Annotation Practices in Societally Impactful Machine Learning Applications

What are these automated systems actually trained on?

Simona Cristina Lupșa[1]
Supervisors: 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: Simona Cristina Lupșa
Final project course: CSE3000 Research Project
Thesis committee: Dr. Cynthia Liem, Andrew M. Demetriou, Jie Yang

Abstract

This study examines dataset annotation practices in influential NeurIPS research. Datasets employed in highly cited NeurIPS papers were assessed based on criteria concerning their item population, labelling schema, and annotation process. While high-level information, such as the presence of human labellers and item population, is present in most cases, procedural details of the annotation process are poorly reported. Notably, 48% of datasets lack details on annotator training, 43% omit inter-rater reliability, and 28% are not publicly accessible. Temporal comparisons show minor improvements, but no substantial progress in reporting annotation methodology. A complementary analysis of 49 NeurIPS papers published since 2020 shows that researchers often discuss the broader impact of their work, yet do not include datasets or their annotations in these assessments. These findings highlight a lack of standardisation in annotation reporting and call for more robust practices that ensure transparency, auditability, and reproducibility in machine learning research.

1 Introduction

Machine Learning (ML) has become a widely influential research field and a dominant industry in recent years. A fundamental component of training and deploying ML models consists of labelled datasets, with the annotations serving as the "ground truth" upon which the model is built. As a result, data annotations directly influence the fairness, validity, and trustworthiness of a model.

Despite this, the research community in the field generally prioritises other criteria to assess the quality of a study or a model, such as performance, generalisation and novelty [1]. ML research either refrains from discussing ethical matters, such as the societal impact of novel discoveries and how the ground truths come to be, or does not address these concerns at all [1]. Moreover, past research on data selection and dataset annotation shows that, in many cases, ground truth creation suffers from insufficient reporting [2] and considerable pitfalls [3, 4]. This inconsistency poses a growing risk, as a lack of transparency when it comes to ground truth creation limits the capacity to assess its quality. Since models are increasingly deployed in high-stakes environments, it is essential to evaluate dataset annotation practices.

Having the objective of providing a robust analysis, this study focuses on the Neural Information Processing Systems (NeurIPS) conference. NeurIPS is, according to Google Scholar metrics, one of the most prolific venues in the field [5]. It covers a wide range of branches within ML, such as deep learning and reinforcement learning, while also actively encouraging research into socioeconomic aspects of ML, including fairness and safety [6]. Taking into account the influence of NeurIPS, it is particularly relevant to examine the annotation practices behind widely used datasets within the venue.

As such, this study addresses the following research question: **"What are the data collection and reporting practices for annotation in societally impactful ML applications from the NeurIPS venue?"**. To find a thorough response, Table 1 presents the sub-questions that provide a framework to support the process.

The remainder of this thesis is structured as follows. Firstly, methodology is described in section 2. Statistical results are illustrated in section 3, followed by a more elaborate

SQ1	How do NeurIPS researchers assess the quality of the datasets that they use for their models? Do they explicitly take annotations into account?
SQ2	What or who is labelling the datasets?
SQ3	What are the relevant criteria for evaluating the transparency of dataset creation?
SQ4	Do the datasets fit the criteria established by **SQ3**?

Table 1: Research sub-questions.

discussion in section 4. Subsequently, section 5 discusses the limitations of this project. At last, the analysis is concluded in section 6.

All materials used to conduct the study are publicly available and listed in Appendix A.

2 Methodology

Based on the research question, the most suitable approach to this study is conducting a structured analysis of NeurIPS publications. This allows for examining the current research landscape within the venue and observing what datasets are frequently encountered. The annotation process of the respective datasets will be documented, with the aim of showcasing the most common patterns arising from labelling and reporting practices.

With this in mind, 75 papers are extracted, 25 each from the last two, five, and fifteen years (starting from 2023, 2020 and 2010, respectively), using citation counts to identify the most influential studies within each period. The 2025 Conference on Neural Information Processing Systems is scheduled for December, which is why the last year to be considered is 2024.

The research process is showcased in a collaborative spreadsheet that combines the efforts of all members of the peer group. The spreadsheet is available in Appendix A. More details regarding the methodology are included in Appendix B.

2.1 Defining "Societally Impactful"

Paper selection only began after a consensus was reached regarding what it means for an ML application to be societally impactful. As such, the h5-index was chosen as a general quantitative measure of relevance. It is defined as "the largest number h, such that h articles from the past five years that have at least h citations each" [7]. This index, therefore, highlights venues that have consistently cited work and is a convenient first glance at the influence of a publication.

Following venue selection, papers are extracted based on the number of citations. While citation count cannot effectively measure the quality of a research paper, it does show that the study had a meaningful contribution, albeit possibly negative, to the scientific process and the literature of the field. Moreover, it is an intuitive extension to the h5-index used for choosing research venues.

In the scope of this project, a societally impactful ML application is defined as a piece of academic work published in a leading field venue, with a high citation count as a testament of recognition within the research community.

2.2 Extracting NeurIPS Papers

As previously mentioned in this chapter, 75 NeurIPS papers from three distinct periods were extracted from Scopus[1], ordered by citation count. Solely using Scopus was a decision taken among the peer group, driven by a couple of considerations. Firstly, different versions and citation counts of the same paper might occur in different databases, which can introduce inconsistencies in the final compilations. Additionally, Scopus is easily accessible for TU Delft students, hence it was an appropriate database choice.

For each paper, the title, publication year, Digital Object Identifier (DOI), and Scopus link were exported. As NeurIPS does not issue DOIs for their published research papers, those were manually extracted from the arXiv preprint[2]. While arXiv can include multiple versions of the same paper, the DOI is used solely as the unique identifier of a paper and has no other purpose. Moreover, NeurIPS papers are publicly accessible on their website[3], hence the version that will be further used is the one available there.

After accounting for duplicates, which were only examined once, and excluding one study that did not use any dataset, a total of 71 NeurIPS papers remain for further analysis.

2.3 Gathering Datasets

Paper selection was followed by compiling the list of datasets used in the respective studies. For this stage, a set of rules was decided to ensure consistency in the datasets that will be included. The purpose of this study is to assess datasets that were either released by the selected papers or used in the training, development, or evaluation of the issued models. As such, datasets that occur as related work on the specific tasks were excluded. Additionally, datasets solely mentioned as possible alternatives, without eventually being used, were not considered for further analysis.

The final step of this stage was deduplication. The most frequent reason for duplicate data was naming, as the same dataset happened to be referenced differently across multiple papers. As a result of this process, reading the NeurIPS papers has led to a total of 351 unique datasets.

2.4 Analysing Dataset Papers

At last, the concluding step of this review is analysing the sources of the datasets, to evaluate their annotation process. Following discussions among the peer group and supervisors, and considering that the midpoint of the allocated nine-week period had been reached, a target

[1] `https://www.scopus.com/`

[2] `https://arxiv.org/`

[3] `https://papers.nips.cc/`

of 60 datasets was set. A method for sorting the extracted datasets has been established in order to assess their relative importance. Based on this, the following weighted formula was used for each dataset and time period:

$$\text{Score}_{d,t} = \sum_{p \in P_{d,t}} \text{Citations}(p)$$

where:

- $\text{Score}_{d,t}$ is the score 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.

Having calculated this, the last remaining step of the methodology was analysing the top 20 datasets from each period according to the computed score. Subsequently, the criteria previously defined by Geiger et al. [2] were adopted, with several additions that would support a more robust assessment of quality and transparency. These additions consist of criteria developed independently by the peer group, capturing aspects of transparency and dataset construction that were not addressed in the original framework. In total, there are 27 criteria, 12 of which were also employed in [2].

Addressing **SQ3**, the final set of criteria was organised into three focal points. The first, *Items*, concerns the description of the item population and the reasoning behind it. The second, *Annotators and Labelling Process*, addresses who the labellers are and how they were selected. and how they interacted during the annotation process, if at all. The third focal point, *Annotation Schema*, elaborates on the rationale behind the chosen schema and the specific task the dataset is designed to support. The exact set of criteria, alongside specific details regarding how to assess them, can be found in subsection B.2 of Appendix B.

2.5 Inspecting how NeurIPS Papers Discuss Annotations

In 2020, NeurIPS mandated that all papers include a "Broader Impact" section in order to be accepted. This only lasted for one year; as of 2021, a separate section was no longer compulsory. However, author instructions did mention that a discussion on societal impact ought to be included [8]. Addressing these concerns, while nevertheless an approach worthy of praise, gives readers an opportunity to observe what NeurIPS researchers prioritise when faced with the demand of responsible research.

With the aim to discover how important the ground truth is in the NeurIPS landscape, a separate analysis was conducted on a subset of the venue papers extracted from Scopus. This consisted of two stages. Firstly, the "Broader Impact" section of a paper was inspected, if it existed, to observe whether researchers assess the impact of their work on human lives. Additionally, it was noted whether the datasets used or their annotations were mentioned as possibly influencing this impact. Secondly, the papers were fully scanned to examine how datasets and annotations are generally approached.

Papers published from 2020 onwards were examined using a set of criteria designed to assess the level of attention given to dataset and annotation transparency. The complete set of criteria, with rules of inspection, is described in subsection B.3 of Appendix B.

3 Findings

3.1 Dataset Annotation Analysis

This section effectively answers *SQ4*.

3.1.1 Overall Statistics

General statistics, compiling all three periods, can be seen in Figure 1.

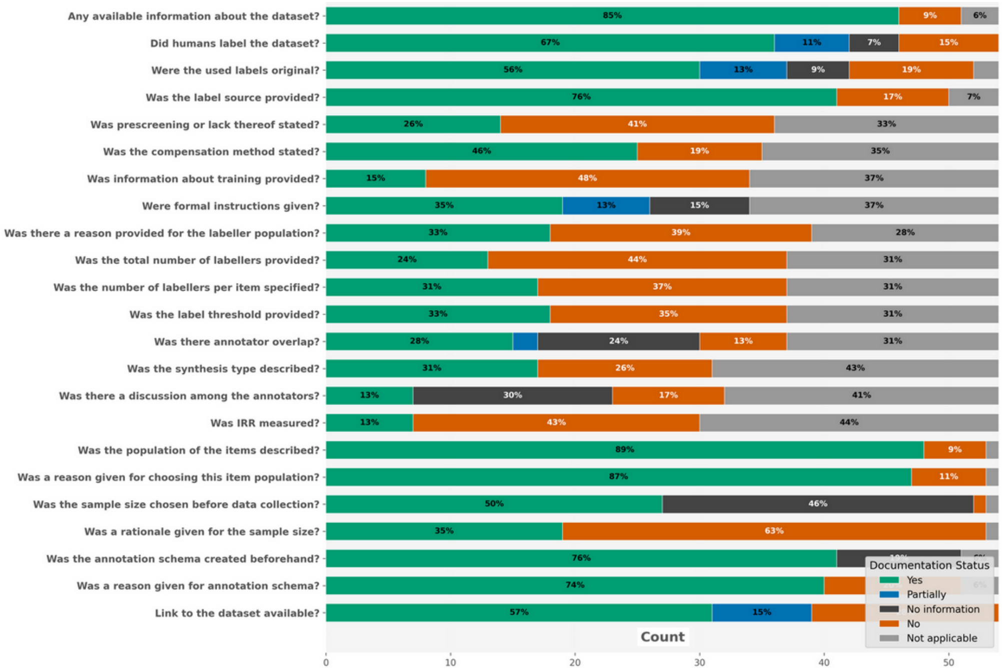


Figure 1: Overall statistics after dataset analysis.

