Dataset quality within a societally impactful machine learning domain

An overview of data collection and annotation practices of the datasets used by papers published by the ACL

Alexandru Fazakas[1]

Supervisor(s): dr. Cynthia Liem[1]**, Andrew M. Demetriou**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Alexandru Fazakas
Final project course: CSE3000 Research Project
Thesis committee: dr. Cynthia Liem, Andrew M. Demetriou, dr. Jie Yang

Abstract

This study gives an overview of the data collection and annotation practices of the datasets used by the most impactful papers published by the Association of Computational Linguistics (ACL). This was achieved by selecting the most highly cited papers published within the ACL anthology across 3 periods (published in the past 2, 5 and 15 years). Afterwards, the datasets used by those papers were extracted and filtered to retain the most impactful ones. Finally, a carefully crafted annotation schema was used to find out information regarding key aspects of the datasets in order to qualitatively analyze them. As a result of this analysis, it was first found that (1) there are fewer datasets used on average in the past 2 years and that there is little overlap with the datasets used by papers published in the past 5 or 15 years. (2) Secondly, there are various concerns related to those key aspects, such as the relatively high (∼36%) and unregulated use of the Amazon Mechanical Turk crowdsourcing platform for the construction of datasets. Another concern is information frequently missing about any rationale regarding labeller population, prescreening, inter-rater reliability and rationale regarding sample size - missing ∼77%, ∼63%, ∼19-56%, and ∼81% of the time. However, reporting practices for most of those issues have slightly improved within datasets used in the past 2 years. (3) Finally, around one third of the information sought was missing across all periods. However, the state of the domain has been generally improving, with a lower one fourth of the information missing from datasets used in the past 2 years. Some recommendations are given in order to overcome those challenges, the most important of which being that each academic organization should require their submissions to include a reporting template in their papers.

1 Introduction

Dataset quality is crucial in order to train successful machine learning (ML) models. The quality of a dataset depends not only on following data collection best practices, but also on ensuring, when applicable, that the ”ground truth” labels of that data are of high quality. The reason dataset quality is so important is because most ML models require a dataset to be trained and then tested on for the task they were designed for. A dataset of low quality would therefore result in a model that underperforms - a phenomenon known as ”garbage in, garbage out”. This is supported by a survey done by (Jain et al., 2020), in which the importance of good data is emphasized, highlighting the necessity of analyzing data quality in terms of its value for machine learning applications.

Unfortunately, previous literature such as (Thyagarajan et al., 2022; Geiger et al., 2021, 2020; Liem et al., 2024) gives reason to be skeptical about the dataset quality of ML papers in a couple domains. (Thyagarajan et al., 2022) argue that labeling errors are fairly commonplace, demonstrating the many errors in the CelebA datasets. (Geiger et al., 2020) survey the annotation practices of machine learning papers that perform classification tasks on twitter data. They found out that, on average, those papers have a normalized information score of 0.441. This score means more than half of the important information regarding annotation procedures was not mentioned. (Geiger et al., 2021) follow up on the work of the earlier paper, sampling 200 papers from 3 different domains, and using a similar method to survey the annotation practices. They find out roughly the same overall results, with some variance between the domains. Finally, (Liem et al., 2024) trace the datasets used in the top 5 most highly cited papers published in the International Conference on Acoustics, Speech, and Signal Processing. Their purpose was to probe the initial origins of the datasets, in order to make a statement about their provenance and quality. They find ”disbalances and unclear origins for the datasets used”, which threatens ”validity and integrity of outcomes trained on these datasets”.

Other reasons to be skeptical of dataset quality include the difficulty of creating good ground truth labels and the many pitfalls of data selection. (Aroyo and Welty, 2015) give an overview of 7 common ”myths” of annotating data, highlighting how easy it is for one to believe them. (Hullman et al., 2022) draw a parallel between social sciences and machine learning, highlighting the various challenges of data collection, such as non-representative samples, bias and non-desirable data.

Bad quality datasets negatively contribute towards the reproducibility crisis currently looming over the machine learning domain. Sampling bias and issues with data quality are noted by (Kapoor and Narayanan, 2023) as one negative contributor towards the aforementioned crisis. (Semmelrock et al., 2025) also notes sharing reproducible data includes documenting its provenance and that limited access to data is another contributing factor to the crisis. Considering the evidence so far, this may be a systemic issue, but only a couple previously mentioned works focus on this issue. As a result, more literature analyzing the state of datasets used by ML models in a societally impactful domain would prove very valuable.

This paper aims to therefore answer the following research question: ”What are the data collection and annotation practices of the datasets present in the most impactful papers of the Association of Computational Linguistics (ACL)?”. The reason for choosing papers within this academic organization was threefold: (1) because papers within this organization tend to have a high impact[1], (2) among other academic organizations, it had the most well documented annotation practices according to Geiger et al. (2021), therefore representing a ”golden standard” within the field and (3) the ACL handles papers in the domain of computational linguistics and natural language processing[2] (two domains with a lot of over-

[1]This can be observed on Google Scholar metrics, with The ”Meeting of the Association of Computational Linguistics” scoring on the 36th place as of the 4th of June 2025. https://scholar.google.com/citations?view_op=top_venues

[2]According to https://www.aclweb.org/portal/what-is-cl, ac-

lap, making analysis less demanding).

The main research question will be broken down in the following subquestions: (1) which datasets are most often used by those papers? what is the overlap across different time periods? (2) how well do the most used datasets report on data collection and annotation practices (if annotated)? do those practices change for more recently impactful datasets? (3) how much information related to annotation practices and data collection is missing from those datasets? does this vary based on when those datasets were used?

2 Methodology

In order to achieve the aim of this study Scopus was firstly used in order to extract papers accepted by the ACL - the details of this procedure are described in Subsection 2.1. Afterwards, all datasets used by those papers were collected and stored by paper that used them. This allowed an overlap calculation in order to answer research subquestion 1. Subsection 2.2 gives the details on those procedures. Finally, a subset of the extracted datasets were annotated and analyzed in order to answer research subquestions 2 and 3. The details on the selection procedure and annotation of the datasets are given in subsection 2.3.

2.1 Selecting the ACL papers

The first step of the analysis process was selecting a number of papers from the aforementioned organization in order to extract the datasets from each one. This was done using Scopus to filter by source title and sort by citation count to extract the first 25 papers for 3 different time periods. An explanation of the each decision will be detailed in the following paragraphs.

Many academic databases are readily available in order to find papers from a variety of sources, such as Google Scholar, Clarivate Web of Science and Scopus, to just name a few. Scopus was chosen because it is query based (therefore search strings are reproducible) and seems to cover the most amount of publications. It was also the only database used in order to avoid the need of deduplication and resolving different citation counts.

Three separate time frames have been selected: published in the past 2, 5 and 15 years (with an upper bound on end 2024). This was because older papers tend to have more citations simply because they have gotten the time to gather more citations. Therefore a period of 2 years was chosen to analyze recent papers, to find out the recently influential datasets as well. In order to only keep reasonably relevant papers, a period of 15 years was chosen. Finally, 5 years was chosen as a reasonable middle point between the two formerly mentioned periods for more diversity.

Citation count was selected as the order of selection for papers, as the objective was focusing on the datasets used in the most influential academic works. The papers that have been selected can be found in ”Tab 1” of the spreadsheet described in Appendix B. The top 25 papers of each period have been selected, as it was believed that this limit would keep the amount of work manageable, while still giving a good overview of the most impactful publications from that period. The precise queries with date executed are available in appendix A. An overview of the papers, along with their citations can be seen in Appendix E.

2.2 Dataset collection and overlap calculation

In order to extract the datasets used by the papers selected in Subsection 2.1, each paper was read and each dataset used was added to the list, with a reference to the paper that mentioned it. Then, the datasets were aggregated by the period of the paper which mentioned them, with some general statistics regarding datasets per period also calculated. Finally, Jaccard and cosine similarity was used to analyze the overlap of the datasets.

There were a couple of rules that were imposed while extracting the datasets. When reading (especially older) papers, datasets are sometimes not given a reference to or are given a reference to a completely different dataset. In such cases, it was attempted to find the dataset that the authors were referring to by searching for it on the internet for a reasonable match (NOTE: datasets that were mentioned but not used were not included). Another rule that was imposed to keep the research work within reason was leaving disproportionately highly cited papers with more than 10 datasets for qualitative analysis. This was because including so many datasets from a heavily cited paper would skew the analysis of the datasets used in a certain period in the next step of the process. The analysis would be skewed since only the datasets of that paper would be analyzed for that period, so diversity would be compromised. Fortunately, such a case did not occur.

As a link was maintained between each dataset and the original ACL paper the referenced it, each dataset mention could be tied to one of the periods of the papers. Using this, the datasets could be aggregated by period. Then, some general statistics regarding datasets per period and overall were calculated. Those were the average datasets per paper, the amount of unique datasets and the amount of total datasets.

With usage of the datasets calculated by period, cosine and Jaccard similarity were chosen to calculate dataset overlap across periods. Cosine similarity was chosen to take the count of the datasets into account as well, and Jaccard was used in order to see what the overlap is without the counts.

2.3 Dataset selection, annotation and analysis

After the datasets were collected, a subset was selected in order to be analyzed. It was decided to analyze the top 20 most influential datasets per period for a total of 60, with the selection procedure detailed in 2.3. In order to analyze the datasets, an annotation schema was first used to extract specific details of interest. Then, those details were aggregated an analyzed in order to find out the answers to the second and third research subquestion.

The dataset selection procedure Since 214 unique datasets were used for all papers selected (more details in Section 3.1), analyzing them all would not be feasible within the allocated time. Therefore, only the most influential datasets from each period were selected for analysis. Initially, it was attempted to sort datasets used within period X by how

cessed 4th of June, 2025

many times they were mentioned by papers within that period. However, such an approach would not include less used datasets that were used by more cited papers.

In order to solve this, a different approach was used, by using "citation sum" instead. It was used to rank each dataset within a certain period by summing up the citations of each paper that mentioned it. The precise formula can be seen in the appendix, under Formula 2. This way, a balance was struck between how many times a dataset was mentioned, but also by which paper it was mentioned. The datasets selected, along with their citation sum and citations can be seen in Appendix F. The precise way the calculations were done for the citation sum for each paper can be seen in Appendix B.

Dataset annotation An annotation schema was created in collaboration with other peers in order to have a structured way to analyze each dataset. Then, the schema was used in a collaborative effort to annotate each of the selected datasets.

The use of an annotation schema in order to analyze the datasets was inspired by the structured content analysis method, a longstanding way to analyze content originating from social sciences. This method is used to turn qualitative content (in this case the information about the datasets) to qualitative data so that it can be easily analyzed. This is exactly what was needed for this analysis.

The creation of the annotation schema was collaborative among peers that worked on a similar annotation effort and is available in full in Appendix C. The questions or items present therein were inspired by the ones used by (Geiger et al., 2020, 2021), but also include some of our own (ours will be marked with an asterisk *). Reasons for the inclusion of the items will be given in the next paragraph, as the schema is presented.

The items in the annotation schema can be succinctly clustered in three categories: annotator transparency, annotation procedures and data collection. The same split will be maintained when the answers will be analyzed. The items/questions of the annotation schema are presented in the following paragraphs with brief justifications.

