the data classification and analysis. It is structured according to the formulated RQs.

### 4.1 Taxonomy of FoS in NLP research (RQ1)

To develop the taxonomy of FoS in NLP shown in Figure 2, we first examined the submission topics of recent years as listed on the websites of major NLP conferences such as ACL, EMNLP, COLING, or IJCNLP. In addition, we reviewed the topics of workshops included in the ACL Anthology to derive further FoS. In order to include smaller topics that are not necessarily mentioned on conference or workshop websites, we manually reviewed all papers from the recently published EMNLP 2022 Proceedings, extracted their FoS, and annotated all 828 papers accordingly. This provided us with an initial set of FoS, which we used to create the first version of the NLP taxonomy. Based on our initial taxonomy, we conducted semi-structured expert interviews with NLP researchers to evaluate and adjust the taxonomy. In the interviews, we placed particular emphasis on the evaluation of the mapping of lower-level FoS to their higher-level FoS, and the correctness and completeness of FoS in the NLP domain. In total, we conducted more than 20 one-on-one interviews with different domain experts. After conducting the interviews, we noticed that experts demonstrated a high degree of agreement on certain aspects of evaluation, while opinions were highly divergent on other aspects. While we easily implemented changes resulting based on high expert agreement, we acted as the final authority in deciding whether to implement a particular change for aspects with low expert agreement. For example, one of the aspects with the highest agreement was that certain lower-level FoS must be assigned not only to one but also to multiple higher-level FoS. Based on the interview results, we subsequently adjusted the annotations of the 828 EMNLP 2022 papers and developed the final NLP-taxonomy, as shown in Figure 2.

### 4.2 Field of Study Classification (RQ2)

We trained a weakly supervised classifier to classify ACL Anthology papers according to the NLP

| **Dataset** → | *Validation* | | | *Test* | | |
|---|---|---|---|---|---|---|
| **Model** ↓ | **P** | **R** | **F**$_1$ | **P** | **R** | **F**$_1$ |
| BERT | **96.57**±**0.14** | 95.43±0.16 | 96.00±0.03 | 89.77±0.20 | 93.58±0.07 | 91.64±0.10 |
| RoBERTa | 95.77±0.19 | 95.19±0.16 | 95.48±0.17 | 87.46±2.75 | 93.29±0.10 | 90.27±1.42 |
| SciBERT | 96.44±0.17 | 95.65±0.14 | 96.05±0.10 | 90.18±3.17 | **94.05**±**0.06** | 92.06±1.65 |
| SPECTER 2.0 | 96.44±0.11 | 95.69±0.14 | **96.06**±**0.08** | **92.46**±**2.58** | 93.99±0.22 | **93.21**±**1.39** |
| SciNCL | 96.39±0.11 | **95.71**±**0.09** | 96.05±0.04 | 89.97±1.85 | 93.74±0.18 | 91.81±0.93 |

Table 1: Evaluation results for classifying papers according to the NLP taxonomy on three runs over different random train/validation splits. Since the distribution of classes is very unbalanced, we report micro scores.