Addressing **SQ2**, Figure 1 illustrates that a majority of the datasets were entirely human-labelled, namely 67%. Another 11% of datasets were labelled both by humans and machines, while 15% were either exclusively machine-labelled or unlabelled. For the remaining 7%, there is no information regarding who or what was involved in dataset annotation.

Some particular patterns already arise from observing these figures. On the one hand, there is a reliable source of information for the majority of datasets, as seen in the first bar of the chart ("Any available information about the dataset?"). In this particular bar, the

"Not applicable" category refers to benchmarks, all of which were found to have accessible documentation in the analysis. As a result, information was available for 91% of the datasets. Additionally, 93% of the papers explicitly state whether the datasets were annotated by humans. Finally, a large proportion of the dataset papers describe the item population (89%) and offer a rationale for it (87%). This indicates that authors generally provide the more high-level aspects of dataset creation.

On the other hand, several important aspects regarding annotation are poorly documented. Generally, taking into account only cases where the criteria are applicable, details concerning the annotation process itself, such as prescreening and training of annotators, rationale for labeller selection, label thresholds and Inter-Rater Reliability (IRR) are predominantly missing. Specifically, 41% of the datasets did not include information on prescreening, 48% on training, 39% on labeller selection rationale, 35% on label threshold and 43% on IRR. This suggests that there is a general lack of documentation of the labelling process. Additionally, 28% of the datasets themselves are inaccessible online. Even though this is by no means a majority, it nevertheless raises transparency and reproducibility concerns.

Overall, authors do report general information about the datasets, such as the presence of human labels, the item population and its reasoning. However, pitfalls are evident when it comes to the annotation process, which is weakly reported. Specifically, the three fields with the least amount of documentation are annotator training, discussion between annotators, and IRR.

3.1.2 Comparing Periods

For each period, the top 20 datasets were inspected. Table 2 presents the datasets that occurred in multiple time frames, showing that the sets of datasets across periods are almost disjoint. Hence, analysing each period separately reveals more information on how transparency levels fluctuate.

Common 2-5-15	**Common 2-5**	**Common 5-15**	**Common 2-15**
COCO [9]	GSM8K [10]	CIFAR-10 [11] ImageNet 2012 [12]	No common datasets

Table 2: Common occurrences in the top 20 datasets across all periods.

Firstly, the lack of reliable information on datasets only concerns the 15-year period. For the remaining time frames, there was either a dataset paper, a website, or a *README* file in a repository linked to the authors, that discussed dataset creation. Consequently, the same rationale applies to reporting who or what labelled the datasets.

Another aspect concerning accessibility is whether the datasets themselves are available online. This is showcased in the last bar of the chart ("Link to the dataset available?"), with "Partially" meaning that a link is provided, but it is either not functional, or redirects to a website that does not actually publish the data. In this case, there are improvements across time frames. While in the 15-year period, 45% of the datasets were not made public at all, this percentage decreased to 30% in the 5-year period, and 10% in the 2-year period.

Further, the percentage of criteria for which no information was provided can be observed in Figure 2. At first glance, it might appear that the more recently used datasets are significantly more documented. Two other aspects support this statement. Firstly, the "Overall" bar, which illustrates that approximately 30% of information is missing across all datasets, is skewed upwards by datasets that were employed in the past. Secondly, as stated beforehand,Table 2 shows that the three periods have only a few datasets in common.

However, this assumption is not fully supported if cases where the criteria are not applicable are excluded. While it is clear that both the 15-year and the 5-year period show the same pitfalls previously discussed, those criteria are not applicable for a majority of the datasets used in the 2-year period. There are two main reasons for why this is the case. Firstly, some datasets from this period are either unlabelled or machine-labelled. As such, criteria that concern human annotators, such as prescreening or discussion, do not fit the context. Secondly, annotations for other datasets used in the 2-year period were taken from external sources that already had those labels. Although it is known that the datasets are human-labelled, criteria such as annotator discussion, label threshold, and implicitly all aspects that concern overlap, are not applicable for those cases. Considering solely the cases where the criteria do apply, the dataset papers used in the 2-year period still report the annotation process weakly. Information about annotator training, labellers per item and IRR is scarce in this period as well. While other aspects of the annotation process show some improvements compared to other periods, those are not substantial.

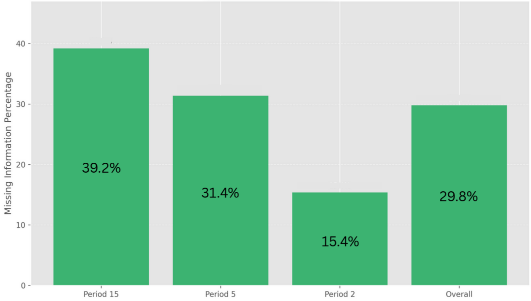


Figure 2: Percentage of fields with no information across periods.

Generally, documentation on datasets seems to have improved over time, albeit with some fluctuations in particular cases. However, these improvements are concentrated in high-level aspects, such as the reasoning behind the item population, where the labellers come from, and data availability. As is the case in the overall statistics, details about the annotation process are poorly reported across all three periods. While there are improvements in criteria such as total labellers and threshold, other details such as IRR and training remain overlooked. This indicates a lack of standardised reporting practices, leaving the inclusion of these details at the discretion of dataset authors.

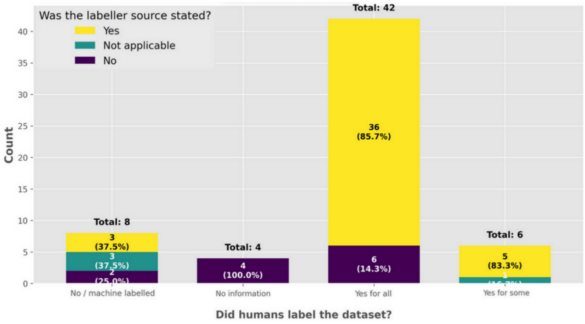


Figure 3: Pairwise analysis between human labels and label source.

Figures concerning each period can be inspected in Appendix D.

3.1.3 Pairwise Analysis

To examine the identified patterns in greater detail, cross-tabulations were conducted for selected pairs of criteria. This approach showcases how different fields interact, which in turn reveals potential inconsistencies in dataset documentation. Analysing these relationships therefore offers a more nuanced understanding of reporting practices.

To begin, for most datasets that involved human annotation, either fully or partially, authors clearly stated the source of their annotators. This is illustrated in Figure 3. The percentage is similar in these categories, indicating a somewhat consistent level of transparency, considering the difference in sample size. When it comes to datasets that were either machine-labelled or unlabelled, conclusions are difficult to draw, since the two are not differentiated.

The next analysis concerns annotator training and formal instructions. Preparatory measures are generally taken prior to annotation, in order to ensure consistency and reliability across labellers. Training and formal instructions are the procedures that were inspected throughout this analysis. They are examined together, in order to reveal the degree of attention allocated to this stage of the annotation process.

Interactive training is one of the most undocumented aspects of this analysis, as discussed beforehand. However, Figure 4 shows that preparation is still taking place, albeit in a different format. Among the 28 dataset papers that do not report on any form of training, 20 still state that written guidelines were provided to labellers. In contrast, all annotators who received training also benefitted from formal instructions. Nevertheless, this shows that dataset authors generally rely on forms of preparation that do not require high degrees of interaction with the labellers.

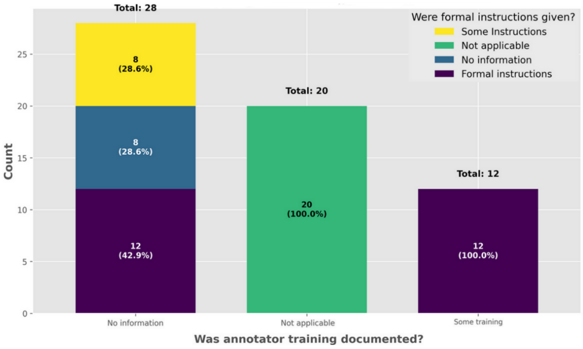


Figure 4: Pairwise analysis between annotator training and formal instructions.

3.2 Perceptions of Societal Impact within NeurIPS

This subsection provides an answer to **SQ1**. Overall statistics after analysing the venue papers can be observed in Figure 5.

These statistics also provide more insight for effectively answering **SQ1**. Out of the 49 venue papers analysed, 55% discussed the annotation of the datasets they used, or lack thereof, in varying degrees of detail. Another 37% briefly touched upon the contents of the datasets. With this in mind, it is reasonable to assume that NeurIPS researchers are generally attentive to the high-level aspects of how datasets are annotated.

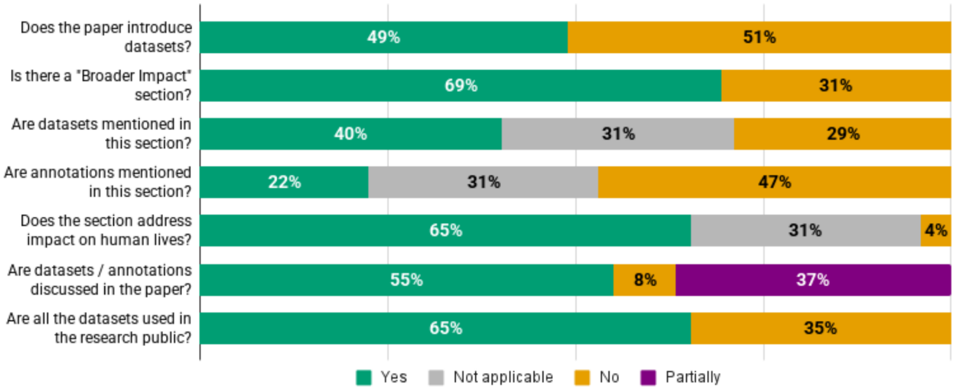


Figure 5: Overall statistics after reviewing NeurIPS papers.

A categorical majority of the papers clearly address the broader impact of the research through a dedicated section. This shows that addressing these topics is a practice that has persisted from 2021 onwards, when the respective section became a voluntary choice (only around 28% of the inspected papers were published in 2020).