Within **annotator transparency** items, the schema captures various elements of utmost importance related to the annotators themselves. The reason those details are important is because the background of a person can significantly influence the way they annotate. The items include whether the labels came from a human, whether they were original (or external), what their source was*, why the population of labellers was selected*, whether there was prescreening and whether compensation was documented.

Items within **annotation procedures** verify the procedure of the annotation effort. Those items were selected to verify whether the annotation experiment was well designed, so that the best quality annotations can be extracted. The most important items include whether there was training and formal instructions, how many annotators were involved, how many annotated each item*, how many labels were required for each item*, whether there was overlap, if inter-rater reliability (IRR) was calculated, and if an annotation schema* along with a rationale* was mentioned.

Finally, **data collection** items focus on the items collected to be annotated. Those concerns are important as having good items is the prerequisite of creating a good dataset, and being mindful of the size shows consideration for the training goal. Items include whether the samples were described*, if a rationale for choosing the samples was given*, if the size was decided beforehand*, if there is a reason for the final size* and if there is a link towards the dataset.

The annotation of the datasets was mainly done by the author of this paper; however $\tilde{2}5\%$ were done by other students working on a similar type of paper, which were looking at a different academic organization. This is because datasets used by papers of the ACL happened to be used in the papers of other academic organizations as well. We have agreed on a common workflow in order to annotate the dataset papers, and each person's individual contribution were roughly equal.

All the annotators were students and there was arguably no direct prescreening of the annotators (including me) There was also no training before or throughout the annotation effort. There have been weekly and asynchronous discussions in order to decide what to do in case uncertainties arose. The "compensation" received for this annotation effort was indirectly the completion of the course within which this research project was created. Due to constraints of time and the scope of the project, there was no overlap of multiple annotators on the dataset papers, therefore no IRR could be calculated.

Dataset analysis After the datasets were labelled, they could finally be analyzed. As the second research subquestion was realated to the data collection and annotation practices of the datasets, various charts were created in order to illustrate them, excluding data that was not applicable. Afterwards, missing information was calculated in a similar fashion.

When calculating the percentage of some information regarding a field, the "Not applicable" answers were excluded. The reason this was done was because this would artificially inflate the amount of information given by a paper. The way the charts are generated is detailed in Appendix B.

The way the missing information percentage was calculated was by taking each of the 20 datasets per period, summing up all the columns where "No information" was present. "Not applicable" was excluded for the same reason mentioned in the paragraph above. The formula can be seen in the Appendix under Formula 1.

3 Findings

This section presents the statistical results related to the three research subquestions mentioned in the Introduction.

3.1 Dataset overlap

This section relates to findings about research subquestion 1: "Which datasets are most often used by those (ACL) papers? What is the overlap across periods?" Some general statistics are first presented. Afterwards, the top 20 most used datasets per period are presented. Finally, the overlap and similarity between across the 3 periods is presented - both for the top 20 datasets and all datasets of those respective periods.

Figure 1 shows some general statistics about the amount of datasets used per period and overall. It can be noticed that there are significantly less total datasets used in period 2 compared to the others, with a significantly smaller average

	2	5	15	overall
Avg datasets	3.478	7.05	4.25	4.821
Unique datasets	67	118	72	211
Total datasets	80	143	103	328
Unique/Total	0.838	0.825	0.699	0.643

Figure 1: Statistics related to the amount of datasets used by papers per period.

	5 - 15	2 - 5	2 - 15
overall cos	0.501	0.116	0.067
overall jaccard	0.224	0.057	0.037
overall common	35.0	10.0	5.0
top20 cos	0.525	0.069	0.0
top20 jaccard	0.29	0.053	0.0
top20 common	9.0	2.0	0.0

Figure 2: The overlap between the 3 periods.

of datasets per paper. It can also be seen that the datasets used in the past 15 years (period 15) were more frequently reused. Overall, papers used ∼4.8 datasets on average.

Figure 10 in the Appendix shows the top 20 datasets used by period. It can be seen that while the *SQuAD* (Stanford question answering dataset, (Rajpurkar et al., 2016a)) dataset is not the most used in all periods, it is the most used overall, with 9/75 papers using it and also 3/75 (not necessarily other) papers using *SQuADv2* (Rajpurkar et al., 2018a). The *GLUE* (General language understanding evaluation, (Wang et al., 2018a)) benchmark is also used very frequently (6/75), and it also contains *squad*. Both *squad* and *glue* are used explicitly in only two papers, therefore 13/75 - 17.3% papers use *squad* directly or indirectly.

The lack of overlap between period 2 and the other two is also noticeable in Figure 10. To highlight this, the Jaccard and cosine similarity among those top 20 and also overall are presented in Figure 2. Both overall and in the top 20, datasets used in the past 5 and 15 have around ∼0.5 cosine similarity, but the overlap between period 2 and either 5 or 15 is really low. However, datasets used in the past 5 and 15 years have a lower Jaccard similarity. This could be explained by the lower amount of unique datasets used by papers in the past 15 years, compared to the ones in the past 5 years, as evidenced by Figure 1.

3.2 Data collection and annotation practices

In order to interpret the data collection and annotation practices of the datasets chosen and answer research subquestion 2, the results of the items in the annotation schema will be analyzed, with changes in the past 2 year being mentioned. This is in accordance with the procedures described in Section 2.3, with each cluster of the items each being mentioned in their

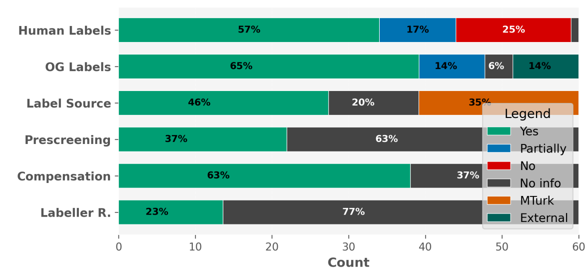

Figure 3: Annotator transparency results.

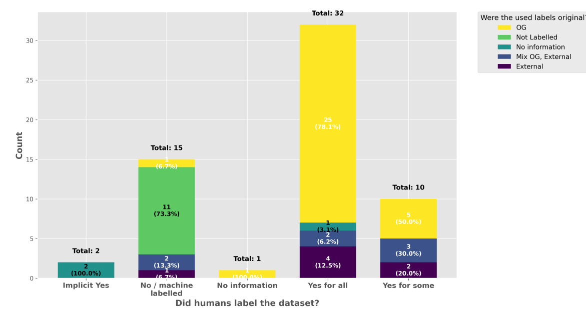

Figure 4: Head to head comparison of the "Human labels" and "OG Labels" items. The answers to "OG Labels" can be seen horizontally, and then based upon that it can be seen vertically what the answers to "Human labels" were.

own part. Figure 8 in the Appendix gives a complete overview of the findings overall and Figure 9 gives an overview specifically for period 2.

Annotator transparency Figure 3 shows the summary of the findings related to annotator transparency. Out of the 60 datasets, ∼57% were annotated by humans, ∼17% had a mix between humans and machines, ∼25% were either not annotated or annotated by machines and ∼1% (one dataset) was unclear. Figure 4 shows the distribution of the originality (i.e. whether the labels were made for the dataset or collected from somewhere else) based on the extent humans contributed to it. It can be seen that when the labels were made by humans, 78.15% of the time the labels were original. However, when there were no human annotators, 73.3% of the time the datasets were not labeled whatsoever[3].

When applicable, the source of the labels was Amazon Mechanical Turk[4] ∼35% of the time, other ∼46% of the time and ∼20% of the time the authors did not specify the source of the labels. When human labelers were involved, the labeler population rationale (i.e. reason for choosing those annotators) was mentioned by few authors (∼23%), the rest (∼77%) giving no reason. Prescreening was mentioned ∼37% of the time, with the most common methods being based on a skill the annotator (12.20% out of all papers) or their performance on the platform so far (12.20% out of all papers). The com-

[3]Note that this does not diminish the quality of the dataset, as some datasets are collections of text used to train models how to form sentences (simplified explanation).

[4]Crowdsourcing marketplace, accessible at: https://www.mturk.com/

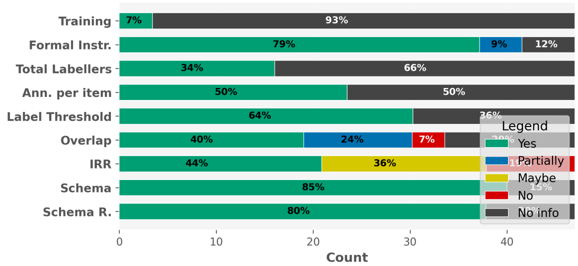

Figure 5: Annotator procedures results.

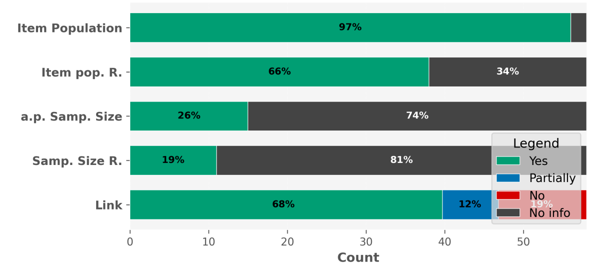

Figure 6: Data collection practices results. "a.p." stands for "a Priori", which means "before".

pensation for the work was mostly money (58.54%), then no compensation (volunteer work) 2.44% and unspecified in 39.02% of the cases.

In the past 2 years, the amount of human annotated datasets has maintained the relative proportion, with slightly more original datasets. While the label sources are relatively similar, the labeler population rational grew to being mentioned by ∼43% of the authors. Prescreening mentions also grew, with ∼58% mentioning some prescreening. Compensation is also more frequently mentioned, as 75% of authors mention it (as opposed to the previous 60.98%).

Annotation procedures A summary of the annotation practices can be found in Figure 5. Regarding individual label quality, few dataset papers documented giving any training (∼7%). However, formal instructions were usually present to at least some extent (∼88%), and there was usually some annotation schema mentioned to guide the labelers (∼72%). Interestingly, when applicable, authors provide a rationale for the way they want the labels annotated significantly more often (∼80%) than mentioning the existence of a annotation schema (∼72%).

When it comes to inter-rater reliability (IRR) measurements, for ∼50% of the papers the number of annotators per item was mentioned or could be inferred. It is to also be noted that only ∼34% of the papers mentioned how many total labelers were present. From the ones that did have more than just one annotator per item, ∼64% mentioned overlap for at least some of the items (and ∼40% for all items). IRR was calculated only ∼44% of the time, with it being potentially not applicable ∼36% of the time because not information about overlap was given (i.e. maybe the authors had just one annotator but did not mention it, so in this case IRR could be not applicable). This means that IRR was calculated at least ∼44% of the time, maybe even more if we could exclude the ∼36% of potentially not applicable instances.

There are however some noticeable changes in the past 2 years. Training is still relatively low (∼17%), however most papers do give formal instructions to their labelers (∼93%). Labelers per item are more frequently mentioned (∼71%), and the same goes for total labelers (∼50% mentioned). More papers mentioned overlap to some extent (75%) and where applicable ∼60% calculated IRR, with ∼20% of papers potentially being not applicable.