| Field of Study | # Papers | Representative Papers | Field of Study | # Papers | Representative Papers |
|---|---|---|---|---|---|
| Machine Translation | 12,922 | Liu et al. (2020), Goyal et al. (2022) | Visual Data in NLP | 2,401 | Tan and Bansal (2019), Xu et al. (2021) |
| Language Models | 11,005 | Devlin et al. (2019), Ouyang et al. (2022) | Ethical NLP | 2,322 | Blodgett et al. (2020), Perez et al. (2022) |
| Representation Learning | 6,370 | Reimers and Gurevych (2019), Gao et al. (2021b) | Question Answering | 2,208 | Karpukhin et al. (2020), Liu et al. (2022b)) |
| Text Classification | 6,117 | Wei and Zou (2019), Hu et al. (2022) | Tagging | 1,968 | Malmi et al. (2019), Wei et al. (2020) |
| Low-Resource NLP | 5,863 | Gao et al. (2021a), Liu et al. (2022a) | Summarization | 1,856 | Liu and Lapata (2019), He et al. (2022) |
| Dialogue Systems & Conversational Agents | 4,678 | Zhang et al. (2020), Roller et al. (2021) | Green & Sustainable NLP | 1,780 | Strubell et al. (2019), Ben Zaken et al. (2022) |
| Syntactic Parsing | 4,028 | Zhou and Zhao (2019), Glavaš and Vulić (2021) | Cross-Lingual Transfer | 1,749 | Conneau et al. (2020), Feng et al. (2022) |
| Speech & Audio in NLP | 3,915 | Baevski et al. (2022), Wang et al. (2020) | Morphology | 1,749 | McCarthy et al. (2020), Goldman et al. (2022) |
| Knowledge Representation | 2,967 | Schneider et al. (2022), Safavi and Koutra (2021) | Explainability & Interpretability in NLP | 1,671 | Danilevsky et al. (2020), Pruthi et al. (2022) |
| Structured Data in NLP | 2,803 | Herzig et al. (2020), Yin et al. (2020) | Robustness in NLP | 1,621 | Hendrycks et al. (2020), Meade et al. (2022) |

Table 2: Overview of the most popular FoS in NLP literature. Representative papers consist of either highly cited studies or comprehensive surveys on the respective FoS.

taxonomy. To obtain a training dataset, we first defined keywords for each FoS included in the final taxonomy to perform a database search for relevant articles. Based on the keywords, we created search strings to query the Scopus and arXiv databases. The search string was applied to titles and author keywords, if available. While we limited the Scopus search results to the NLP domain with additional restrictive keywords such as "NLP", "natural language processing", or "computational linguistics", we limited the search in arXiv to the cs.CL domain. We subsequently merged duplicate articles to create a multi-label dataset and removed articles included in the EMNLP 2022 proceedings, as this dataset is used as test set. Finally, we applied a fuzzy string matching heuristic and added missing classes based on the previously defined FoS keywords that appear twice or more in the article titles or abstracts. The final training dataset consists of 178,521 articles annotated on average with 3.65 different FoS. On average, each class includes 7936.50 articles, while the most frequent class is represented by 63728 articles and the least frequent class by 141 articles. We split this unevenly distributed dataset into three different random 90/10 training/validation sets and used the human-annotated EMNLP 2022 articles as the test dataset.

For multi-label classification, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), SciBERT (Beltagy et al., 2019), SPECTER 2.0 (Cohan et al., 2020; Singh et al., 2022), and SciNCL (Ostendorff et al., 2022) models were fine-tuned in their base versions on the three different training datasets and evaluated on their respective validation and test datasets. We trained all models for three epochs, using a batch size of 8, a learning rate of $5e-5$, and the AdamW optimizer (Loshchilov and Hutter, 2019).

The evaluation results are shown in Table 1. SPECTER 2.0 shows significant performance on both validation and test data. Therefore, we selected SPECTER 2.0 as our final classification model, which we subsequently trained with the same parameters on the combined training, validation, and test data. Using the final model, we classified all papers included in the ACL Anthology from 1952 to 2022. To obtain our final dataset for analysis, we removed the articles that were not truly research articles, such as prefaces; articles that were not written in English; and articles where the classifier was uncertain and simply predicted every class possible. This final classified dataset includes a total of 74,279 research papers. Table 2 shows the final classification results with respect to the number of publications for each of the most popular FoS.