However, datasets, and implicitly their annotation, are often not mentioned in the researchers' efforts to address the impact of their work. Even though most of the papers discuss potential effects on human lives, there is often no correlation made between how ML systems impact humans and the fact that training and experimental data constitute a fundamental component of said systems. This is illustrated in Figure 6.

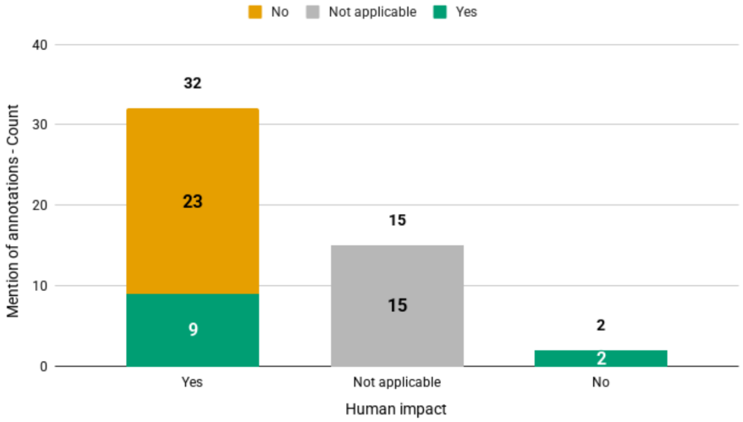


Figure 6: Pairwise analysis between mentions of impact on humans and mentions of ground truth in the “Broader Impact” sections.

Lastly, 35% of the venue papers make use of datasets that are not public. Consulting the comparison shown in Figure 7 reveals that most of these proprietary datasets are not the focal point of their respective papers, but rather introduced to support the research, either by training or evaluating a ML model.

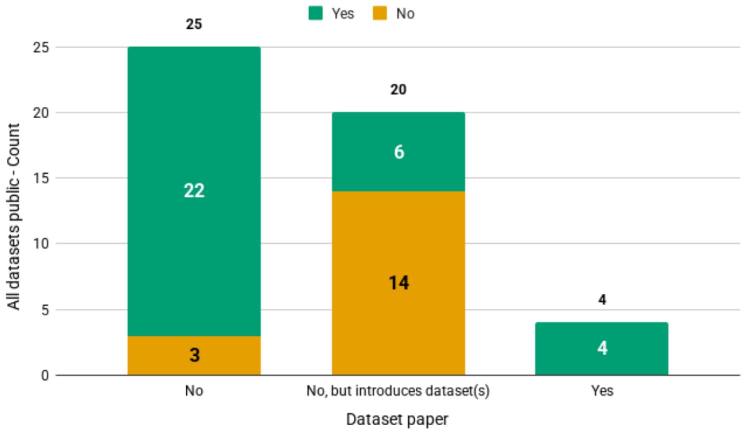


Figure 7: Pairwise analysis between papers that introduce datasets and use of public datasets.

4 Discussion

4.1 Lacking Information on the Annotation Process

As previously addressed, most dataset papers examined do provide general information, such as the item population and the rationale behind its selection. Unfortunately, similar levels of coverage do not extend to annotation. High-level aspects, such as the presence of human annotators and their source, are commonly reported. However, elaborations on the procedural aspects of annotation are overall scarce.

Firstly, the lack of interactive training procedures suggests limited engagement between dataset creators and annotators. Authors prefer formal written instructions as an alternative, yet this is an impersonal and formulaic approach by nature. The overall tendency towards excluding collaborative mechanisms, also suggested by lack of information on other criteria, implies a generally unsupervised annotation process. A possible pitfall of such an approach is that it encourages simplistic choices, which can reduce ground truth quality by making it devoid of nuance [4]. This is further supported by the general lack of interaction among annotators, who are assigned tasks individually without opportunities for discussion. As a result, annotation becomes an encapsulated activity, which only enables flawed datasets [4].

Additionally, IRR metrics are also largely unreported. Measurements of IRR show how consistently different annotators label the same data, which in turn reveal how consistent the labelling task actually is. Without them, one cannot assess whether the labels are reliable, or the result of agreement by chance between annotators. Most dataset papers do not include any IRR statistics or mention of overlap between annotators. This aspect, or lack thereof, raises doubts on the quality of the labels.

Within NeurIPS, there have been emerging efforts to make annotation and dataset construction more transparent, while also involving authors more directly in the process. Particularly, the OpenAssistant Conversations project adopts an unconventional framework. Over 13,000 volunteers contributed to the dataset, combining the actual labelling task with item quality reviews and filtering harmful content [13]. Discussion among labellers was facilitated through a Discord server, where authors, among other contributors, had a clear moderation role and promptly addressed issues raised by the annotators. The project does suffer from a set of limitations, particularly under-representation of some demographics. Nevertheless, it explores a labelling procedure that is community-driven, democratised, and aligned with the "Crowd Truth" framework [4], which makes it unique in the list of datasets that were analysed. Such initiatives additionally show that similar efforts are possible with limited funding, as all labellers were volunteers.

4.2 "What are these automated systems actually trained on?"

The field of computer vision has seen major advancements due to the use of large-scale, web-scraped datasets. Among them, ImageNet [14] is regarded as the benchmark that launched the widespread adoption of convolutional neural networks (CNNs) in both industry and academia [12]. CNNs have since been employed in applications ranging from medical imaging to product recommendations.

A significant amount of datasets included in this analysis share similarities with ImageNet, either through the data collection procedure, share of common items, or aim to provide all-encompassing image corpora. However, the scale of these datasets effectively implies that they are difficult to audit exhaustively.

Nevertheless, efforts have been made in discovering the intricacies of such datasets [15, 16]. Both studies have found vast quantities of harmful content in the datasets they inspected (i.e. ImageNet and LAION-400M), including explicit imagery, racial and ethnic discrimination, and propagation of other toxic stereotypes. At large, the corpora that dominate the field are inherently unfiltered and poorly reported, as further confirmed by this study.

Lack of information on this particular category of corpora has been recognised as a limitation within NeurIPS, due to the risks of harm, unintended spread of biases, and concerns regarding consent [17, 18, 19, 20]. Although the technological progress cannot be doubted, there is a definite need for more awareness and action to ensure that these practices do not persist and do not negatively affect human lives.

4.3 Private Datasets

Both the dataset annotation analysis and the targeted NeurIPS analysis have shown that 28% of the employed datasets that are not made public. Barriers to dataset access directly undermine the reproducibility of the presented findings, as researchers do not have the resources to verify results or inspect data quality.

As shown in the analysis of NeurIPS papers, the datasets that are not published are typically supporting resources for training or evaluating the proposed models. While the core contributions may be valid, the use of inaccessible datasets make the conditions under which results were produced uncertain. Moreover, NeurIPS has published papers that were authored by strong participants within the industry of ML, whose products are used by numerous people around the world. OpenAI is a particular example, having more than half a billion monthly users [21]. Upon inspecting the datasets used to create one of OpenAI's most acclaimed products, i.e. GPT-3 [22], one can observe that 9 out of the 36 employed datasets, from both training and evaluation stages, are proprietary. Hence, datasets that cannot be evaluated have shaped products used by people around the world. This draws attention to the transparency standards that research venues are expected to uphold, and whether current practices adequately support reproducibility and robust reviewing processes.

5 Limitations

5.1 Time Constraints

As this research project was conducted under a strict nine-week time frame, several methodological choices were driven by this aspect. For instance, the decision to focus solely on one research venue was taken with the objective of presenting a robust, in-depth analysis while also taking time into consideration. The remaining members of the peer group pursued similar reviews on other venues.

Perhaps the most significant decision that was influenced by time constraints was the number of papers and datasets selected for inspection. With this in mind, as an effort to balance feasibility and depth of analysis, the study targets 75 venue papers and 60 datasets. Analysing all the papers and compiling the relevant information took approximately five weeks. To justify the inclusion of specific items, we established selection criteria that prioritised relevance. Namely, citation count for the papers, and the weighted citation-based score for the datasets, discussed in subsection 2.4.

At last, the choice of examining how impact is discussed in NeurIPS research was limited to the last five years, due to the same time considerations. Moreover, it was motivated by the conference decision to mandate such a discussion in 2020 and include it in the author instructions one year later [8]. Approximately one week of the project was spent on this analysis. While a broader temporal scope might have yielded more insights, focusing on recent publications allowed for a targeted analysis within the available time frame.

Restricting the amount of items that were examined implicitly required sacrificing the breadth of the study. However, it allowed for a more robust analysis given the allocated time. Future research can build upon it by reviewing annotation practices on a larger scale.

5.2 Choice of Impact Indicators

Given the scope of the project, as well as the aim to capture the annotation practices across datasets that are widely used in NeurIPS, the chosen definition of societal impact has proven itself useful. Due to it being strictly quantitative, it provided a means to prioritise the venue papers and the datasets that constitute the subject of analysis.

Nevertheless, quantitative measures do have a few disadvantages. For instance, a drawback of the h5-index is that it has a bias towards venues that publish a significant number of papers, since a small conference or journal that publishes few papers will implicitly have a low index, even if those papers are highly cited. Thus, the decision to focus on prolific venues within the field was deliberate.

More complete definitions of societal impact would require qualitative assessments. Generally, societal impact can be understood as a perceivable effect on human lives. To this extent, NeurIPS encompasses a wide range of ML fields, many of which involve more technical contributions whose societal implications may not be immediately visible. Throughout the analysis, it became clear that many areas of Machine Learning, such as advances in computer vision, prediction systems, large language models, and speech recognition technologies, hold substantial impact, even if not directly observable in everyday life.

As such, while qualitative measures could offer a richer and more nuanced understanding of societal impact, the use of a quantitative metric in this project allowed for a scalable way to prioritize venue papers and datasets, which was in line with the study's focus on annotation practices. Future research tailored to specific application domains may build on this work and develop more context-dependent definitions of impact that also incorporate qualitative dimensions.

6 Conclusions and Future Work

This study was set up as a means to investigate the common reporting and collection practices concerning annotation, in the case of datasets frequently employed in NeurIPS research. In all, 71 venue papers and 60 datasets constituted the subject of this analysis. Subsequently, datasets were examined on a set of 27 criteria concerning the items, annotators, and annotation schema.

The findings reveal that, while general information is well reported, the procedural aspects of annotation are insufficiently documented. Omitting these stages of the labelling process limits the ability to assess the quality and reliability of the datasets in question. Moreover, the targeted review of recent NeurIPS papers indicated that, although societal impact is frequently approached, ground truth construction is rarely discussed in this context. This reveals that, while labelled datasets are a foundational part of ML applications, researchers do not take them into consideration when assessing the implications of their studies. The analysis also raises concerns about the accessibility of datasets, a notable proportion of which are private or only partially accessible. This practice is present even in recent venue research, which undermines ongoing efforts to ensure that research is reproducible.