Data collection practices A summary of the results can be found in Figure 6. With few exceptions (∼3% and ∼5%), papers describe their item population and item source. However, not as many give a reason for choosing such an item population or choosing to collect that data in such a way, with only 66% giving a rationale for the item population.

Sample size is not also generally paid much attention to. Only ∼26% of the papers mention aiming for a predetermined sample size and even fewer (∼19%) actually give a reason for collecting so many items.

The link to the datasets were not always available either. Throughout all the periods, a link to the dataset was available on the paper or on the ACL abstract[5] only ∼68% of the time, with a broken link being available ∼12% of the time. This leaves ∼20% of the papers without any link to the dataset produced.

There are some differences in the past two years. Now, all papers describe their item population and mention their item source and almost all (∼95%) give an item population rationale. Almost the same proportion of papers decide their sample size beforehand, but fewer (∼10%) give a reason for going for this amount of items. A working link is available more often (∼79%) and a broken link is available ∼10% of the time, leaving less papers (∼11%) without any link whatsoever.

3.3 Missing information

This subsection relates to research subquestion 3, with the missing information per period being displayed in Figure 7. The pattern to be pointed out is the amount of missing information per period, and the fact that it getting better. In order to illustrate this, a missing information percentage per period will be presented.

Figure 7 shows the distribution of the missing information per periods and overall. The overall missing information amounts to 33.03%. This means that almost one third of the fields considered for analysis could not be filled due to lack of information given by the authors.

It also be seen that the most cited papers from the past 15 years used datasets with a higher percentage of missing information than papers in the past 5 years. Furthermore, papers in the past two years are doing the best, with their datasets having 24.9% missing information according to our metrics.

[5]The page where the abstract and other meta information of the paper can be seen within the ACL anthology

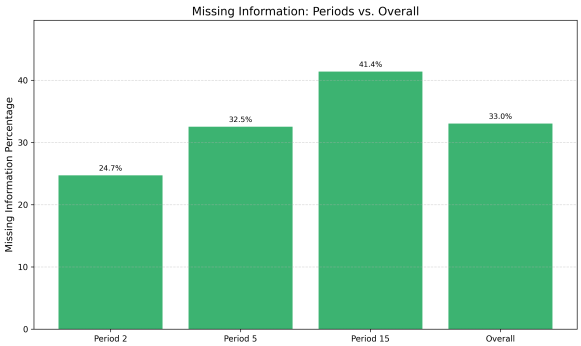

Figure 7: Information missing from the datasets by period and overall.

4 Discussion

As a followup of Section 3, this part will put the findings in context, giving more meaning to the answers given to the research subquestions. It will go over each subsection present in the findings and talk about the most concerning aspects and give reasons as to why they might be alarming and why they should be paid attention to. Section 4.4 ties the results of the findings together and gives brief recommendations on what should be done.

4.1 Dataset overlap

The main trend visible from the results presented in Subsection 3.1 is that there has been a radical shift in the datasets used in the past 2 years, as compared to the ones in the past 5 or 15 years. This is most likely due to the increased attention LLMs have gotten after the release of ChatGPT-3 in 2022[6]. Additional reasons for this, as mentioned by (Orr and Crawford, 2024) are because some datasets can become obsolete, better datasets get released or older datasets may get "solved" or even retracted.

It can also be seen that papers in the past two years use a lower amount of datasets, signifying that either (1) the tasks have become more specific, so less datasets are available, (2) authors now prefer to use more specific datasets or (3) that there is currently a lower amount of of datasets for the tasks authors now focus on. The actual reason is most likely a combination of the previously mentioned reasons, with a hint that more recent papers are now dealing with LLMs. Therefore, more sophisticated and challenging benchmarks are required, which may be harder to construct.

Those datasets used by the papers in the past 2 years generally contain more information about their annotation practices and data collection, as substantiated by the other subsections in Section 3. While this does represent a step in the right direction of quality datasets, this does not undo the potential harm done by using lower quality datasets in the past. Those datasets used by older papers could have given a false boost in the performance of the model evaluated, artificially increasing the validity of those models. This is actually supported by (Le Bras et al., 2020; Niven and Kao, 2019), who have shown that manipulating the test data of the datasets that BERT (Devlin et al., 2019a) was trained on such as GLUE leads to significantly decreased performance pf the model (while human performance remains stable).

[6]According to https://openai.com/index/chatgpt/, accessed 16th June 2025

4.2 Data collection and annotation practices

This subsection will interpret the findings of all the parts of Subsection 3.2, namely annotator transparency, annotation practices and data collection practices.

Annotator transparency

The results discussed in the first part of Subsection 3.2 show that there is reason for concern in a couple aspects regarding annotator transparency, even for more recently used datasets. The most concerning are the lack of any mention for prescreening or labeler population rationale, occasional lack of a annotator source and the frequent use of MTurk as the annotator source, all the while seldom giving details about a reason for choosing such an annotator source and any details about prescreening.

Giving details about the annotation source, rationale and prescreening is absolutely crucial as the diversity of the dataset can be explained. While a label source is not given sometimes (∼20%), the rationale for choosing those labelers is usually skipped (∼77%) and the prescreening (or lack thereof) is not stated ∼63% of the time. As (Orr and Crawford, 2024) relay from the opinions of other expert dataset creators in their first recommendation for building better datasets, ensuring diversity is crucial for a responsibly created dataset. Diversifying annotator backgrounds is one way to achieve that, but given the low information usually present about the labellers, it can be assumed that annotator diversity is usually not achieved.

MTurk is also used quite frequently (32.61% of the time) as a label source for the generation of labels. (Aguinis et al., 2021) notes that scholarly opinions of the usage of MTurk are mixed, with some journals even sometimes refusing papers that used the platform for their work. The reason for MTurk being so controversial is that there are a plethora of pitfalls to using MTurk without necessary precautions, as described in the aforementioned paper. The generally low amount of prescreening and frequently unsubstantiated choice of labelers leads to believe the use of the platform was not used responsibly enough, and as a result the annotation quality could have suffered.

Reporting of labeller population rationale and prescreening has improved in datasets used in the recent years. Labeller population rationale saw a jump from being mentioned only ∼23% of the time to being mentioned ∼43% of the time. Prescreening is now stated more often, at higher ∼58% compared to the previous ∼37%.

Annotation procedures

The second part of Subsection 3.2 sheds light on frequent oversights regarding annotation procedures. What is most alarming is that authors either do not always calculate IRR metrics or do not set up their data formation in such a way that it can be calculated (i.e. single annotator per item). Another concern is the lack of a mention of an annotation schema for the annotators to use.

As (Aroyo and Welty, 2015) mention, one single annotator is often not enough to annotate data reliably, but in the results it can be seen that only ∼64% of the papers have at least some overlap, with only ∼40% having overlap for all items. IRR might be calculated as low as ∼44% of the time when applicable, which leaves potentially more than half of the papers without IRR or with the possibility of shaping the experiment in such a shape that IRR could be calculated. The reason the lack of IRR is such a big issue as it represents a part of the confidence that you can have in the data, as suggested by (McHugh, 2012). In datasets used in the past two years this has however improved, with more papers (∼60%) calculating IRR, suggesting scientists are designing more reliable experiments.

Not mentioning an annotation schema or lack thereof (which happens ∼28% of the time) is also harmful for the replicability and reproducibility of the data. Even though authors do give some mention of their rationale for annotating the items in a certain way (∼80% overall), this is not enough to replicate or reproduce the experiment.

Data collection practices

When analyzing the findings of Subsubsection 3.2, authors usually overlook a couple of important points. Among those are the general lack of attention to details related to the amount of items and the rationale given for item populations and the not uncommon lack of a (working) link to the dataset.

Only ∼26% of the papers actually give evidence for having decided the amount of items they would like to gather beforehand, usually just "ending up" with that amount of items. Even fewer (∼19%) give a reason after collection of why some many items have been gathered. This shows that the scientists usually do not consider how many samples their dataset would need in order to properly train their model. While not having a link to the dataset happens 68% of the time overall, this has gotten better with datasets used in the past 2 years, with ∼79% of papers having a working link.

4.3 Missing information

Perhaps the most troubling of findings is presented in Subsection 3.3. That subsection shows the amount of data that no information could be gathered about during the annotation effort, with the highlights being the fact that very close to 1/3 of the information is missing regardless of period and 24.03% is missing in the past 2 years.

The fact that there is less missing information in datasets used the past 2 years as compared to the past 5 or 15 years suggests that datasets are getting increasingly more well documented, and scientists are beginning to agree on a standard of a better quality of datasets. This makes sense, as authors submitting papers for the ACL are required to complete a checklist before submitting their paper[7], ensuring best practices have been followed. Other academic organizations such as NeurIPS have started implementing mandatory checklists[8]. There have even been efforts to correct "Documentation debt" by publishing retroactive papers about datasets already realeased, such as (Bandy and Vincent, 2021).

However, the recent improvements in the amount of missing information does not diminish the fact the there is still 24.03% missing in the past two years, and ∼1/3 missing overall. This is especially concerning as previously used datasets lack the information to make them reproducible.

4.4 Tying the results together and recommendations

Following the discussion based on all the research subquestions, an answer can be given to the research question posed: "What are the data collection and annotation practices of the datasets present in the most impactful papers of the ACL?". (1) Papers in the past tended to reuse more datasets than they did in the past 2 years, with this period having the least amount of total datasets used as well. Datasets in the past 2 years are also very different compared to the ones in the past 5 and 15 years, probably because of the rise of ChatGPT and other LLMs. (2) Within the datasets used, there are various concerns, mostly regarded to information missing in many places, but also the relatively high(∼36%) and unregulated use of MTurk. The most important missing aspects are, with how often they are missing: labeller population rationale(∼77%), prescreening(∼63%), IRR(19%-56%) and sample size rationale(∼81%). In the datasets used only in the past 2 years, the amount of missing information from those fields is generally lower: labeller population rationale(∼57%), prescreening(∼40%), IRR (20% -40%), but sample size rationale has a higher amount of missing information (∼90%) (3) Finally, it has been seen that information is missing 33.03% of the time overall, with a lower 24.7% of the time in the past 2 years, which signifies a development towards better quality datasets.

While making an actionable plan on how to make datasets of better quality is outside the scope of this paper, some resources and examples will be given. Some papers such as (Orr and Crawford, 2024; Gebru et al., 2018) already exist, which give recommendations on how to how to better report on datasets or even how to make them of better quality (which also implies better documenting them). Those represent good starting points for other academic organizations to begin with enforcing better dataset standards.

I believe academic organizations should also play a role in creating better regulated datasets, with each one requiring their papers to include a checklist for their datasets. Both ACL and NeurIPS have some guidelines in place on how to report on datasets, as mentioned in Section 4.3. However, when reading the most highly cited papers selected, only one of them, namely (Suzgun et al., 2023a) included the ACL checklist in the paper itself, with the rest of the papers not having such a section. Therefore, better enforcement of this checklist inclusion is needed, with each paper including such a checklist.