### 4.3 Characteristics and Developments of the Research Landscape (RQ3)

Considering the literature on NLP, we start our analysis with the number of studies as an indicator of research interest. The distribution of publications over the 50-year observation period is

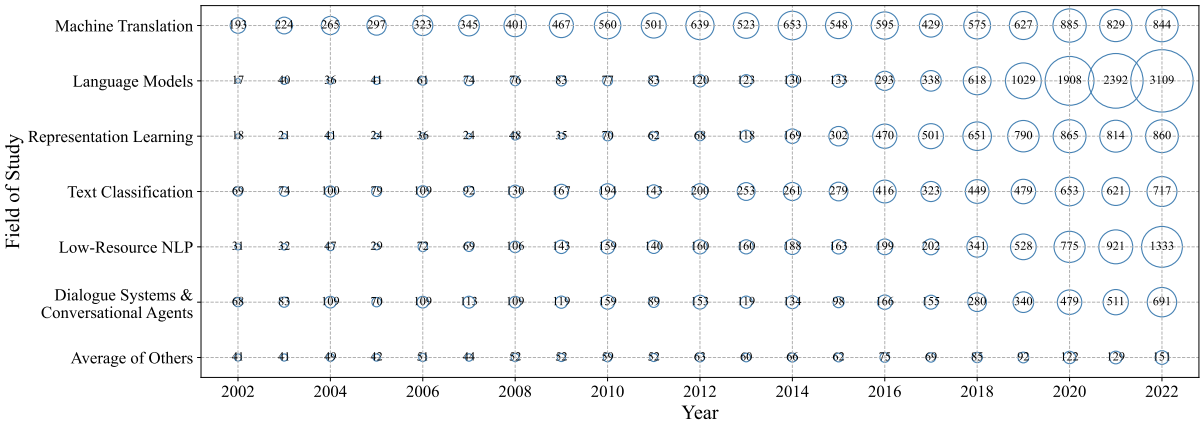


Figure 3: Distribution of number of papers by most popular FoS from 2002 to 2022.

shown in Figure 1. While the first publications appeared in 1952, the number of annual publications grew slowly until 2000. Accordingly, between 2000 and 2017, the number of publications roughly quadrupled, whereas in the subsequent five years it has doubled again. We therefore observe a near-exponential growth in the number of NLP studies, indicating increasing attention from the research community.

Examining Table 2 and Figure 3, the most popular FoS in the NLP literature and their recent development over time are revealed. While the majority of studies in NLP are related to machine translation or language models, the developments of both FoS are different. Machine translation is a thoroughly researched field that has been established for a long time and has experienced a modest growth rate over the last 20 years. Language models have also been researched for a long time. However, the number of publications on this topic has only experienced significant growth since 2018. Similar differences can be observed when looking at the other popular FoS. Representation learning and text classification, while generally widely researched, are partially stagnant in their growth. In contrast, dialogue systems & conversational agents and particularly low-resource NLP continue to exhibit high growth rates in the number of studies. Based on the development of the average number of studies on the remaining FoS in Figure 3, we observe a slightly positive growth overall. However, the majority of FoS are significantly less researched than the most popular FoS. We conclude that the distribution of research across FoS is extremely unbalanced and that the development of NLP research is largely shaped by advances in a few highly popular FoS.

### 4.4 Research Trends and Directions for Future Work (RQ4)

Figure 4 shows the growth-share matrix of FoS in NLP research inspired by Henderson (1970). We use it to examine current research trends and possible future research directions by analyzing the growth rates and total number of papers related to the various FoS in NLP between 2018 and 2022. The upper right section of the matrix consists of FoS that exhibit a high growth rate and simultaneously a large number of papers overall. Given the growing popularity of FoS in this section, we categorize them as *trending stars*. The lower right section contains FoS that are very popular but exhibit a low growth rate. Usually, these are FoS that are essential for NLP research but already relatively mature. Hence, we categorize them as *foundational FoS*. The upper left section of the matrix contains FoS that exhibit a high growth rate but only very few papers overall. Since the progress of these FoS is rather promising, but the small number of overall papers renders it difficult to predict their further developments, we categorize them as *rising question marks*. The FoS in the lower left of the matrix are categorized as *niche FoS* owing to their low total number of papers and their low growth rates.