As a concluding remark, it is shown that both the annotation process itself and its reporting are still largely unstandardised. Future work should aim to establish more rigorous and uniform procedures for documenting dataset annotation. Additionally, the research community in the field should intensify efforts to audit and critically examine widely used datasets, both in terms of labelling procedure and item population. At last, future research can extend upon this study, either by focusing it on specific branches of ML, or by broadening its scope to include more venues and different impact assessments.

Responsible Research

Reproducibility and Replicability

This literature review mainly aims to address issues of transparency in ML research. As such, reproducibility and replicability were, intuitively, focal points in conducting this study. Documenting the process was done extensively throughout the project, with the precise aim to present as detailed a description as possible.

Given the thorough documentation of the methodology, as well as the availability of all the resources that were used, it is reasonable to assume that this project is replicable. Researchers who wish to either extend upon this analysis, or apply the same methods for other venues or time frames, have all the resources at hand to conduct such a study. Paper and dataset extraction, along with dataset analysis and aggregate statistics, have well-set rules that can be followed in the future.

However, such an assumption cannot be applied to conclude that this review is fully reproducible. Venue papers were extracted from Scopus on a certain date, having citation count as the sorting criterion. As time passes, the number of citations will undoubtedly differ, which might alter the final set of papers to be reviewed. Citation count was additionally used to decide which datasets will be examined, hence there is no guarantee that this list will be the same in the future. Still, citations were used as the primary measurement of impact, so excluding them from the methodology would remove a fundamental component of this study.

On this basis, while replicability is guaranteed, there are parts of this study that need to be altered to also ensure reproducibility. Measuring impact differently, both in terms of papers and datasets, can mitigate this issue in future research on the topic.

Use of Large Language Models

The only Large Language Model (LLM) that was used during this project is GPT-4o, made by OpenAI. GPT-4o has been used to assist in writing the code that produced the statistics shown in section 3.

No LLM was used for the personal contribution to the codebase that generated the statistics shown in the project. Further, LLMs did not substitute any personal work or analysis in other stages of the analysis. Examples of exact prompts given to GPT-4o are shown in Appendix E.

Ethical Concerns

This thesis does not introduce a novel ML model or release any dataset intended for training or evaluation. As such, ethical concerns related to human use or impact – which are rightfully considered when deploying ML systems – do not directly apply here. Instead, this study mainly aims to contribute to an ongoing discussion around annotation practices, and how these are insufficiently addressed within the research community.

References

[1] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, “The values encoded in machine learning research,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, (New York, NY, USA), pp. 173–184, Association for Computing Machinery, 2022.

[2] R. S. Geiger, D. Cope, J. Ip, M. Lotosh, A. Shah, J. Weng, and R. Tang, ““garbage in, garbage out” revisited: What do machine learning application papers report about human-labeled training data?,” *Quantitative Science Studies*, vol. 2, pp. 795–827, 11 2021.

[3] J. Hullman, S. Kapoor, P. Nanayakkara, A. Gelman, and A. Narayanan, “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, (New York, NY, USA), pp. 335–348, Association for Computing Machinery, 2022.

[4] L. Aroyo and C. Welty, “Truth is a lie: Crowd truth and the seven myths of human annotation,” *AI Mag.*, vol. 36, pp. 15–24, Mar. 2015.

[5] Google, “Top publications – google scholar metrics.” `https://scholar.google.com/citations?view_op=top_venues`, 2025. Accessed: June 18, 2025.

[6] NeurIPS Proceedings, “Neurips 2025 call for papers.” `https://neurips.cc/Conferences/2025/CallForPapers`, 2025. Accessed: June 18, 2025.

[7] Wake Forest University School of Medicine, “Journal-level metrics.” `https://libguides.wakehealth.edu/researchmetrics/journal`, 2024. Accessed: 2025-06-22.

[8] C. Ashurst, E. Hine, P. Sedille, and A. Carlier, “Ai ethics statements: Analysis and lessons learnt from neurips broader impact statements,” in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, (New York, NY, USA), pp. 2047–2056, Association for Computing Machinery, 2022.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in *Computer Vision – ECCV 2014* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), (Cham), pp. 740–755, Springer International Publishing, 2014.

[10] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, “Training verifiers to solve math word problems,” 2021.

[11] A. Krizhevsky, “Learning multiple layers of features from tiny images,” *University of Toronto*, 05 2012.

[12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “Imagenet large scale visual recognition challenge,” 2015.

[13] A. Köpf, Y. Kilcher, D. von Rütte, S. Anagnostidis, Z. R. Tam, K. Stevens, A. Barhoum, D. Nguyen, O. Stanley, R. Nagyfi, S. ES, S. Suri, D. Glushkov, A. Dantuluri, A. Maguire, C. Schuhmann, H. Nguyen, and A. Mattick, "Openassistant conversations - democratizing large language model alignment," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 47669–47681, Curran Associates, Inc., 2023.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[15] A. Birhane and V. Prabhu, "Large image datasets: A pyrrhic win for computer vision?," pp. 1536–1546, 01 2021.

[16] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," 2021.

[17] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 9694–9705, Curran Associates, Inc., 2021.

[18] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 22199–22213, Curran Associates, Inc., 2022.

[19] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 36479–36494, Curran Associates, Inc., 2022.

[20] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 12077–12090, Curran Associates, Inc., 2021.

[21] M. Paris, "Chatgpt hits 1 billion users, openai ceo says, doubled in weeks," Apr. 2025. Accessed: 2025-06-22.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 1877–1901, Curran Associates, Inc., 2020.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[25] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds.), vol. 32, Curran Associates, Inc., 2019.

[28] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[29] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[30] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[31] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[32] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[33] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[34] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[36] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, “Improved training of wasserstein gans,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[37] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Advances in Neural Information Processing Systems* (C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, eds.), vol. 26, Curran Associates, Inc., 2013.

[38] X. SHI, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[39] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[40] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[41] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?,” in *Advances in Neural Information Processing Systems* (Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, eds.), vol. 27, Curran Associates, Inc., 2014.

[42] J. Snell, K. Swersky, and R. Zemel, “Prototypical networks for few-shot learning,” in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[43] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[44] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[45] O. Vinyals, C. Blundell, T. Lillicrap, k. kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[46] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in Neural Information Processing Systems* (F. Pereira, C. Burges, L. Bottou, and K. Weinberger, eds.), vol. 25, Curran Associates, Inc., 2012.

[47] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 27730–27744, Curran Associates, Inc., 2022.

[48] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 24824–24837, Curran Associates, Inc., 2022.

[49] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 8780–8794, Curran Associates, Inc., 2021.

[50] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, B. Piot, k. kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent - a new approach to self-supervised learning," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 21271–21284, Curran Associates, Inc., 2020.

[51] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 12449–12460, Curran Associates, Inc., 2020.

[52] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information*

Processing Systems (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 18661–18673, Curran Associates, Inc., 2020.

[53] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 9459–9474, Curran Associates, Inc., 2020.

[54] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 9912–9924, Curran Associates, Inc., 2020.

[55] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 596–608, Curran Associates, Inc., 2020.

[56] I. O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 24261–24272, Curran Associates, Inc., 2021.

[57] Y. You, T. Chen, Y. Sui, T. Chen, Z. Wang, and Y. Shen, "Graph contrastive learning with augmentations," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 5812–5823, Curran Associates, Inc., 2020.

[58] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 22419–22430, Curran Associates, Inc., 2021.

[59] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 7462–7473, Curran Associates, Inc., 2020.

[60] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. L. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. a. Bińkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems* (S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, eds.), vol. 35, pp. 23716–23736, Curran Associates, Inc., 2022.

[61] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, "Unsupervised data augmentation for consistency training," in *Advances in Neural Information Processing Systems*

(H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6256–6268, Curran Associates, Inc., 2020.

[62] M. Tancik, P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 7537–7547, Curran Associates, Inc., 2020.

[63] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 17022–17033, Curran Associates, Inc., 2020.

[64] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 22118–22133, Curran Associates, Inc., 2020.

[65] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 34892–34916, Curran Associates, Inc., 2023.

[66] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 53728–53741, Curran Associates, Inc., 2023.

[67] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 46595–46623, Curran Associates, Inc., 2023.

[68] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 10088–10115, Curran Associates, Inc., 2023.

[69] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 11809–11822, Curran Associates, Inc., 2023.

[70] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: language agents with verbal reinforcement learning," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 8634–8652, Curran Associates, Inc., 2023.

[71] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark, "Self-refine: Iterative refinement with self-feedback," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 46534–46594, Curran Associates, Inc., 2023.

[72] W. Dai, J. Li, D. LI, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 49250–49267, Curran Associates, Inc., 2023.

[73] T. Schick, J. Dwivedi-Yu, R. Dessi, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 68539–68551, Curran Associates, Inc., 2023.

[74] C. Wang, W. He, Y. Nie, J. Guo, C. Liu, Y. Wang, and K. Han, "Gold-yolo: Efficient object detector via gather-and-distribute mechanism," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 51094–51112, Curran Associates, Inc., 2023.

[75] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 28541–28564, Curran Associates, Inc., 2023.

[76] Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 38154–38180, Curran Associates, Inc., 2023.

[77] Z. Wang, C. Lu, Y. Wang, F. Bao, C. LI, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 8406–8441, Curran Associates, Inc., 2023.

[78] X. Ma, G. Fang, and X. Wang, "Llm-pruner: On the structural pruning of large language models," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 21702–21720, Curran Associates, Inc., 2023.

[79] X. Zou, J. Yang, H. Zhang, F. Li, L. Li, J. Wang, L. Wang, J. Gao, and Y. J. Lee, "Segment everything everywhere all at once," in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 19769–19782, Curran Associates, Inc., 2023.

[80] A. Wei, N. Haghtalab, and J. Steinhardt, “Jailbroken: How does llm safety training fail?,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 80079–80110, Curran Associates, Inc., 2023.

[81] L. Ke, M. Ye, M. Danelljan, Y. liu, Y.-W. Tai, C.-K. Tang, and F. Yu, “Segment anything in high quality,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 29914–29934, Curran Associates, Inc., 2023.

[82] J. Liu, C. S. Xia, Y. Wang, and L. ZHANG, “Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 21558–21572, Curran Associates, Inc., 2023.

[83] T. Zhou, P. Niu, x. wang, L. Sun, and R. Jin, “One fits all: Power general time series analysis by pretrained lm,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 43322–43355, Curran Associates, Inc., 2023.

[84] N. Gruver, M. Finzi, S. Qiu, and A. G. Wilson, “Large language models are zero-shot time series forecasters,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 19622–19635, Curran Associates, Inc., 2023.

[85] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Defossez, “Simple and controllable music generation,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 47704–47720, Curran Associates, Inc., 2023.

[86] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, H. He, N. An, D. Lian, L. Cao, and Z. Niu, “Frequency-domain mlps are more effective learners in time series forecasting,” in *Advances in Neural Information Processing Systems* (A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, eds.), vol. 36, pp. 76656–76679, Curran Associates, Inc., 2023.