5 Responsible Research

This research has involved other human subjects as annotators besides myself whenever there was dataset overlap (i.e. a

[7]According to https://aclrollingreview.org/responsibleNLPresearch/, accessed on the 16th of June 2025

[8]According to https://neurips.cc/public/guides/PaperChecklist, accessed on the 16th of June 2025

dataset was used by both our domains tackled). As with any annotation work, subjectivity and interpretation are unavoidable, but those were attempted to be minimized by using a well crated annotation schema, as described in Subsubsection 2.3.

Although LLMs were used for purposes related to this study, they have never replaced critical thinking and have not been used to generate full sentences. The purposes I used LLMs for were the following: (1) typo and grammar correction, along with stylistic suggestions for the text and (2) as inspiration for the creation of the scripts used to analyze that data. The code generated by the LLMs was always double checked and modified when there were inconsistencies. I have not used LLMs to summarize or extract information out of any paper in this or related to this study.

In the spirit of replicability, the general methods through which the data has been collected and selected are available in Section 2. Additional details such as the papers surveyed, the datasets annotated and the query used to collect the papers are available in Appendix E, F and A. The actual data, how to use it and how the statistics have been calculated in Appendix B. The annotation schema used for the datasets is available in Appendix C.

However, reproducibility is difficult to argue for in this case, because the outcome is highly dependent on the domain of the papers selected. For instance, papers posted by a different organization or academic journal may pay more attention to using good datasets. What can be argued however, is that using different defensible measures of impact and selection criteria for the same domain and periods we selected will lead to approximately the same answers to the research questions we used. This is because the following have been well argumented: (1) selection of the papers pointing towards influential works within 3 different periods, (2) selection of the datasets extracted from the influential works and (3) the items selected for the annotation schema.

6 Limitations and future work

This section will present the limitations of this paper, which will be paired with some future work and suggestions for the future. The limitations presented are: (1) domain related limitations, (2) dataset annotation related limitations, (3) lack of further in depth MTurk usage analysis and (4) only brief suggestions of improvement.

This paper has only analyzed papers published by only one academic organization, namely ACL, which deals with Natural Language Processing and Computational Linguistics. This was to make the analysis process simpler. In future analyses, other domains should be chosen, or maybe even multiple domains. This would help the science community get a better overview of the state of annotation and data collection practices.

Although this paper critiques the fact that less than half of the datasets papers that had annotation had overlap, this paper included human annotation and did not have overlap. This was because a bigger importance was placed on having enough papers to analyze and because resources were limited in the context of this short research project. In the future, I would urge authors that undertake such an annotation effort to have multiple annotators annotate at least a validation subset to calculate IRR, and give further validity to their claims.

The use of MTurk in the analyzed datasets was only cautiously critiqued, as there was no special attention paid to the correlation between MTurk usage and how carefully the experiment was designed. In the future, I would suggest that other scientists take this into account, and pay attention to it, so that a more powerful statement can be made upon the usage of the platform within ACL.

Finally, although a lot of critique was given, suggestions for improvement have only been given in passing, with no evidence collected as to whether they would make a difference. The aim of this study was to only give an overview of the datasets used by the ACL, so the suggestion of solutions was not the main concern. In order to solve a problem it needs to be specifically identified, which was done in this case. I will leave the next step to future work - which is finding a solution of the problem.

7 Conclusions

The purpose of this study was to give an overview of the annotation and data collection practices of the datasets used by the most impactful papers published within the ACL. This was explored by looking at multiple aspects: Which datasets are most often used by those papers? how well do the most used datasets report on annotation practices? What about data collection practices? Finally, how much information is missing from those datasets?

Dataset overlap and the most used datasets were identified and it was found out that there has been a major shift regarding the specific datasets in the past two years. Issues with annotation practices were found, such as the frequent use of MTurk($\sim$36%) without any additional precautions, the labeller population and prescreening being frequently not reported (missing $\sim$77% and $\sim$63% of the time, respectively) and IRR not being calculated between $\sim$19% and $\sim$56% of the time. Regarding data collection, it was found that papers seldom ($\sim$19% of the time) give any reason for their sample size. The issues regarding lack of reporting for those fields have generally improved. Overall, it was found that $\sim$1/3 of the information sought was missing from those datasets, with a slightly lower $\sim$1/4 for datasets in the past 2 years.

The findings were also discussed and interpreted. It was explained which results were most concerning and why they were alarming. Although some suggestions were given in passing for better quality datasets (i.e. better reporting or additional precautions), those were only given in passing and not supported by evidence that they would work. While having guidelines or a reporting template as Orr and Crawford (2024); Gebru et al. (2018) have suggested would greatly help, I believe the best solution would be having guidelines from each venue to include such a template in their paper.

References

Aguinis, H., Villamor, I., and Ramani, R. S. (2021). Mturk research: Review and recommendations. *Journal of Management*, 47(4):823–837.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Bandy, J. and Vincent, N. (2021). Addressing "documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In Vanschoren, J. and Yeung, S., editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Koehn, P., Malmasi, S., Monz, C., Müller, M., Pal, S., Post, M., and Zampieri, M. (2019). Findings of the 2019 conference on machine translation (WMT19). In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Yepes, A. J., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics. ds: WMT19; citation sum: 2284.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Névéol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 conference on machine translation. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Guillou, L., Haddow, B., Huck, M., Yepes, A. J., Névéol, A., Neves, M., Pecina, P., Popel, M., Koehn, P., Monz, C., Negri, M., Post, M., Specia, L., Verspoor, K., Tiedemann, J., and Turchi, M., editors, *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics. ds: WMT16; citation sum: 2002.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015a). A large annotated corpus for learning natural language inference. page 632 – 642. Cited by: 2589; All Open Access, Green Open Access.

Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015b). A large annotated corpus for learning natural language inference. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics. ds: SNLI; citation sum: 3757.

Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity - multilingual and cross-lingual focused evaluation. *CoRR*, abs/1708.00055. ds: STSb; citation sum: 47482; expansion of: GLUE.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., and Koehn, P. (2013). One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, abs/1312.3005. ds: 1B word; citation sum: 2354.

Chiang, C.-H. and Lee, H.-Y. (2023). Can large language models be an alternative to human evaluation? volume 1, page 15607 – 15631. Cited by: 194; All Open Access, Green Open Access, Hybrid Gold Open Access.

Chinchor, N. (2001). Message understanding conference (muc) 7. Web Download. ds: MUC-7; citation sum: 27202.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. page 1724 – 1734. Cited by: 12056.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., and Schulman, J. (2021). Training verifiers to solve math word problems. ds: GSM8k; citation sum: 197.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. page 8440 – 8451. Cited by: 3349.

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2020). Transformer-xl: Attentive language models beyond a fixed-length context. page 2978 – 2988. Cited by: 1330.

Davis, E., Morgenstern, L., and Ortiz, C. (n.d.). The winograd schema challenge. ds: WNLI; citation sum: 47482; expansion of: GLUE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *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.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). Bert: Pre-training of deep bidirectional transformers for language understanding. volume 1, page 4171 – 4186. Cited by: 45837.

Dolan, W. B. and Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. ds: MRPC; citation sum: 47482; expansion of: GLUE.

Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., and Tang, J. (2022). Glm: General language model pretraining with autoregressive blank infilling. volume 1, page 320 – 335. Cited by: 824; All Open Access, Green Open Access, Hybrid Gold Open Access.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). Summeval: Reevaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. ds: SummEval; citation sum: 233.

Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics. ds: WritingPrompts; citation sum: 176.

Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., and Zhou, M. (2020). Codebert: A pre-trained model for programming and natural languages. page 1536 – 1547. Cited by: 1070.

Gao, L., Dai, Z., Pasupat, P., Chen, A., Chaganty, A. T., Fan, Y., Zhao, V. Y., Lao, N., Lee, H., Juan, D.-C., and Guu, K. (2023). Rarr: Researching and revising what language models say, using language models. volume 1, page 16477 – 16508. Cited by: 115; All Open Access, Green Open Access, Hybrid Gold Open Access.

Gao, T., Fisch, A., and Chen, D. (2021a). Making pre-trained language models better few-shot learners. volume 1, page 3816 – 3830. Cited by: 1056; All Open Access, Green Open Access, Hybrid Gold Open Access.

Gao, T., Yao, X., and Chen, D. (2021b). Simcse: Simple contrastive learning of sentence embeddings. page 6894 – 6910. Cited by: 1886.

Gardent, C., Shimorina, A., Narayan, S., and Perez-Beltrachini, L. (2017). The WebNLG challenge: Generating text from RDF data. In Alonso, J. M., Bugarín, A., and Reiter, E., editors, *Proceedings of the 10th International Conference on Natural Language Generation*, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics. ds: WebNLG 2017; citation sum: 2653.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., III, H. D., and Crawford, K. (2018). Datasheets for datasets. *CoRR*, abs/1803.09010.

Geiger, R. S., Cope, D., Ip, J., Lotosh, M., Shah, A., Weng, J., and Tang, R. (2021). "garbage in, garbage out" revisited: What do machine learning application papers report about human-labeled training data? *CoRR*, abs/2107.02278.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., and Huang, J. (2020). Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 325–336, New York, NY, USA. Association for Computing Machinery.

Geva, M., Khashabi, D., Segal, E., Khot, T., Roth, D., and Berant, J. (2021). Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361. ds: strategyQA; citation sum: 403.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tur, D. (2023). Topical-chat: Towards knowledge-grounded open-domain conversations. ds: Topical-Chat; citation sum: 233.

Gururangan, S., Marasovic, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., and Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. page 8342 – 8360. Cited by: 1268.

Ho, X., Duong Nguyen, A.-K., Sugawara, S., and Aizawa, A. (2020). Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics. ds: 2WikiMultiHop; citation sum: 195.

Hsieh, C.-Y., Li, C.-L., Yeh, C.-K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.-Y., and Pfister, T. (2023). Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. page 8003 – 8017. Cited by: 113; All Open Access, Green Open Access, Hybrid Gold Open Access.

Huang, J. and Chen-Chuan Chang, K. (2023). Towards reasoning in large language models: A survey. page 1049 – 1065. Cited by: 155; All Open Access, Green Open Access, Hybrid Gold Open Access.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., and Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, page 335–348, New York, NY, USA. Association for Computing Machinery.

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., and Munigala, V. (2020). Overview and importance of data quality for machine learning tasks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3561–3562, New York, NY, USA. Association for Computing Machinery.

Jawahar, G., Sagot, B., and Seddah, D. (2020). What does bert learn about the structure of language? page 3651 – 3657. Cited by: 971.

Jiang, Z., Xu, F. F., Araki, J., and Neubig, G. (2020). How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423 – 438. Cited by: 995; All Open Access, Gold Open Access, Green Open Access.

Jiang, Z., Xu, F. F., Gao, L., Sun, Z., Liu, Q., Dwivedi-Yu, J., Yang, Y., Callan, J., and Neubig, G. (2023). Active retrieval augmented generation. page 7969 – 7992. Cited by: 118; All Open Access, Green Open Access, Hybrid Gold Open Access.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., and Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64 – 77. Cited by: 1377; All Open Access, Gold Open Access, Green Open Access.

Joshi, M., Choi, E., Weld, D., and Zettlemoyer, L. (2017). TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Barzilay, R. and Kan, M.-Y., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics. ds: TriviaQA; citation sum: 3121.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. volume 2, page 427 – 431. Cited by: 2186; All Open Access, Green Open Access, Hybrid Gold Open Access.