Figure 4 shows that language models are currently receiving the most attention, which is also consistent with the observations from Table 2 and Figure 3. Based on the latest developments in this area, this trend is likely to continue and accelerate in the near future. Text classification, machine translation, and representation learning rank among the most popular FoS, but only show marginal growth. In the long term, they may be replaced by faster-growing fields as the most popular FoS.

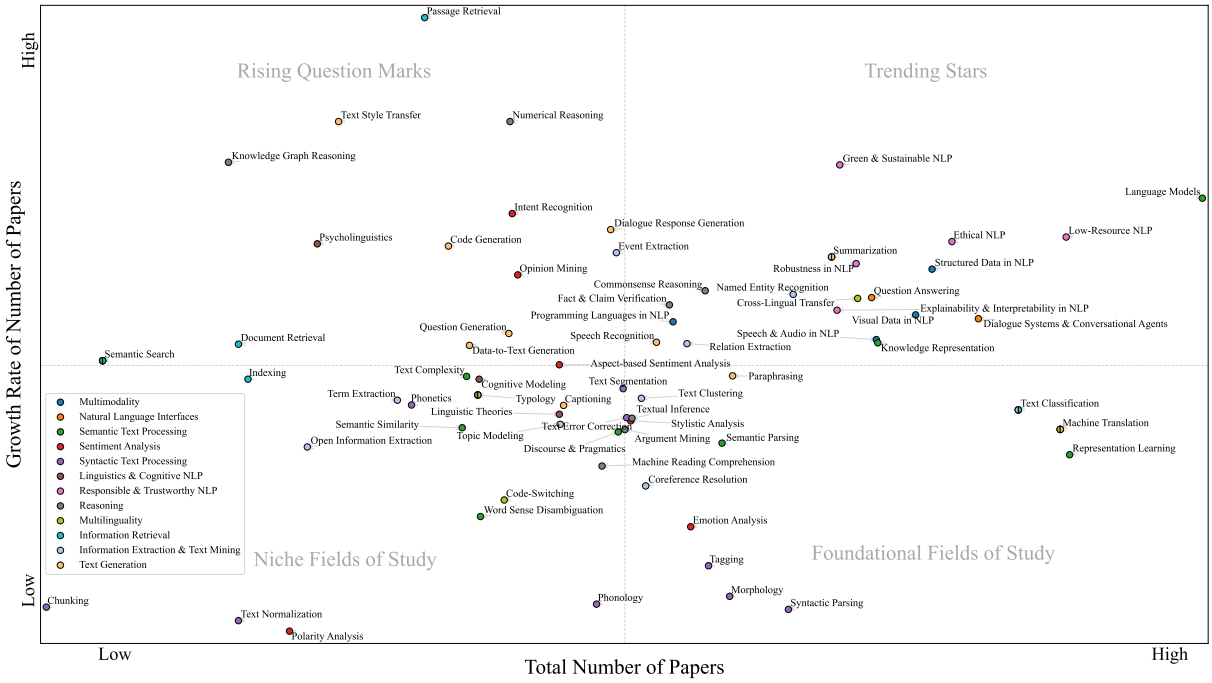


Figure 4: Growth-share matrix of FoS in NLP. The growth rates and total number of works for each FoS are calculated from the start of 2018 to the end of 2022. To obtain a more uniform distribution of the data, we apply the Yeo-Johnson transformation (Yeo and Johnson, 2000).

In general, FoS related to syntactic text processing exhibit negligible growth and low popularity overall. Conversely, FoS concerned with responsible & trustworthy NLP, such as green & sustainable NLP, low-resource NLP, and ethical NLP tend to exhibit a high growth rate and also high popularity overall. This trend can also be observed in the case of structured data in NLP, visual data in NLP, and speech & audio in NLP, all of which are concerned with multimodality. In addition, natural language interfaces involving dialogue systems & conversational agents and question answering are becoming increasingly important in the research community. We conclude that in addition to language models, responsible & trustworthy NLP, multimodality, and natural language interfaces are likely to characterize the NLP research landscape in the near future.