[87] O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, J. Leveling, C. Monz, P. Pecina, M. Post, H. Saint-Amand, R. Soricut, L. Specia, and A. Tamchyna, “Findings of the 2014 workshop on statistical machine translation,” in *Proceedings of the Ninth Workshop on Statistical Machine Translation* (O. Bojar, C. Buck, C. Federmann, B. Haddow, P. Koehn, C. Monz, M. Post, and L. Specia, eds.), (Baltimore, Maryland, USA), pp. 12–58, Association for Computational Linguistics, June 2014.

[88] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[89] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz, “Building a large annotated corpus of English: The Penn Treebank,” *Computational Linguistics*, vol. 19, no. 2, pp. 313–330, 1993.

[90] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014.

[91] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. E. Hinton, "Grammar as a foreign language," *CoRR*, vol. abs/1412.7449, 2014.

[92] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[93] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *CoRR*, vol. abs/1902.06162, 2019.

[94] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.

[95] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao, "LSUN: construction of a large-scale image dataset using deep learning with humans in the loop," *CoRR*, vol. abs/1506.03365, 2015.

[96] A. Patel, S. Bhattamishra, and N. Goyal, "Are nlp models really able to solve simple math word problems?," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Association for Computational Linguistics, June 2021.

[97] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela, "Adversarial NLI: A new benchmark for natural language understanding," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 4885–4901, Association for Computational Linguistics, July 2020.

[98] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the AI2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018.

[99] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the AI2 reasoning challenge," *CoRR*, vol. abs/1803.05457, 2018.

[100] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *CoRR*, vol. abs/2005.14165, 2020.

[101] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Technical Report 0, University of Toronto, Toronto, Ontario, 2009.

[102] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *2008 Sixth Indian Conference on Computer Vision, Graphics Image Processing*, pp. 722–729, 2008.

[103] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. V. Jawahar, "Cats and dogs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3498–3505, 2012.

[104] T. Berg, J. Liu, S. W. Lee, M. L. Alexander, D. W. Jacobs, and P. N. Belhumeur, "Birdsnap: Large-scale fine-grained visual categorization of birds," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2019–2026, 2014.

[105] F.-F. Li, M. Andreeto, M. Ranzato, and P. Perona, "Caltech 101," Apr 2022.

[106] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing textures in the wild," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

[107] S. Maji, E. Rahtu, J. Kannala, M. B. Blaschko, and A. Vedaldi, "Fine-grained visual classification of aircraft," *CoRR*, vol. abs/1306.5151, 2013.

[108] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *Proceedings of the 13th European Conference on Computer Vision (ECCV 2014), Part VI* (D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds.), vol. 8694 of *Lecture Notes in Computer Science*, (Zurich, Switzerland), pp. 446–461, Springer, Cham, Sept. 2014.

[109] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, "3d object representations for fine-grained categorization," in *2013 IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013.

[110] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

[111] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom, "Program induction by rationale generation: Learning to solve and explain algebraic word problems," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (R. Barzilay and M.-Y. Kan, eds.), (Vancouver, Canada), pp. 158–167, Association for Computational Linguistics, July 2017.

[112] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," *CoRR*, vol. abs/2201.11903, 2022.

[113] A. Talmor, J. Herzig, N. Lourie, and J. Berant, "CommonsenseQA: A question answering challenge targeting commonsense knowledge," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4149–4158, Association for Computational Linguistics, June 2019.

[114] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, *et al.*, "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *arXiv preprint*, 2022.

[115] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," 2022.

[116] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[117] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," 2022.

[118] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[119] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

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

[121] Stanford CRFM, "Overview." `https://crfm.stanford.edu/2023/03/13/alpaca.html`, Mar 2023. Accessed: 2025-06-22.

[122] R. Nallapati, B. Zhou, C. dos Santos, Ç. Gu̇lçehre, and B. Xiang, "Abstractive text summarization using sequence-to-sequence RNNs and beyond," in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (S. Riezler and Y. Goldberg, eds.), (Berlin, Germany), pp. 280–290, Association for Computational Linguistics, Aug. 2016.

[123] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Association for Computational Linguistics, 2011.

[124] N. Stiennon, L. Ouyang, J. Wu, D. M. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano, "Learning to summarize from human feedback," *CoRR*, vol. abs/2009.01325, 2020.

[125] M. Völske, M. Potthast, S. Syed, and B. Stein, "TL;DR: Mining Reddit to learn automatic summarization," in *Proceedings of the Workshop on New Frontiers in Summarization* (L. Wang, J. C. K. Cheung, G. Carenini, and F. Liu, eds.), (Copenhagen, Denmark), pp. 59–63, Association for Computational Linguistics, Sept. 2017.

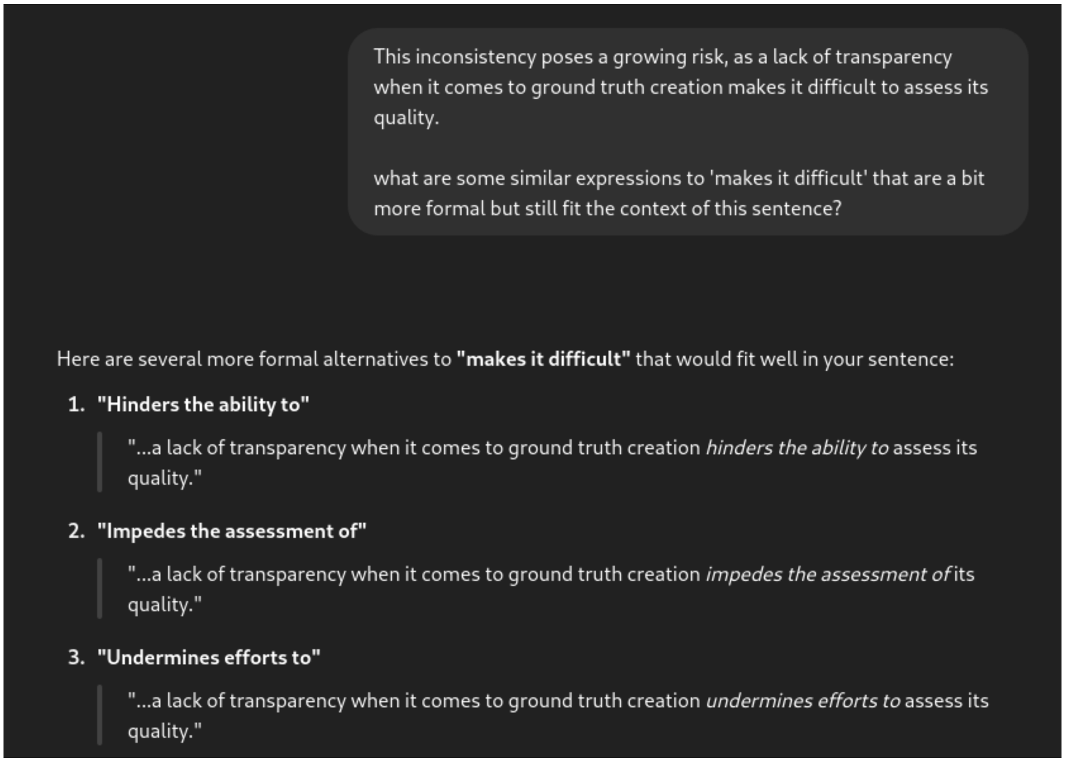
[126] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023.

[127] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, "Judging llm-as-a-judge with mt-bench and chatbot arena," 2023. NeurIPS 2023 Datasets and Benchmarks Track.

[128] LAION-AI, "Open-instruction-generalist: Open instruction generalist is an assistant trained on massive synthetic instructions to perform many millions of tasks." GitHub repository. Accessed: 2025-06-22.

[129] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-pairs: A challenge dataset for measuring social biases in masked language models," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (B. Webber, T. Cohn, Y. He, and Y. Liu, eds.), (Online), pp. 1953–1967, Association for Computational Linguistics, Nov. 2020.

[130] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei, and A. Roberts, "The flan collection: Designing data and methods for effective instruction tuning," 2023.



This inconsistency poses a growing risk, as a lack of transparency when it comes to ground truth creation makes it difficult to assess its quality.

what are some similar expressions to 'makes it difficult' that are a bit more formal but still fit the context of this sentence?

Here are several more formal alternatives to **"makes it difficult"** that would fit well in your sentence:

1. **"Hinders the ability to"**

   > "...a lack of transparency when it comes to ground truth creation *hinders the ability to* assess its quality."

2. **"Impedes the assessment of"**

   > "...a lack of transparency when it comes to ground truth creation *impedes the assessment of* its quality."

3. **"Undermines efforts to"**

   > "...a lack of transparency when it comes to ground truth creation *undermines efforts to* assess its quality."

Figure 8: Enter Caption

A Resources

The collaborative spreadsheet, containing all the venue papers and datasets that have been inspected, can be accessed here: `https://docs.google.com/spreadsheets/d/16MkuS-upEQxkAj-poZO5ggPqmu_UIDbwi7HWS3-21HE/edit?usp=sharing`.

Here, *Tab 1* was compiled as mentioned in subsection 2.2, *Tab 2* as described in subsection 2.3 and *Tab 3* as discussed in subsection 2.4. The code that was used for aggregate statistics is available on GitHub: `https://github.com/Gargant0373/DatasetAnalysis`.

B Methodology Details

This appendix is meant to provide more details regarding paper extraction and the criteria used in analysing datasets and NeurIPS papers.