Kapoor, S. and Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*. Publisher: Elsevier.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-T. (2020). Dense passage retrieval for open-domain question answering. page 6769 – 6781. Cited by: 1847.

Kim, Y. (2014). Convolutional neural networks for sentence classification. page 1746 – 1751. Cited by: 8279; All Open Access, Green Open Access, Hybrid Gold Open Access.

Kirk, H. R., Yin, W., Vidgen, B., and Röttger, P. (2023). Semeval-2023 task 10: Explainable detection of online sexism. page 2193 – 2210. Cited by: 122.

Koncel-Kedziorski, R., Hajishirzi, H., Sabharwal, A., Etzioni, O., and Ang, S. D. (2015). Parsing algebraic word problems into equations. *Transactions of the Association for Computational Linguistics*, 3:585–597. ds: SingleEQ; citation sum: 197.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. page 66 – 71. Cited by: 2151; All Open Access, Green Open Access, Hybrid Gold Open Access.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466. ds: nq; citation sum: 231.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. page 260 – 270. Cited by: 2642; All Open Access, Green Open Access, Hybrid Gold Open Access.

Le Bras, R., Swayamdipta, S., Bhagavatula, C., Zellers, R., Peters, M. E., Sabharwal, A., and Choi, Y. (2020). Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Lee, K., Chang, M.-W., and Toutanova, K. (2019). Latent retrieval for weakly supervised open domain question answering. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics. ds: NaturalQuestions-Open; citation sum: 234.

Lester, B., Al-Rfou, R., and Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. page 3045 – 3059. Cited by: 1674.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. page 7871 – 7880. Cited by: 5688.

Li, J., Cheng, X., Zhao, W. X., Nie, J.-Y., and Wen, J.-R. (2023a). Halueval: A large-scale hallucination evaluation benchmark for large language models. page 6449 – 6464. Cited by: 169.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. volume 1, page 4582 – 4597. Cited by: 1885; All Open Access, Green Open Access, Hybrid Gold Open Access.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. (2023b). Evaluating object hallucination in large vision-language models. page 292 – 305. Cited by: 155.

Li, Y., Lin, Z., Zhang, S., Fu, Q., Chen, B., Lou, J.-G., and Chen, W. (2023c). Making large language models better reasoners with step-aware verifier. volume 1, page 5315 – 5333. Cited by: 103; All Open Access, Hybrid Gold Open Access.

Liem, C. C. S., Taşcılar, D., and Demetriou, A. M. (2024). A quest through interconnected datasets: Lessons from highly-cited icassp papers. In *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*, pages 1–8.

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157 – 173. Cited by: 297.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726 – 742. Cited by: 1261; All Open Access, Gold Open Access, Green Open Access.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). G-eval: Nlg evaluation using gpt-4 with better human alignment. page 2511 – 2522. Cited by: 263; All Open Access, Green Open Access, Hybrid Gold Open Access.

Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. page 1412 – 1421. Cited by: 4182; All Open Access, Green Open Access, Hybrid Gold Open Access.

Luong, T., Socher, R., and Manning, C. (2013). Better word representations with recursive neural networks for morphology. In Hockenmaier, J. and Riedel, S., editors, *Pro-*

ceedings of the Seventeenth Conference on Computational Natural Language Learning, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics. ds: RW; citation sum: 27202.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. volume 1, page 142 – 150. Cited by: 4108.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2023a). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. volume 1, page 9802 – 9822. Cited by: 213; All Open Access, Green Open Access, Hybrid Gold Open Access.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., and Hajishirzi, H. (2023b). When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics. ds: PopQA; citation sum: 197.

Manakul, P., Liusie, A., and Gales, M. J. (2023). Selfcheck-gpt: Zero-resource black-box hallucination detection for generative large language models. page 9004 – 9017. Cited by: 130; All Open Access, Green Open Access, Hybrid Gold Open Access.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The stanford corenlp natural language processing toolkit. volume 2014-June, page 55 – 60. Cited by: 5893.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA). ds: SICK; citation sum: 2660.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*, 22(3):276–282. STAT: MEDLINE; PMID: 23092060; PMC: PMC3900052.

Miao, S.-y., Liang, C.-C., and Su, K.-Y. (2020). A diverse corpus for evaluating and developing English math word problem solvers. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 975–984, Online. Association for Computational Linguistics. ds: ASDiv; citation sum: 195.

Mikolov, T., Yih, W.-T., and Zweig, G. (2013). Linguistic regularities in continuous spaceword representations. page 746 – 751. Cited by: 2699.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28. ds: mc; citation sum: 27202.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. page 12076 – 12100. Cited by: 164; All Open Access, Green Open Access, Hybrid Gold Open Access.

Mitchell, A., Strassel, S., Przybocki, M., Davis, J. K., Doddington, G. R., Grishman, R., Meyers, A., Brunstein, A., Ferro, L., and Sundheim, B. (2004). Tides extraction (ace) 2003 multilingual training data. ds: ace-2003; citation sum: 27202.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2023a). Mteb: Massive text embedding benchmark. page 2006 – 2029. Cited by: 158.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M. S., Shen, S., Yong, Z.-X., Schoelkopf, H., Tang, X., Radev, D., Aji, A. F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., and Raffel, C. (2023b). Crosslingual generalization through multitask finetuning. volume 1, page 15991 – 16111. Cited by: 218; All Open Access, Green Open Access, Hybrid Gold Open Access.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. (2020). Adversarial NLI: A new benchmark for natural language understanding. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annl Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics. ds: ANLI; citation sum: 1801.

Niven, T. and Kao, H.-Y. (2019). Probing neural network comprehension of natural language arguments. In Korhonen, A., Traum, D., and Màrquez, L., editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Orr, W. and Crawford, K. (2024). Building better datasets: Seven recommendations for responsible design from dataset creators.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. page 48 – 53. Cited by: 2207.

Parker, R., Graff, D., Kong, J., Chen, K., and Maeda, K. (2011). English gigaword fifth edition. Web Download. ds: gigaword-5; citation sum: 27202.

Patel, A., Bhattamishra, S., and Goyal, N. (2021). Are NLP models really able to solve simple math word problems? In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 2080–2094, Online. Association for Computational Linguistics. ds: SVAMP; citation sum: 297.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. page 1532 – 1543. Cited by: 27640.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. volume 1, page 2227 – 2237. Cited by: 6983.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. (2019). Language models as knowledge bases? In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *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 2463–2473, Hong Kong, China. Association for Computational Linguistics. ds: lama; citation sum: 1816.

Piskorski, J., Stefanovitch, N., Da San Martino, G., and Nakov, P. (2023). Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. page 2343 – 2361. Cited by: 107; All Open Access, Hybrid Gold Open Access.

Press, O., Zhang, M., Min, S., Schmidt, L., Smith, N. A., and Lewis, M. (2023). Measuring and narrowing the compositionality gap in language models. page 5687 – 5711. Cited by: 151.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A python natural language processing toolkit for many human languages. page 101 – 108. Cited by: 1086.

Quora (2012). First quora dataset release: Question pairs. ds: QQP; citation sum: 47482; expansion of: GLUE.

Radev, D. R., Zhang, R., Rau, A., Sivaprasad, A., Hsieh, C., Rajani, N. F., Tang, X., Vyas, A., Verma, N., Krishna, P., Liu, Y., Irwanto, N., Pan, J., Rahman, F., Zaidi, A., Mutuma, M., Tarabar, Y., Gupta, A., Yu, T., Tan, Y. C., Lin, X. V., Xiong, C., and Socher, R. (2020). DART: open-domain structured data record to text generation. *CoRR*, abs/2007.02871. ds: dart; citation sum: 3592.

Rajpurkar, P., Jia, R., and Liang, P. (2018a). Know what you don't know: Unanswerable questions for SQuAD. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Rajpurkar, P., Jia, R., and Liang, P. (2018b). Know what you don't know: Unanswerable questions for squad. *CoRR*, abs/1806.03822. ds: SQuADv2; citation sum: 44198.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016a). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016b). Squad: 100,000+ questions for machine comprehension of text. page 2383 – 2392. Cited by: 4101; All Open Access, Green Open Access, Hybrid Gold Open Access.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016c). SQuAD: 100,000+ questions for machine comprehension of text. In Su, J., Duh, K., and Carreras, X., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics. ds: SQuAD; citation sum: 5946.

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316 – 1331. Cited by: 151; All Open Access, Gold Open Access, Green Open Access.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. page 3982 – 3992. Cited by: 6520.

Rogers, A., Kovaleva, O., and Rumshisky, A. (2020). A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842 – 866. Cited by: 939; All Open Access, Gold Open Access, Green Open Access.

Roy, S. and Roth, D. (2015). Solving general arithmetic word problems. In Màrquez, L., Callison-Burch, C., and Su, J., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1743–1752, Lisbon, Portugal. Association for Computational Linguistics. ds: MultiArith; citation sum: 198.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy. *Commun. ACM*, 8(10):627–633. ds: RG; citation sum: 27202.

Schick, T. and Schütze, H. (2021). Exploiting cloze questions for few shot text classification and natural language inference. page 255 – 269. Cited by: 953; All Open Access, Green Open Access, Hybrid Gold Open Access.

Sciavolino, C., Zhong, Z., Lee, J., and Chen, D. (2021). Simple entity-centric questions challenge dense retrievers. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. ds: entityquestions; citation sum: 197.

See, A., Liu, P. J., and Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. volume 1, page 1073 – 1083. Cited by: 2671; All Open Access, Green Open Access, Hybrid Gold Open Access.

Sellam, T., Das, D., and Parikh, A. P. (2020). Bleurt: Learning robust metrics for text generation. page 7881 – 7892. Cited by: 883.

Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., and Kowald, D. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. *AI Magazine*, 46(2):e70002.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. volume 3, page 1715 – 1725. Cited by: 4567; All Open Access, Green Open Access, Hybrid Gold Open Access.

Shin, T., Razeghi, Y., Logan, R. L., Wallace, E., and Singh, S. (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. page 4222 – 4235. Cited by: 893; All Open Access, Green Open Access, Hybrid Gold Open Access.

Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., and Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (rag) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1 – 17. Cited by: 116; All Open Access, Gold Open Access, Green Open Access.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013a). Recursive deep models for semantic compositionality over a sentiment treebank. In Yarowsky, D., Baldwin, T., Korhonen, A., Livescu, K., and Bethard, S., editors, *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics. ds: SST-2; citation sum: 47482; expansion of: GLUE.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013b). Recursive deep models for semantic compositionality over a sentiment treebank. page 1631 – 1642. Cited by: 6276.

Strubell, E., Ganesh, A., and McCallum, A. (2020). Energy and policy considerations for deep learning in nlp. page 3645 – 3650. Cited by: 973.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q., Chi, E., Zhou, D., and Wei, J. (2023a). Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H., Zhou, D., and Wei, J. (2023b). Challenging big-bench tasks and whether chain-of-thought can solve them. page 13003 – 13051. Cited by: 117; All Open Access, Green Open Access, Hybrid Gold Open Access.