Further notable developments can be observed in the area of reasoning, specifically with respect to knowledge graph reasoning and numerical reasoning and in various FoS related to text generation. Although these FoS are currently still relatively small, they apparently attract more and more interest from the research community and show a clear positive tendency toward growth.

## 5 Discussion

The observations of our comprehensive study reveal several insights that we can situate to related work. Since the first publications in 1952, researchers have paid increasing attention to the field of NLP, particularly after the introduction of Word2Vec (Mikolov et al., 2013) and accelerated by BERT (Devlin et al., 2019). This observed growth in research interest is in line with the study of Mohammad (2020b). Historically, machine translation was one of the first research fields in NLP (Jones, 1994), which continues to be popular and steadily growing nowadays. However, recent advances in language model training have sparked increasing research efforts in this field, as shown in Figure 3 and Figure 4. Since scaling up language models significantly enhance performance on downstream tasks (Brown et al., 2020; Kaplan et al., 2020; Wei et al., 2022a; Hoffmann et al., 2022), researchers continue to introduce increasingly larger language models (Han et al., 2021). However, training and using these large language models involves significant challenges, including computational costs (Narayanan et al., 2021), environmental issues (Strubell et al., 2019), and ethical considerations (Perez et al., 2022). As a result, a recent increase in research efforts has been noted

to render language models and NLP more responsible & trustworthy in general, as shown in Figure 4. Additionally, recent advances aim to train large-scale multimodal language models capable of understanding and generating natural language text and performing all types of downstream tasks while interacting with humans through natural language input prompts (OpenAI, 2023). From our observations in Figure 4, we again find support for this trend in NLP literature for multimodality, text generation, and natural language interfaces.

Although language models have achieved remarkable success on various NLP tasks, their inability to reason is often seen as a limitation that cannot be overcome by increasing the model size alone (Rae et al., 2022; Wei et al., 2022b; Wang et al., 2023). Although reasoning capabilities are a crucial prerequisite for the reliability of language models, this field is still relatively less researched and receives negligible attention. While Figure 4 exhibits high growth rates for knowledge graph reasoning and numerical reasoning in particular, research related to reasoning is still rather underrepresented compared to the more popular FoS.

## 6 Conclusion

Recent years have witnessed an increasing prominence of NLP research. To summarize recent developments and provide an overview of this research area, we defined a taxonomy of FoS in NLP and analyzed recent research developments.

Our findings show that a large number of FoS have been studied, including trending fields such as multimodality, responsible & trustworthy NLP, and natural language interfaces. While recent developments are largely a result of recent advances in language models, we have noted a lack of research pertaining to teaching these language models to reason and thereby afford more reliable predictions.

## 7 Limitations

Constructing the taxonomy highly depends on the personal decisions of the authors, which can bias the final result. The taxonomy may not cover all possible FoS and offers potential for discussions, as domain experts have inherently different opinions. As a countermeasure, we aligned the opinions of multiple domain experts and designed the taxonomy at a higher level, allowing non-included FoS to be considered as possible subtopics of existing ones.

For this study, we limited our analysis to papers published in the ACL Anthology, which typically feature research presented at major international conferences and are written in English. However, research communities that publish their work in regional venues exist, often in languages other than English. In addition, NLP research is also presented at other prominent global conferences such as AAAI, NeurIPS, ICLR, or ICML. Therefore, the findings we report in this study pertain specifically to NLP research presented at major international conferences and journals in English.

Furthermore, the accuracy of the classification results poses another threat to the validity of our study. Data extraction bias and classification model errors may negatively affect the results. To mitigate this risk, the authors regularly discussed the used classification schemes and conducted a thorough evaluation of the performance of the classification model.