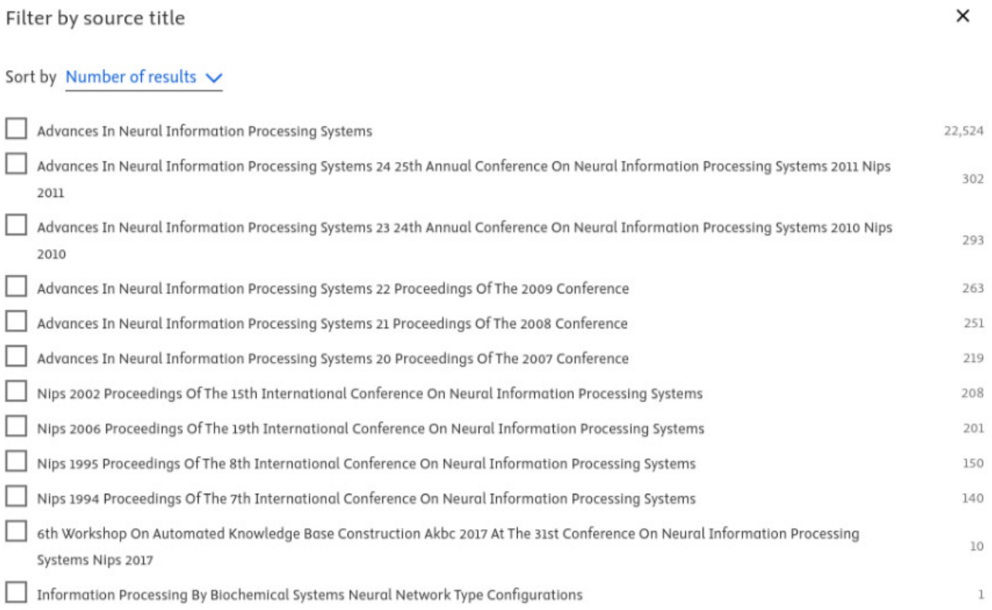
B.1 Using Scopus to retrieve NeurIPS papers

Paper extraction took place on April 25, 2025; as such, all the query results were recorded that day. The first Scopus query was used to observe the venues (in Scopus, "source title"s) that appear after simply querying Neural Information Processing Systems:

```
SRCTITLE ( neural AND information AND processing AND systems )
```

The query mostly yielded NeurIPS papers, except for one venue: "Information Processing by Chemical Systems Neural Network Type Configurations" as seen in Figure 9. As such, the query was tweaked to exclude this venue:

```
SRCTITLE ( neural AND information AND processing AND systems AND NOT biochemical )
```



Filter by source title ×

Sort by Number of results

□	Advances In Neural Information Processing Systems	22,524
□	Advances In Neural Information Processing Systems 24 25th Annual Conference On Neural Information Processing Systems 2011 Nips 2011	302
□	Advances In Neural Information Processing Systems 23 24th Annual Conference On Neural Information Processing Systems 2010 Nips 2010	293
□	Advances In Neural Information Processing Systems 22 Proceedings Of The 2009 Conference	263
□	Advances In Neural Information Processing Systems 21 Proceedings Of The 2008 Conference	251
□	Advances In Neural Information Processing Systems 20 Proceedings Of The 2007 Conference	219
□	Nips 2002 Proceedings Of The 15th International Conference On Neural Information Processing Systems	208
□	Nips 2006 Proceedings Of The 19th International Conference On Neural Information Processing Systems	201
□	Nips 1995 Proceedings Of The 8th International Conference On Neural Information Processing Systems	150
□	Nips 1994 Proceedings Of The 7th International Conference On Neural Information Processing Systems	140
□	6th Workshop On Automated Knowledge Base Construction Akbc 2017 At The 31st Conference On Neural Information Processing Systems Nips 2017	10
□	Information Processing By Biochemical Systems Neural Network Type Configurations	1

Figure 9: List of source titles shown in Scopus after applying the first query.

This new query only yielded NeurIPS papers. Having sorted this out, to extract papers from the certain time periods, the following queries were used:

- 15-year period:

  ```
  SRCTITLE ( neural AND information AND processing AND systems
      AND NOT biochemical )
  AND PUBYEAR > 2009 AND PUBYEAR < 2025
  ```

- 5-year period:

```
SRCTITLE ( neural AND information AND processing AND systems
    AND NOT biochemical )
AND PUBYEAR > 2019 AND PUBYEAR < 2025
```

- 2-year period:

```
SRCTITLE ( neural AND information AND processing AND systems
    AND NOT biochemical )
AND PUBYEAR > 2022 AND PUBYEAR < 2025
```

The "Sort by" feature of Scopus was used to get the most cited papers from each period, as shown in Figure 10.

21,155 documents found

Sort by Cited by (highest)

Analyze results
□ All
Export
Download
Citation overview
••• More
Show all abstracts

Figure 10: Sorting by citation count in Scopus.

B.2 Tab 3: Dataset assessment criteria

All the criteria that were used to inspect the annotation process of the datasets are discussed below, in the same order they occurred in the collaborative spreadsheet. As mentioned before, some of them were taken from the analysis of Geiger et. al [2], while others are new criteria. Here, the new criteria will be listed in **bold**.

Some general rules of thumb were taken into account during this process. Each cell in the sheet could be completed with "No information", "Not applicable", or "Unsure", depending on the context.

"No information" means that the dataset paper does not mention or discuss the criterion at all. With this in mind, "No" is different from "No information", as "No" means that the dataset paper explicitly states that something was not used or not considered (e.g. it is stated that there was no sample size that was aimed for).

"Not applicable" was used for criteria that do not fit the context of the specific dataset (e.g. there is no need for labeller selection criteria if the dataset was machine-labelled).

Finally, "Unsure" was used when there was uncertainty in the information provided by the dataset papers, which would be further discussed in a meeting with the peer group. If the uncertainty could not be resolved, the cell would remain completed as "Unsure".

With this in mind, these are the criteria that were considered when reading the dataset papers:

- **Available**. Is there a reliable source for information about the dataset? Dropdown criterion, with the following options:

  - *Yes*, if there is a reliable source of information on the dataset, i.e. a dataset paper, or a comprehensive README file in a repository that contains the dataset;
  - *No*, otherwise.

- Outcome. The ML task the dataset is aimed for.

- Human Labels. Was the dataset labelled by humans? Dropdown criterion, with the following options:

  - *Yes for all*, if humans labelled all the items of the dataset;
  - *Yes for some*, if human labellers were present, but not for all items;
  - *No / machine labelled*, if the dataset was either unlabelled, or entirely machine-labelled;
  - *Implicit Yes*, if this information is not explicitly stated, but the dataset paper mentions other things, such that we can infer human labellers were present (e.g. how many annotators labelled an item, overlap, etc.);
  - *No information.*
  - *Unsure*;

  "Not applicable" was not used here, since this list of options is exhaustive for this criterion.

- OG Labels. Were the labels original (i.e. created by people who are directly involved in the creation of the dataset)? Dropdown criterion, with the following options:

  - *OG*, if the labels themselves were created either by the authors, or by the labellers;
  - *Mix OG, External*, if there are both labels created by authors/labellers and labels that are taken from an external source that is already available;
  - *External*, if all the labels were taken from an external source;
  - *Not Labelled*, if the dataset was not labelled at all. This option is a replacement for "Not applicable";
  - *No information.*
  - *Unsure*;

- Label Source. Where were the annotators hired from? / Who are the annotators? If the dataset is machine-labelled, what algorithm / model was used for annotation?

- Prescreening. How were the labellers selected? Dropdown criterion, with the following options:

  - *Generic skill-based*, if annotators were selected based on a specific skill set (e.g. language proficiency for a translation dataset);
  - *No pre-screening (stated)*, if it is explicitly mentioned that no selection took place;
  - *Project Specific*, if the labellers are known by the authors, or if authors did their own pre-screening;
  - *Location Qualification*, if the labellers were selected based on their location;

  - *Previous Platform*, if labellers were chosen based on how well they performed for other jobs (e.g. HIT accuracy on MTurk);
  - *No information*;
  - *Not applicable.*
  - *Unsure*;

- Compensation. How were the annotators compensated? Dropdown criterion, with the following options:

  - *Money*, if the labellers were paid;
  - *Authorship*, if the labellers were mentioned as paper authors;
  - *Course Credit*, if the labellers were, for example, students, and the annotation task was part of coursework;
  - *Other compensation*, if other compensation method was stated;
  - *Volunteer*, if the labellers were not given any compensation;
  - *No information*;
  - *Not applicable*;
  - *Unsure.*

- Training. Did the labellers receive any sort of interactive training prior to the annotation process? Formal instructions do not count as training. Dropdown criterion, with the following options:

  - *Some training*, if interactive training took place;
  - *No information*;
  - *Not applicable*;
  - *Unsure.*

- Formal Instructions. Did the annotators receive any formal instructions prior to annotation? Dropdown criterion, with the following options:

  - *No instructions*, if it is explicitly stated that there were no instructions provided;
  - *Some Instructions*, if it is mentioned that labellers were given instructions, but this is not elaborated on;
  - *Formal Instructions*, if it is clearly stated that the annotators received formal instructions, with those instructions also stated in the paper;
  - *No information*;
  - *Not applicable*;
  - *Unsure.*

- **Labeller Population Rationale**. Do the authors of the dataset paper provide any reasoning as to why they chose those specific labellers?

- Total Labellers. How many labellers worked on annotation?

- **Annotators per item**. How many labellers worked on each item? This number can either be exact, or a measurement of central tendency, i.e. mean, mode, or median.
- **Label Threshold**. What is the minimum amount of labellers each item needed?
- Overlap. Was there overlap in the annotation process? In other words, did more labellers work on the same items? Dropdown criterion, with the following options:
  - *Yes for all*, if each item in the dataset was labelled by more than one annotator;
  - *Yes for some*, if some, but not all items were labelled by more than one annotator;
  - *No*, if there was no overlap;
  - *No information*;
  - *Not applicable*;
  - *Unsure*.
- Overlap Synthesis. How were disagreements solved between annotators, if it was the case? Dropdown criterion, with the following options:
  - *Quantitative*, if some quantitative criterion was used to solve overlap (e.g. majority vote);
  - *Qualitative*, if some qualitative criterion was used to solve overlap (e.g. discussion between annotators);
  - *Other*, if there were other methods to solve overlap;
  - *No information*;
  - *Not applicable*;
  - *Unsure*.
- **Synthesis Type**. What was the exact method used to solve disagreements?
- **Discussion**. Was there any discussion among the annotators? Dropdown criterion, with the following options:
  - *Yes*, if discussions took place;
  - *No*, if it is explicitly stated that no discussions took place;
  - *No information*;
  - *Not applicable*;
  - *Unsure*.
- Inter-Rater Reliability (IRR). Was there any IRR reported, if there was overlap? Dropdown criterion, with the following options:
  - *Yes*, if IRR was reported;
  - *No*, if no IRR was reported;
  - *No information*;
  - *Not applicable*;
  - *Unsure*.

- **Metric**. What was the exact metric for computing IRR?
- **Item population**. Briefly describe what the dataset contains.
- **Item population rationale**. Why was this specific item population chosen?
- **Item source**. Where were the items taken form?
- **A priori sample size**. Was the sample size chosen before the creation of the dataset?
- **Item sample size rationale**. Did the authors decide on the sample size for some specific reasons?
- **A priori annotation schema**. Was the annotation schema (i.e. the labels) decided before the creation of the dataset?
- **Annotation schema rationale**. Did the authors choose the annotation schema for any specific reasons?
- Link to data. Is the dataset publicly available online? Dropdown criterion, with the following options:

  - *Yes*, if there is a provided link to the dataset, and it directs to the dataset;
  - *Yes, but broken*, if there is a link provided, but it does not direct to the dataset;
  - *No*, if there is no link provided;
  - *Not applicable*;
  - *Unsure.*

  "No information" is not an option here, since the list of options is exhaustive for this criterion.

B.3 NeurIPS societal impact analysis

This study also inspects how NeurIPS researchers assess the societal impact of their work. Criteria used in the analysis are listed below, in the order they occurred in the spreadsheet. Once again, "Not applicable" is used if the criterion does not fit the context (e.g. there is no discussion in the "Broader Impact" section if there is no such section).