Talmor, A., Herzig, J., Lourie, N., and Berant, J. (2019). CommonsenseQA: A question answering challenge targeting commonsense knowledge. In Burstein, J., Doran, C., and Solorio, T., editors, *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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics. ds: CommonSenseQA; citation sum: 297.

Thyagarajan, A., Snorrason, E., Northcutt, C., and Mueller, J. (2022). Identifying incorrect annotations in multi-label classification data.

Tjong Kim Sang, E. F. (2002). Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*. ds: CoNLL-2002; citation sum: 1838.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. ds: CoNLL-2003; citation sum: 1838.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Zico Kolter, J., Morency, L.-P., and Salakhutdinov, R. (2020). Multimodal transformer for unaligned multimodal language sequences. page 6558 – 6569. Cited by: 1249.

Unknown (n.d.a). English wikipedia. ds: English Wikipedia; citation sum: 2221.

Unknown (n.d.b). Recognizing textual entailment. ds: RTE; citation sum: 47482; expansion of: GLUE.

Wadhwa, S., Amir, S., and Wallace, B. C. (2023). Revisiting relation extraction in the era of large language models. volume 1, page 15566 – 15589. Cited by: 139; All Open Access, Green Open Access, Hybrid Gold Open Access.

Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics. ds: QAGS; citation sum: 233.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018a). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018b). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Linzen, T., Chrupała, G., and Alishahi, A., editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics. ds: GLUE; citation sum: 4092.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., and Lim, E.-P. (2023a). Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. volume 1, page 2609 – 2634. Cited by: 127; All Open Access, Green Open Access, Hybrid Gold Open Access.

Wang, Y., Huang, M., Zhao, L., and Zhu, X. (2016). Attention-based lstm for aspect-level sentiment classification. page 606 – 615. Cited by: 2220; All Open Access, Hybrid Gold Open Access.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023b). Self-instruct: Aligning language models with self-generated instructions. volume 1, page 13484 – 13508. Cited by: 427; All Open Access, Green Open Access, Hybrid Gold Open Access.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. (2023c). Self-instruct: Aligning language models with self-generated instructions. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics. ds: Self-Instruct; citation sum: 392.

Wang, Y., Mishra, S., Alipoormolabashi, P., Kordi, Y., Mirzaei, A., Naik, A., Ashok, A., Dhanasekaran, A. S., Arunkumar, A., Stap, D., Pathak, E., Karamanolakis, G., Lai, H., Purohit, I., Mondal, I., Anderson, J., Kuznia, K., Doshi, K., Pal, K. K., Patel, M., Moradshahi, M., Parmar, M., Purohit, M., Varshney, N., Kaza, P. R., Verma, P., Puri, R. S., Karia, R., Doshi, S., Sampat, S. K., Mishra, S., Reddy A, S., Patro, S., Dixit, T., and Shen, X. (2022). Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In Goldberg, Y., Kozareva, Z., and Zhang, Y., editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. ds: superni; citation sum: 392.

Warstadt, A., Singh, A., and Bowman, S. R. (2018). Neural network acceptability judgments. *CoRR*, abs/1805.12471. ds: CoLA; citation sum: 47482; expansion of: GLUE.

Wenzek, G., Lachaux, M., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., and Grave, E. (2019). Ccnet: Extracting high quality monolingual datasets from web crawl data. *CoRR*, abs/1911.00359. ds: ccnet; citation sum: 2448.

Williams, A., Nangia, N., and Bowman, S. (2018a). A broad-coverage challenge corpus for sentence understanding through inference. In Walker, M., Ji, H., and Stent, A., editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. ds: multi-nli; citation sum: 3575; expansion of: GLUE.

Williams, A., Nangia, N., and Bowman, S. R. (2018b). A broad-coverage challenge corpus for sentence understanding through inference. volume 1, page 1112 – 1122. Cited by: 2652.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. page 38 – 45. Cited by: 8003.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2021). mt5: A massively multilingual pre-trained text-to-text transformer. page 483 – 498. Cited by: 1294; All Open Access, Green Open Access, Hybrid Gold Open Access.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical attention networks for document classification. page 1480 – 1489. Cited by: 4354; All Open Access, Hybrid Gold Open Access.

Zellers, R., Bisk, Y., Schwartz, R., and Choi, Y. (2018). SWAG: A large-scale adversarial dataset for grounded commonsense inference. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics. ds: SWAG; citation sum: 44198.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2023). Benchmarking large language models for news summarization. ds: news summarization; citation sum: 233.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., and Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, 12:39 – 57. Cited by: 131.

Zhang, X., Zhao, J. J., and LeCun, Y. (2015). Character-level convolutional networks for text classification. *CoRR*, abs/1509.01626. ds: AG's news; citation sum: 233.

Zhang, Y., Zhong, V., Chen, D., Angeli, G., and Manning, C. D. (2017). Position-aware attention and supervised data improve slot filling. In Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics. ds: TACRED; citation sum: 2221.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2020). Ernie: Enhanced language representation with informative entities. page 1441 – 1451. Cited by: 884.

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27. ds: bookcorpus; citation sum: 44198.

A Search queries

The following search query has been used to filter for all ACL journals (which can be seen on the ACL website[9]) and filter for the past 15 years:

```
SRCTITLE ( ( association AND for AND computational AND linguistics ) OR arabicnlp OR ( computational AND natural AND language AND learning ) OR ( empirical AND methods AND in AND natural AND language AND processing ) OR ( international AND conference AND on AND spoken AND language AND translation ) OR ( workshop AND on AND semantic AND evaluation ) OR ( joint AND conference AND on AND lexical AND computational AND semantics ) OR ( workshop AND on AND statistical AND machine AND translation ) ) AND PUBYEAR > 2009 AND PUBYEAR < 2025
```

For the past 5 years and 2 years, the year in `PUBYEAR > 2009` would be changed to 2019 and 2022, respectively. For each individual search query, the results were sorted by citations and then the top 25 results were exported. The query was done on the 25th of April, 2025.

B The spreadsheet setup and analysis code

The spreadsheet mentioned in the paper, along with the scripts used to generate any plots or statistics based on the paper are available here[10][11] or in the repository associated with this research project. Instructions are also available in the codebase on how to generate plots using data from the spreadsheet. It is to be noted however that the spreadsheet has other data from other academic journals/organizations (CVPR, TPAMI, Neurips and AIII), but only the ACL data should be considered.

The spreadsheet is organized in multiple tabs: *Tab 1*, *Tab 2*, *Dataset DB*, *Tab 3*, *Dataset Leaderboard*, *Overall Statistics*, *ACL Statistics* and other tabs that are not used by this project. Tab 1, 2 and 3 are the same as described in Section 2. *Dataset DB* was used in order to aggregate all of the datasets gathered from *Tab 2*. *Dataset Leaderboard* was used to gather all the dataset papers to be read and papers marked with gray are excluded because they are not in the top 20 (either because their citation sum is not high enough, or because there was an expansion of a benchmark paper into multiple papers). *Overall Statistics* and *ACL Statistics* are used to select the papers based on either their occurrences or their citation sum.

The code to generate the plots was written in Python and generates graphs and statistics based on CSVs that can be downloaded from the spreadsheet. Instructions on how to used in can be found on the `README.md` file within the codebase. The main analyzers used were `annotation_statistics_analyzer`, `documntation_completeness_analyzer`, `head-to-head_analyzer`, `missing_information_analyzer`, `paper_dataset_stats_analyzer` and `top_dataset_frequency_analyzer`.

[9] https://aclanthology.org/

[10] https://docs.google.com/spreadsheets/d/16MkuS-upEQxkAj-poZO5ggPqmu_UIDbwi7HWS3-21HE/edit?usp=sharing

[11] https://github.com/Gargant0373/DatasetAnalysis

C The annotation schema

Each column dropdown has a "Unsure", "No information" and "Not applicable" option, unless otherwise stated. "Unsure" signifies the entry is marked for discussion for the next meeting.

"No information" means the author does not give any information about this question and "Not applicable" means this question does not make sense to be asked (e.g. no reason to ask ourselves about the overlap metric if there is no overlap for annotations). "No" means the author has stated explicitly the is no. . . (e.g. it was stated that no prescreening was done).

Rules of thumb:

- If the dataset is a benchmark that contains multiple datasets, report on each dataset within the paper (each dataset within the benchmark would count as 1 dataset for the top 20 within that period).
- If a dataset X from a benchmark Y is composed of a collection of datasets, answer question about collection as a whole based on what dataset X says about all datasets. Can also look at what benchmark Y says about dataset X as a whole.

Schema items:

- **Empty** - "Yes" if there is no information about the dataset (i.e. author does not reference it and is not findable on the web/private dataset etc.), "No" if there is information available, "Unsure" if it might be out of scope, "Benchmark" if it is a benchmark (to signify it was expanded)
- **Outcome** - what was the purpose of this dataset? I.e. ImageNet made for object recognition
- **Human Labels** - "Yes for all" all of the items collected were annotated; "Yes for some" some items annotated, but others (e.g. in the dev set etc.) left unannotated; "No / Machine labelled" item unannotated (e.g. Wikipedia text for pretraining LMs) or annotated by a machine (synthetic means), "Unknown" the author does not specify how the dataset was annotated, "Implicit Yes" We know based on the subject matter that it had to be human labeled (e.g. patient data)
- **OG Labels** - "OG" they made the labels themselves (through crowdworkers etc.) "External" labels were taken from another place already available, "Not Labelled" there are no annotations (the latter replaces "Not applicable")
- **Label source** - where were the labels taken from? MTurk, other crowdsourcing websites, students, no information, not applicable etc. (this could be turned into a dropdown later, for now just be consistent for your publication)
- **Labeller population rationale** - did they give a rationale for why they picked those specific labellers?

- **Prescreening** - "Generic skill based" they state that the workers were filtered on their skills i.e. basic spanish skills etc. "Previous platform performance" hired based on how good they were on the platform i.e. 97% HIT accuracy, "Project-specific prescreening" e.g. inviting good crowdworkers back, doing their own prescreening
- **Compensation** - how were the workers compensated? We assume hiring somebody on a crowdsourcing platform implies money. If annotated by authors, put "authorship". Options are "Money", "Authorship", "Course Credit", "Other Compensation", "Volunteer", "No information", "Not applicable", "Unsure".
- **Training** - whether annotators receive interactive training for this specific annotation task / research project - simple formal instructions are not training
- **Formal instructions** - whether or not annotators received formal instructions on how to annotate the data
- **Total labellers** - How many people annotated the items? "Not applicable" and "No information" are valid options.
- **Annotators per item** - do the authors say how many authors they had per label? Can be average etc.
- **Label threshold** - what is the minimum amount of labels each item needed?
- **Overlap** - did multiple annotators work on the same item? Sometimes you could theoretically infer that they had at most one annotator per item, but if it is not clear enough use "no information"
- **Overlap synthesis** - in what manner was the overlap solved? "Qualitative" (discussion), "Quantitative" (no discussion), "Other" Synthesis type - what method did they use? E.g. majority vote for quantitative or discussion for qualitative
- **Discussion** - was there a discussion among the annotators? (sometimes researchers look at the annotation)
- **IRR** - was there IRR reported if there was overlap? If no overlap, put "not applicable".
- **Metric** - if IRR was reported, what was the metric? E.g. F1 or Cohen Kappa etc. Put "not applicable" only if there is no overlap (i.e. 1 annotator, machine labelled)
- **A priori annotation schema** - "yes", "yes, from external source" "no" (if they make it up as they go, like iNaturalist)
- **Annotation schema rationale** - did they put any thought into why they use this schema?
- **Item population** - briefly describe the item population
- **Item population rationale** - why did they go for this item population?
- **Item source** - where did they take the items from?
- **A priori sample size** - did they decide the sample size before they started collecting the items?
- **Item sample size rationale** - why did they choose to collect this amount of items?
- **Link to Data** - Was a link to the dataset provided? "Yes", "Yes, but broken", "No", "Not applicable" (when the dataset is created synthetically). "No information" means "No" here.