## Acknowledgments

We would like to thank Phillip Schneider, Stephen Meisenbacher, Mahdi Dhaini, Juraj Vladika, Oliver Wardas, Anum Afzal, Wessel Poelman, and Alexander Blatzheim of sebis for helpful discussions and valuable feedback.

## References

James Allen. 1995. *Natural Language Understanding*. Benjamin Cummings.

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.

Alexei Baevski, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2022. Unsupervised speech recognition.

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.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. In

*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. 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.

Jacob Eisenstein. 2019. *Introduction to natural language processing*. MIT press.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making pre-trained language models better few-shot learners. 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 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3090–3104, Online. Association for Computational Linguistics.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 864–870, Dublin, Ireland. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*,

pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bruce Henderson. 1970. The product portfolio.

Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. 2020. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Jingang Wang, Juanzi Li, Wei Wu, and Maosong Sun. 2022. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2225–2240, Dublin, Ireland. Association for Computational Linguistics.

John Hutchins. 1999. Retrospect and prospect in computer-based translation. In *Proceedings of Machine Translation Summit VII*, pages 30–36, Singapore, Singapore.

Karen Sparck Jones. 1994. Natural language processing: a historical review. *Current issues in computational linguistics: in honour of Don Walker*, pages 3–16.

Daniel Jurafsky and James H. Martin. 2009. *Speech and language processing*, 2. ed., [pearson international edition] edition. Prentice Hall series in artificial intelligence. Prentice Hall, Pearson Education International, London [u.a.].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022a. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2022b. Challenges in generalization in open domain question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2014–2029, Seattle, United States. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Eric Malmi, Sebastian Krause, Sascha Rothe, Daniil Mirylenka, and Aliaksei Severyn. 2019. Encode, tag, realize: High-precision text editing. 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 5054–5065, Hong Kong, China. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA.

Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena

Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3922–3931, Marseille, France. European Language Resources Association.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

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: An interactive visual explorer for natural language processing literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 232–255, Online. Association for Computational Linguistics.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia. 2021. Efficient large-scale language model training on GPU clusters. *CoRR*, abs/2104.04473.

OpenAI. 2023. Gpt-4 technical report.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11670–11688, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Monarch Parmar, Naman Jain, Pranjali Jain, P. Jayakrishna Sahit, Soham Pachpande, Shruti Singh, and Mayank Singh. 2020. Nlpexplorer: Exploring the universe of nlp papers. In *Advances in Information Retrieval*, pages 476–480, Cham. Springer International Publishing.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Danish Pruthi, Rachit Bansal, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary C. Lipton, Graham Neubig, and William W. Cohen. 2022. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Mukund Rungta, Janvijay Singh, Saif M. Mohammad, and Diyi Yang. 2022. Geographic citation gaps in NLP research. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1371–1383, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tara Safavi and Danai Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1053–1067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. 2022. A decade of knowledge graphs in natural language processing: A survey. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 601–614, Online only. Association for Computational Linguistics.

Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A web-scale system for scientific knowledge exploration. In *Proceedings of ACL 2018, System Demonstrations*, pages 87–92, Melbourne, Australia. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *ArXiv*, abs/2211.13308.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Simon Šuster, Stéphan Tulkens, and Walter Daelemans. 2017. A short review of ethical challenges in clinical natural language processing. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 80–87, Valencia, Spain. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

L. Tunstall, L. von Werra, and T. Wolf. 2022. *Natural Language Processing with Transformers*. O'Reilly Media.

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.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq S2T: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39, Suzhou, China. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh 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. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022b. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. 2020. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, Online. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. 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 2579–2591, Online. Association for Computational Linguistics.

In-Kwon Yeo and Richard A. Johnson. 2000. A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4):954–959.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Junru Zhou and Hai Zhao. 2019. Head-Driven Phrase Structure Grammar parsing on Penn Treebank. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2396–2408, Florence, Italy. Association for Computational Linguistics.