- **Dataset paper**. Is the research presented in the paper centred around releasing a dataset or benchmark? Drowdown criterion, with the following options:

  - *Yes*, if this is the case;
  - *No, but introduces datasets*, if the paper is not centred around a dataset / benchmark, but datasets were created as part of the research;
  - *No*, otherwise.

- **Societal / Broader Impact section**. Does the paper have a (sub-)section strictly named this way? *Yes* / *No* answers only.

- **Datasets mentioned**. Are datasets discussed in this section? *Yes / No / Not Applicable* answers only.
- **DS details**. Brief description of how datasets are discussed.
- **Dataset annotations/labels mentioned**. Are annotations discussed in this section? Discussions on item population or sample size also count. *Yes / No / Not Applicable* answers only.
- **Label details**. Brief description of how annotations are discussed.
- **Human impact**. Does the section address other aspects that concern impact on human lives? This includes, but is not limited to: bias and fairness, legality and safety, health, environment, impact on labour, etc. *Yes / No / Not Applicable* answers only.
- **Datasets / annotations in other sections**. Does the paper discuss the datasets used, and / or their annotation process? Solely mentioning the datasets counts towards a "No" answer. Dropdown criterion, with the following options:
  - *Yes*, if this is the case;
  - *Yes, but only contents*, if the contents of the datasets are described, but there is no other discussion;
  - *No*, otherwise.
- **Discussion details**. Brief description of what is discussed.
- **Internal dataset(s)**. Does the paper use proprietary datasets that are not made public? *Yes / No* answers only.

C Analysed Papers and Datasets

This appendix compiles the list of NeurIPS papers and dataset papers that were inspected as part of the project.

Table 3: NeurIPS papers that were analysed, without duplicates.
**According to Scopus on April 25, 2025.*

Title	**Year**	**Period**	**Cited by***
ImageNet classification with deep convolutional neural networks [23]	2012	15	85,220
Attention is all you need [24]	2017	15	83,482
Generative adversarial nets [25]	2014	15	48,297
Faster R-CNN: Towards real-time object detection with region proposal networks [26]	2015	15	33,127
PyTorch: An imperative style, high-performance deep learning library [27]	2019	15	29,810
Distributed representations of words and phrases and their compositionality [28]	2013	15	25,301

Table 3, continued from previous page

Title	Year	Period	Cited by*
Language models are few-shot learners [22]	2020	15	19,613
A unified approach to interpreting model predictions [29]	2017	15	16,200
Sequence to sequence learning with neural networks [30]	2014	15	15,234
Inductive representation learning on large graphs [31]	2017	15	11,673
Denoising diffusion probabilistic models [32]	2020	15	9,246
LightGBM: A highly efficient gradient boosting decision tree [33]	2017	15	8,998
GANs trained by a two time-scale update rule converge to a local Nash equilibrium [34]	2017	15	8,643
PointNet++: Deep hierarchical feature learning on point sets in a metric space [35]	2017	15	8,244
Improved training of wasserstein GANs [36]	2017	15	7,291
Translating embeddings for modeling multi-relational data [37]	2013	15	7,106
Convolutional LSTM network: A machine learning approach for precipitation nowcasting [38]	2015	15	6,900
Convolutional neural networks on graphs with fast localized spectral filtering [39]	2016	15	6,665
Two-stream convolutional networks for action recognition in videos [40]	2014	15	6,287
How transferable are features in deep neural networks? [41]	2014	15	6,227
Prototypical networks for few-shot learning [42]	2017	15	6,172
Improved techniques for training GANs [43]	2016	15	6,114
Spatial transformer networks [44]	2015	15	6,099
Matching networks for one shot learning [45]	2016	15	5,721
Practical Bayesian optimization of machine learning algorithms [46]	2012	15	5,701
Training language models to follow instructions with human feedback [47]	2022	5	4,795
Chain-of-Thought Prompting Elicits Reasoning in Large Language Models [48]	2022	5	4,314
Diffusion Models Beat GANs on Image Synthesis [49]	2021	5	4,045
Bootstrap your own latent: a new approach to self-supervised learning [50]	2020	5	3,942
SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers [20]	2021	5	3,896
wav2vec 2.0: A framework for self-supervised learning of speech representations [51]	2020	5	3,605
Supervised contrastive learning [52]	2020	5	3,228
Retrieval-augmented generation for knowledge-intensive NLP tasks [53]	2020	5	2,587
Unsupervised learning of visual features by contrasting cluster assignments [54]	2020	5	2,314

Table 3, continued from previous page

Title	Year	Period	Cited by*
FixMatch: Simplifying semi-supervised learning with consistency and confidence [55]	2020	5	2,275
Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding [19]	2022	5	2,207
MLP-Mixer: An all-MLP Architecture for Vision [56]	2021	5	1,730
Graph contrastive learning with augmentations [57]	2020	5	1,645
Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting [58]	2021	5	1,629
Large Language Models are Zero-Shot Reasoners [18]	2022	5	1,569
Implicit neural representations with periodic activation functions [59]	2020	5	1,511
Flamingo: a Visual Language Model for Few-Shot Learning [60]	2022	5	1,486
Unsupervised data augmentation for consistency training [61]	2020	5	1,384
Fourier features let networks learn high frequency functions in low dimensional domains [62]	2020	5	1,338
Align before Fuse: Vision and Language Representation Learning with Momentum Distillation [17]	2021	5	1,328
HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis [63]	2020	5	1,273
Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models	2024	5	1,256
Open graph benchmark: Datasets for machine learning on graphs [64]	2020	5	1,255
Visual Instruction Tuning [65]	2023	2	740
Direct Preference Optimization: Your Language Model is Secretly a Reward Model [66]	2023	2	531
Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena [67]	2023	2	516
QLORA: Efficient Finetuning of Quantized LLMs [68]	2023	2	504
Tree of Thoughts: Deliberate Problem Solving with Large Language Models [69]	2023	2	387
Reflexion: Language Agents with Verbal Reinforcement Learning [70]	2023	2	288
SELF-REFINE: Iterative Refinement with Self-Feedback [71]	2023	2	214
InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning [72]	2023	2	213
Toolformer: Language Models Can Teach Themselves to Use Tools [73]	2023	2	212
LIMA: Less Is More for Alignment	2023	2	175
Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism [74]	2023	2	149
LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day [75]	2023	2	142

Table 3, continued from previous page

Title	Year	Period	Cited by*
HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face [76]	2023	2	141
ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation [77]	2023	2	136
LLM-Pruner: On the Structural Pruning of Large Language Models [78]	2023	2	134
Segment Everything Everywhere All at Once [79]	2023	2	124
Jailbroken: How Does LLM Safety Training Fail? [80]	2023	2	115
Segment Anything in High Quality [81]	2023	2	114
Is Your Code Generated by ChatGPT Really Correct? Rigorous Evaluation of Large Language Models for Code Generation [82]	2023	2	113
One Fits All: Power General Time Series Analysis by Pretrained LM [83]	2023	2	99
Large Language Models Are Zero-Shot Time Series Forecasters [84]	2023	2	91
Simple and Controllable Music Generation [85]	2023	2	88
OpenAssistant Conversations - Democratizing Large Language Model Alignment [13]	2023	2	86
Frequency-domain MLPs are More Effective Learners in Time Series Forecasting [86]	2023	2	82

Table 4: Table

Dataset	**Period**
cifar-10 [11]	15
imagenet 2012 [12]	15
wmt14 [87]	15
mnist [88]	15
ptb [89]	15
imagenet 2010 [90]	15
imagenet fall 2009	15
berkeleyparser [91]	15
tfd	15
coco [9]	15
pascal voc 2007 [92]	15
pascal voc 2012 [93]	15
google news	15
word analogy task [94]	15
lsun [95]	15
svhn [96]	15
anli [97]	15
arc-challenge [98]	15
arc-easy [99]	15
books1 [100]	15
imagenet [14]	5
cifar-10 [11]	5
cifar-100 [101]	5
oxford flowers 102 [102]	5
oxford-iiit pets [103]	5
pascal voc 2007 [92]	5
imagenet 2012 [12]	5
birdsnap [104]	5
caltech101 [105]	5
dtd [106]	5
fgvc aircraft [107]	5
food-101 [108]	5
stanfordcars [109]	5
sun397 [110]	5
pascal voc 2012 [93]	5
aqua [111]	5
coin flip [112]	5
commonsenseqa [113]	5
date understanding [114]	5
gsm8k [10]	5
coco [9]	2
hh-rlhf [115]	2
llava-instruct-158k [116]	2
scienceqa [117]	2
cc-595k [118]	2

llava-bench [119]	2
superni [120]	2
alpaca [121]	2
gsm8k [10]	2
openassistant conversations [13]	2
cnn/dm modified [122]	2
imdb- [123]	2
n/a [124]	2
webis-tldr-17 [125]	2
chatbot arena [126]	2
mt-bench [127]	2
chip2 [128]	2
crows-pairs [129]	2
flan v2 [130]	2
llava-bench COCO[119]	2
llava-bench in-the-wild [119]	2

D Additional Figures

Other relevant statistics on the findings can be found here.

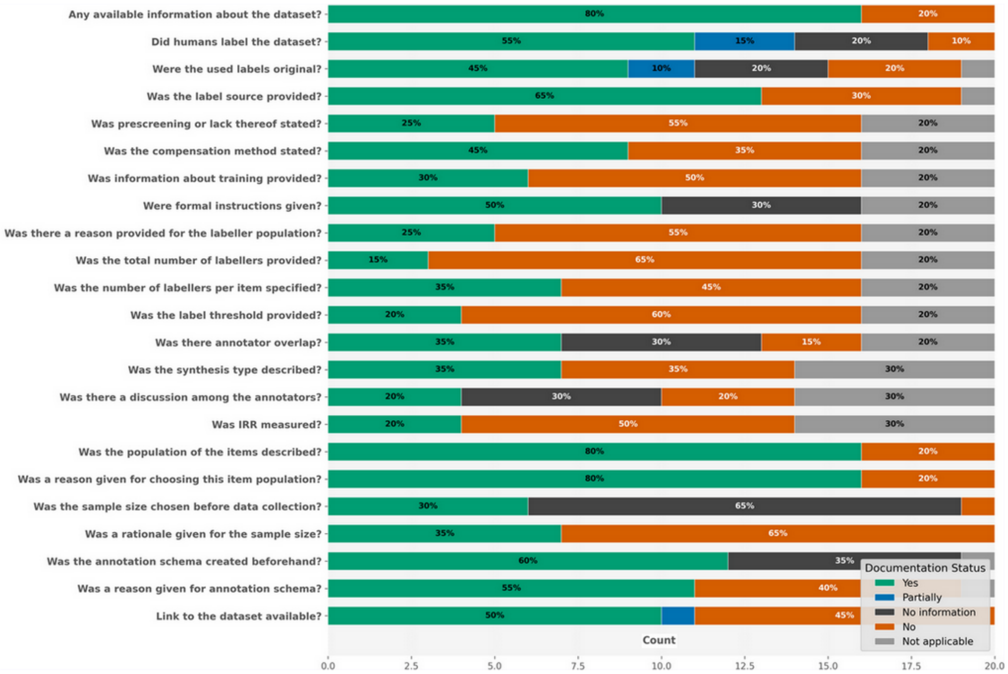


Figure 11: Overall statistics for the 15-year period.