D Formulas or figures too big to be introduced in the main paper

This section contains some formulas or figure that were too big to be introduced in the main content of the paper.

$$\text{missing}_{\text{period}} = \frac{\sum_{i=1}^{20} \text{MissingFields}_i}{\sum_{i=1}^{20} \text{ApplicableFields}_i} \times 100 \tag{1}$$

where:

- MissingFields_i is the number of fields in dataset i (within the period) that contain missing information, such as *"No information"*, *"Unknown"*, or *"Unsure"*.
- $\text{ApplicableFields}_i$ is the number of fields in dataset i that are applicable (i.e., not marked as *"Not applicable"*).

$$Score_{d,t} = \sum_{p \in P_{d,t}} \text{Citations}(p) \tag{2}$$

where:

- $\text{Score}_{d,t}$ is the usage count for dataset d in time period t.
- $P_{d,t}$ is the set of papers in time period t that used dataset d.
- $\text{Citations}(p)$ is the number of citations of paper p.

E Papers read in order to extract the datasets

The next tables mention which papers have been read to extract the datasets, per period. Those are Tables 1, 2, 3. The citations are slightly higher than in the spreadsheet used, as the data has been re-extracted on the 22nd of June.

F Datasets extracted along with citation sums

The next tables mention which dataset papers have been read and annotated, per period. Those are Tables 4, 5, 6. The citation sums are slightly higher than in the spreadsheet used, as the data has been re-extracted on the 22nd of June.

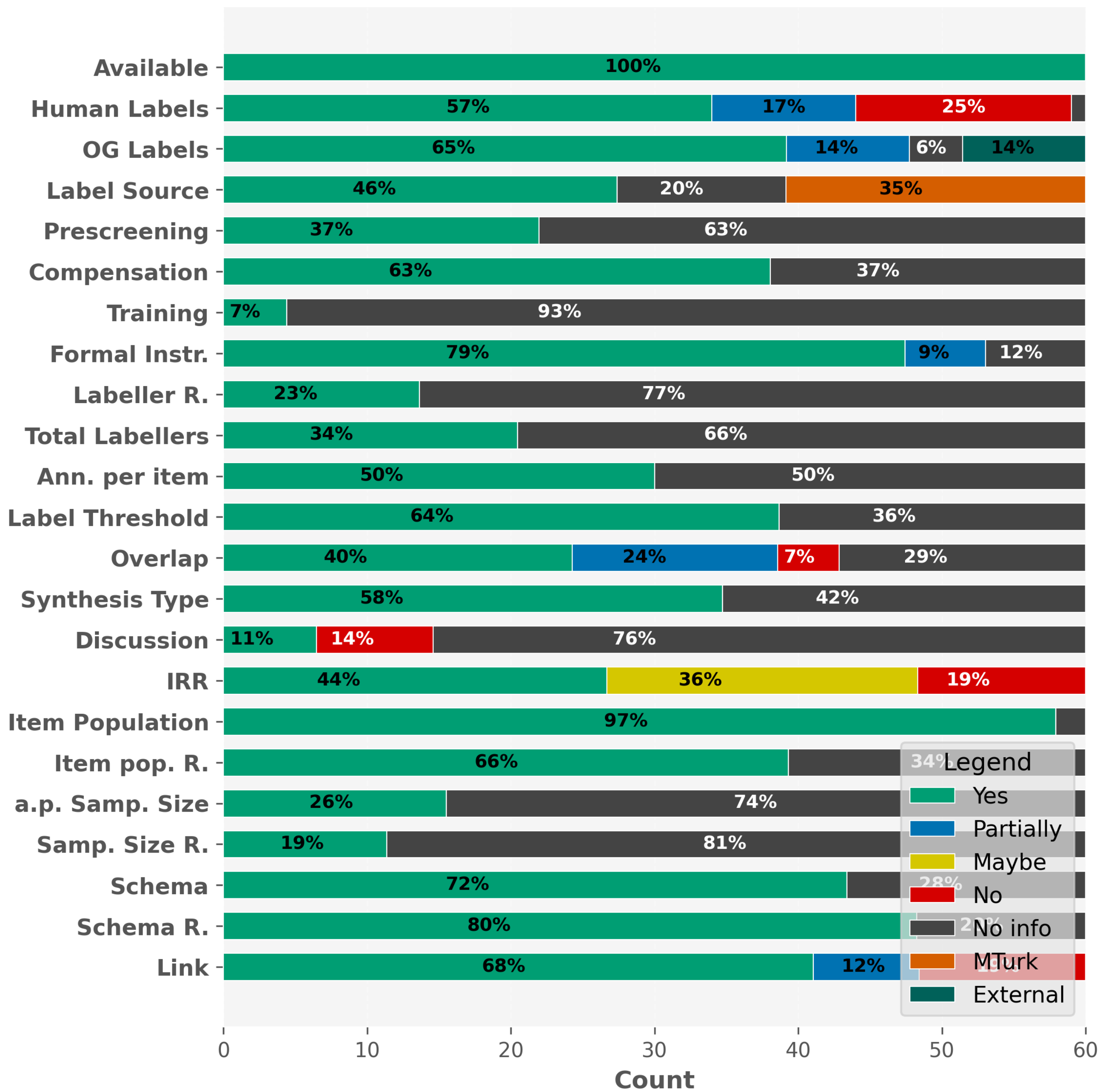

Figure 8: Quantitative summary of the annotation of the most used dataset papers indifferent of period. ”Not applicable” is excluded.

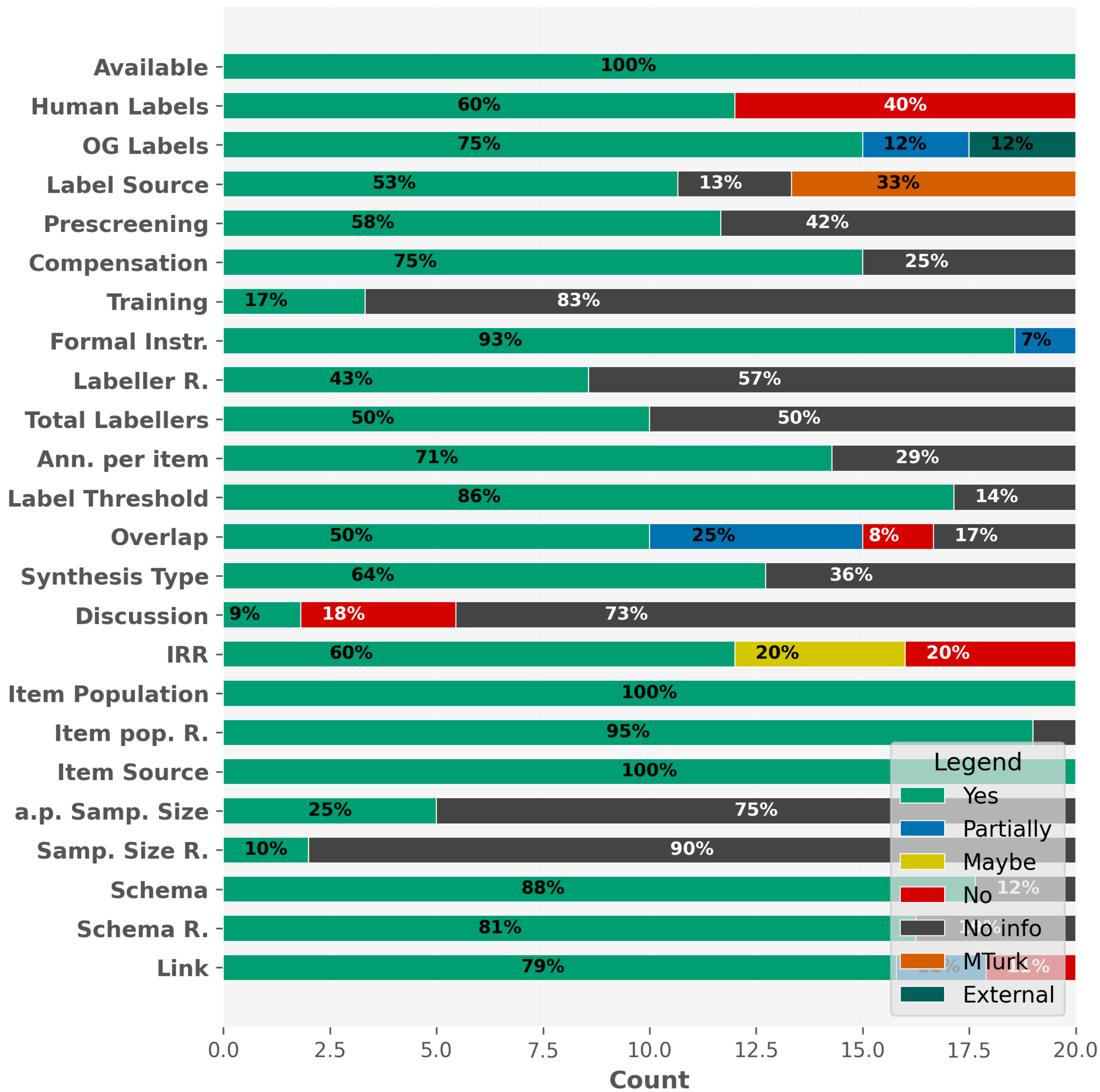

Figure 9: Quantitative summary the annotation of the most used dataset papers used in the past 2 years. ”Not applicable” is excluded.

Dataset	Count
strategyqa	4
svamp	3
commonsens	3
2wikimultihop	2
gsm8k	2
multiarith	2
nq	2
singleeq	2
asdiv	2
opendomainq	1
qrecc	1
qags	1
proofwriter	1
prontoqa	1
popqa	1
news summa	1
opendialkg	1
nyt	1
realnews	1
naturalquestic	1

(a) Top 20 Datasets – Period 2

Dataset	Count
glue	4
squad	4
multi-nli	3
snli	3
lama	2
wmt19	2
wmt16	2
webnlg 2017	2
tacred	2
sick	2
nq	2
triviaqa	2
english wikipe	2
conll-2003	2
conll-2002	2
ccnet	2
ag's news	2
1b word	2
dart	2
sts2012	1

(b) Top 20 Datasets – Period 5

Dataset	Count
conll-2003	5
squad	4
cnn/dm	3
snli	3
sst-1	3
subj	3
wmt14	3
mpqa	2
yahoo answe	2
wmt15	2
trec	2
multi-nli	2
mr	2
sick	2
glue	2
cr	2
conll-2002	2
amazon-5	2
1b word	2
english wikipe	2

(c) Top 20 Datasets – Period 15

Dataset	Count
squad	9
conll-2003	7
glue	6
snli	6
multi-nli	5
1b word	4
strategyqa	4
subj	4
sick	4
cnn/dm	4
conll-2002	4
nq	4
english wikipe	4
ag's news	4
squadv2	3
yahoo answe	3
sst-1	3
wmt16	3
svamp	3
trec	3

(d) Top 20 Datasets – Overall

Figure 10: Top 20 dataset usage frequency grouped by period and overall.

Title	Citations	Citation
SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions	427	(Wang et al., 2023b)
Lost in the Middle: How Language Models Use Long Contexts	297	(Liu et al., 2024)
G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment	263	(Liu et al., 2023)
Crosslingual Generalization through Multitask Finetuning	218	(Muennighoff et al., 2023b)
When Not to Trust Language Models: Investigating Effectiveness of Parametric and Non-Parametric Memories	213	(Mallen et al., 2023a)
Can Large Language Models Be an Alternative to Human Evaluation?	194	(Chiang and Lee, 2023)
HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models	169	(Li et al., 2023a)
FACTSCORE: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation	164	(Min et al., 2023)
MTEB: Massive Text Embedding Benchmark	158	(Muennighoff et al., 2023a)
Evaluating Object Hallucination in Large Vision-Language Models	155	(Li et al., 2023b)
Towards Reasoning in Large Language Models: A Survey	155	(Huang and Chen-Chuan Chang, 2023)
Measuring and Narrowing the Compositionality Gap in Language Models	151	(Press et al., 2023)
In-Context Retrieval-Augmented Language Models	151	(Ram et al., 2023)
Revisiting Relation Extraction in the era of Large Language Models	139	(Wadhwa et al., 2023)
Benchmarking Large Language Models for News Summarization	131	(Zhang et al., 2024)
SELFCHECKGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models	130	(Manakul et al., 2023)
Plan-and-Solve Prompting: Improving Zero-Shot Chain-of-Thought Reasoning by Large Language Models	127	(Wang et al., 2023a)
SemEval-2023 Task 10: Explainable Detection of Online Sexism	122	(Kirk et al., 2023)
Active Retrieval Augmented Generation	118	(Jiang et al., 2023)
Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them	117	(Suzgun et al., 2023b)
Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering	116	(Siriwardhana et al., 2023)
RARR: Researching and Revising What Language Models Say, Using Language Models	115	(Gao et al., 2023)
Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes	113	(Hsieh et al., 2023)
SemEval-2023 Task 3: Detecting the Category, the Framing, and the Persuasion Techniques in Online News in a Multi-lingual Setup	107	(Piskorski et al., 2023)
Making Large Language Models Better Reasoners with Step-Aware Verifier	103	(Li et al., 2023c)

Table 1: Papers, citation count and their references from past 2 years

Title	Citations	Citation
Transformers: State-of-the-Art Natural Language Processing	8003	(Wolf et al., 2020)
BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension	5688	(Lewis et al., 2020)
Unsupervised cross-lingual representation learning at scale	3349	(Conneau et al., 2020)
SimCSE: Simple Contrastive Learning of Sentence Embeddings	1886	(Gao et al., 2021b)
Prefix-tuning: Optimizing continuous prompts for generation	1885	(Li and Liang, 2021)
Dense passage retrieval for open-domain question answering	1847	(Karpukhin et al., 2020)
The Power of Scale for Parameter-Efficient Prompt Tuning	1674	(Lester et al., 2021)
Spanbert: Improving pre-training by representing and predicting spans	1377	(Joshi et al., 2020)
Transformer-XL: Attentive language models beyond a fixed-length context	1330	(Dai et al., 2020)
mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer	1294	(Xue et al., 2021)
Don't stop pretraining: Adapt language models to domains and tasks	1268	(Gururangan et al., 2020)
Multilingual denoising pre-training for neural machine translation	1261	(Liu et al., 2020)
Multimodal transformer for unaligned multimodal language sequences	1249	(Tsai et al., 2020)
Stanza: A Python natural language processing toolkit for many human languages	1086	(Qi et al., 2020)
CodeBERT: A pre-trained model for programming and natural languages	1070	(Feng et al., 2020)
Making pre-trained language models better few-shot learners	1056	(Gao et al., 2021a)
How can we know what language models know?	995	(Jiang et al., 2020)
Energy and policy considerations for deep learning in NLP	973	(Strubell et al., 2020)
What does BERT learn about the structure of language?	971	(Jawahar et al., 2020)
Exploiting cloze questions for few shot text classification and natural language inference	953	(Schick and Schütze, 2021)
A primer in bertology: What we know about how bert works	939	(Rogers et al., 2020)
AUTOPROMPT: Eliciting knowledge from language models with automatically generated prompts	893	(Shin et al., 2020)
ErniE: Enhanced language representation with informative entities	884	(Zhang et al., 2020)
BLEURT: Learning robust metrics for text generation	883	(Sellam et al., 2020)
GLM: General Language Model Pretraining with Autoregressive Blank Infilling	824	(Du et al., 2022)

Table 2: Papers, citation count and their references from past 5 years

Title	Citations	Citation
BERT: Pre-training of deep bidirectional transformers for language understanding	45837	(Devlin et al., 2019b)
GloVe: Global vectors for word representation	27640	(Pennington et al., 2014)
Learning phrase representations using RNN encoder-decoder for statistical machine translation	12056	(Cho et al., 2014)
Convolutional neural networks for sentence classification	8279	(Kim, 2014)
Transformers: State-of-the-Art Natural Language Processing	8003	(Wolf et al., 2020)
Deep contextualized word representations	6983	(Peters et al., 2018)
Sentence-BERT: Sentence embeddings using siamese BERT-networks	6520	(Reimers and Gurevych, 2019)
Recursive deep models for semantic compositionality over a sentiment treebank	6276	(Socher et al., 2013b)
The stanford CoreNLP natural language processing toolkit	5893	(Manning et al., 2014)
BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension	5688	(Lewis et al., 2020)
Neural machine translation of rare words with subword units	4567	(Sennrich et al., 2016)
Hierarchical attention networks for document classification	4354	(Yang et al., 2016)
Effective approaches to attention-based neural machine translation	4182	(Luong et al., 2015)
Learning word vectors for sentiment analysis	4108	(Maas et al., 2011)
SQuad: 100,000+ questions for machine comprehension of text	4101	(Rajpurkar et al., 2016b)
Unsupervised cross-lingual representation learning at scale	3349	(Conneau et al., 2020)
Linguistic regularities in continuous spaceword representations	2699	(Mikolov et al., 2013)
Get to the point: Summarization with pointer-generator networks	2671	(See et al., 2017)
A broad-coverage challenge corpus for sentence understanding through inference	2652	(Williams et al., 2018b)
Neural architectures for named entity recognition	2642	(Lample et al., 2016)
A large annotated corpus for learning natural language inference	2589	(Bowman et al., 2015a)
Attention-based LSTM for aspect-level sentiment classification	2220	(Wang et al., 2016)
Fairseq: A fast, extensible toolkit for sequence modeling	2207	(Ott et al., 2019)
Bag of tricks for efficient text classification	2186	(Joulin et al., 2017)
SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing	2151	(Kudo and Richardson, 2018)

Table 3: Papers, citation count and their references from past 15 years

Dataset name	Citation sum	Citation
strategyQA	403	(Geva et al., 2021)
Self-Instruct	392	(Wang et al., 2023c)
superni	392	(Wang et al., 2022)
CommonSenseQA	297	(Talmor et al., 2019)
SVAMP	297	(Patel et al., 2021)
NaturalQuestions-Open	234	(Lee et al., 2019)
AG's news	233	(Zhang et al., 2015)
news summarization	233	(Zhang et al., 2023)
QAGS	233	(Wang et al., 2020)
SummEval	233	(Fabbri et al., 2021)
Topical-Chat	233	(Gopalakrishnan et al., 2023)
nq	231	(Kwiatkowski et al., 2019)
MultiArith	198	(Roy and Roth, 2015)
entityquestions	197	(Sciavolino et al., 2021)
GSM8k	197	(Cobbe et al., 2021)
MultiArith	197	(Roy and Roth, 2015)
PopQA	197	(Mallen et al., 2023b)
SingleEQ	197	(Koncel-Kedziorski et al., 2015)
2WikiMultiHop	195	(Ho et al., 2020)
ASDiv	195	(Miao et al., 2020)
WritingPrompts	176	(Fan et al., 2018)

Table 4: Most used datasets in the past 2 years, with their citation sum and citation.

Dataset name	Citation sum	Citation
SQuAD	5946	(Rajpurkar et al., 2016c)
GLUE	4092	(Wang et al., 2018b)
SNLI	3757	(Bowman et al., 2015b)
dart	3592	(Radev et al., 2020)
multi-nli	3575	(Williams et al., 2018a)
nq	3121	(Kwiatkowski et al., 2019)
TriviaQA	3121	(Joshi et al., 2017)
SICK	2660	(Marelli et al., 2014)
WebNLG 2017	2653	(Gardent et al., 2017)
ccnet	2448	(Wenzek et al., 2019)
1B word	2354	(Chelba et al., 2013)
WMT19	2284	(Barrault et al., 2019)
English Wikipedia	2221	(Unknown, nda)
TACRED	2221	(Zhang et al., 2017)
AG's news	2160	(Zhang et al., 2015)
WMT16	2002	(Bojar et al., 2016)
CoNLL-2002	1838	(Tjong Kim Sang, 2002)
CoNLL-2003	1838	(Tjong Kim Sang and De Meulder, 2003)
lama	1816	(Petroni et al., 2019)
ANLI	1801	(Nie et al., 2020)

Table 5: Most used datasets in the past 5 years, with their citation sum and citation.

Dataset name	Citation sum	Citation
CoNLL-2003	86025	(Tjong Kim Sang and De Meulder, 2003)
English Wikipedia	71400	(Unknown, nda)
SQuAD	60674	(Rajpurkar et al., 2016c)
multi-nli	47482	(Williams et al., 2018a)
CoLA	47482	(Warstadt et al., 2018)
SST-2	47482	(Socher et al., 2013a)
MRPC	47482	(Dolan and Brockett, 2005)
STSb	47482	(Cer et al., 2017)
QQP	47482	(Quora, 2012)
RTE	47482	(Unknown, ndb)
GLUE	47482	(Wang et al., 2018b)
WNLI	47482	(Davis et al., nd)
bookcorpus	44198	(Zhu et al., 2015)
SQuADv2	44198	(Rajpurkar et al., 2018b)
SWAG	44198	(Zellers et al., 2018)
ace-2003	27202	(Mitchell et al., 2004)
gigaword-5	27202	(Parker et al., 2011)
mc	27202	(Miller and Charles, 1991)
MUC-7	27202	(Chinchor, 2001)
RG	27202	(Rubenstein and Goodenough, 1965)
RW	27202	(Luong et al., 2013)

Table 6: Most used datasets in the past 15 years, with their citation sum and citation.