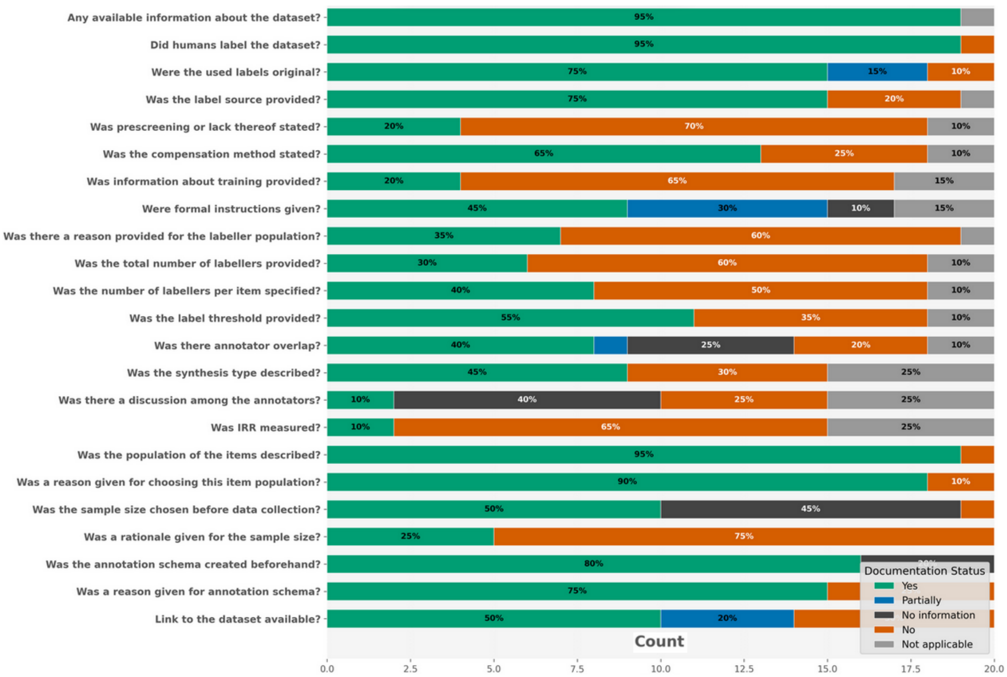


Figure 12: Overall statistics for the 5-year period.

E Examples of GPT-4o prompts

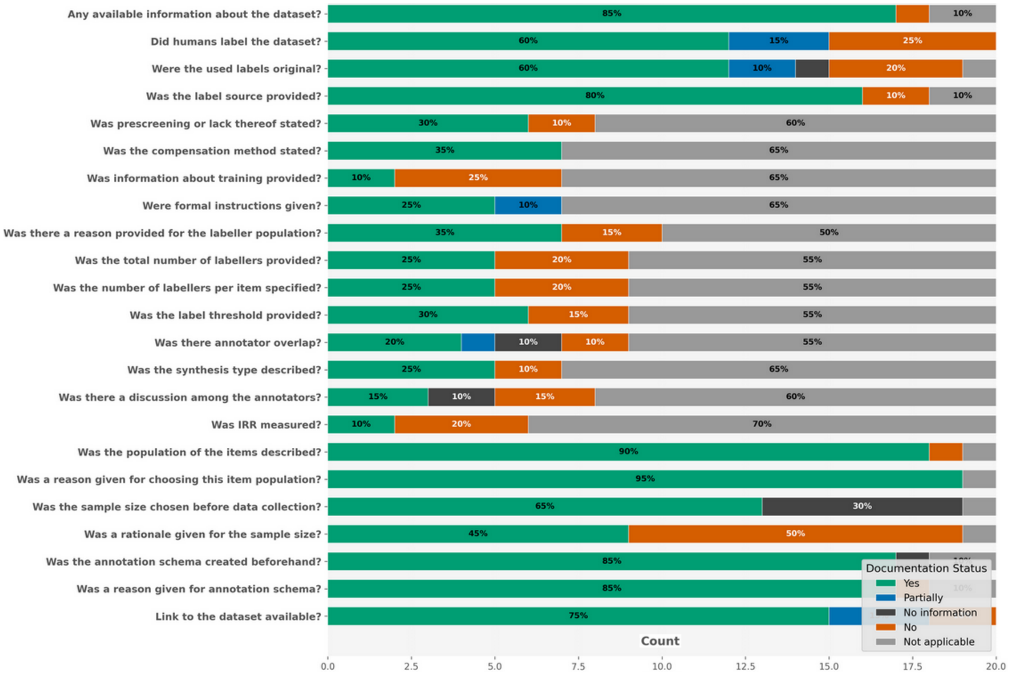
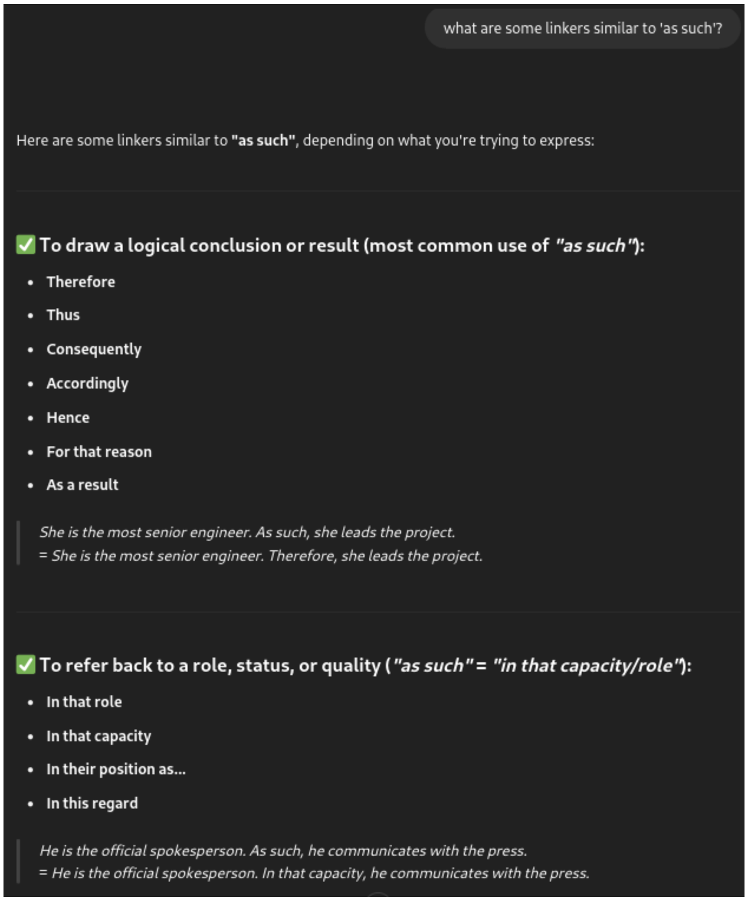


Figure 13: Overall statistics for the 2-year period.



what are some linkers similar to 'as such'?

Here are some linkers similar to **"as such"**, depending on what you're trying to express:

✅ To draw a logical conclusion or result (most common use of *"as such"*):

- **Therefore**
- **Thus**
- **Consequently**
- **Accordingly**
- **Hence**
- **For that reason**
- **As a result**

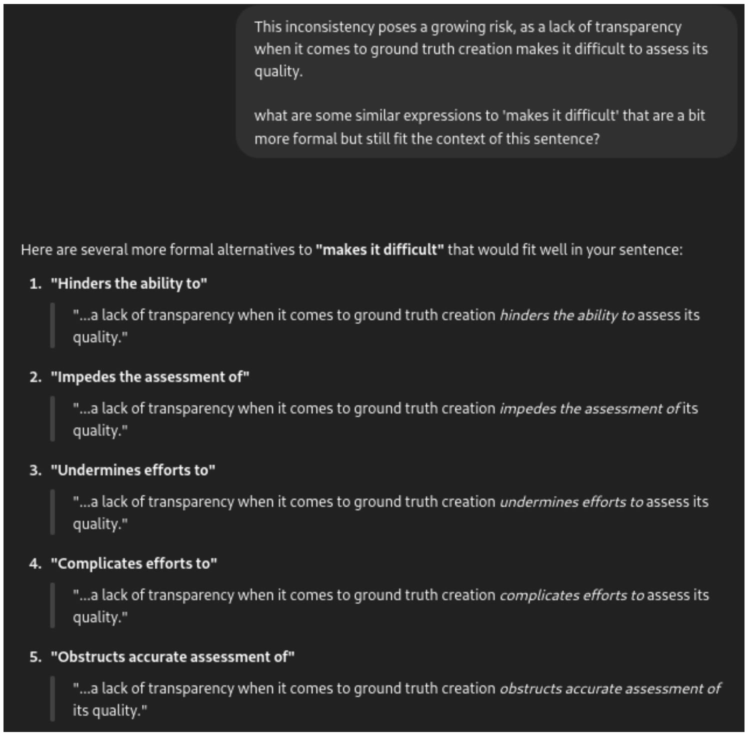
> *She is the most senior engineer. As such, she leads the project.*
> *= She is the most senior engineer. Therefore, she leads the project.*

✅ To refer back to a role, status, or quality (*"as such"* = *"in that capacity/role"*):

- **In that role**
- **In that capacity**
- **In their position as...**
- **In this regard**

> *He is the official spokesperson. As such, he communicates with the press.*
> *= He is the official spokesperson. In that capacity, he communicates with the press.*

Figure 14: GPT-4o suggesting connecting words.



This inconsistency poses a growing risk, as a lack of transparency when it comes to ground truth creation makes it difficult to assess its quality.

what are some similar expressions to 'makes it difficult' that are a bit more formal but still fit the context of this sentence?

Here are several more formal alternatives to **"makes it difficult"** that would fit well in your sentence:

1. **"Hinders the ability to"**

   > "...a lack of transparency when it comes to ground truth creation *hinders the ability to* assess its quality."

2. **"Impedes the assessment of"**

   > "...a lack of transparency when it comes to ground truth creation *impedes the assessment of* its quality."

3. **"Undermines efforts to"**

   > "...a lack of transparency when it comes to ground truth creation *undermines efforts to* assess its quality."

4. **"Complicates efforts to"**

   > "...a lack of transparency when it comes to ground truth creation *complicates efforts to* assess its quality."

5. **"Obstructs accurate assessment of"**

   > "...a lack of transparency when it comes to ground truth creation *obstructs accurate assessment of* its quality."

Figure 15: GPT-4o suggesting similar expressions.