A Need for a Contemporary Field

## Abstract

AI systems are shaped by the training data, algorithm, and evaluation data used to build them. Recent work has shown a need for greater amounts of data, and greater effort in the design, collection, analysis and reporting of data used for building AI systems. Solving these problems may require more effort and resources than the field can offer. Present work makes suggestions aimed at making better use of labor and resources in the AI research field to meet these needs.

## Introduction

The future quality of AI systems will be shaped by the data, evaluation, and reporting practices we establish today. AI systems are a product of the data used to train and evaluate them, thus we indirectly shape their behavior by the processes we design to collect training and evaluation data (Hullman et al., 2022). Training and evaluation data have ongoing effects as they are often re-used (Geiger et al., 2021), likely because of their high cost. Data professionals spend more time on data preparation and cleansing than they do on model selection, training and deployment [[1]](#footnote-21) [[2]](#footnote-22) [[3]](#footnote-23), and yet the focus remains on the latter (Birhane et al., 2022) with data professionals indicating that they prefer the ‘model work’ over the ‘data work’ (Sambasivan et al., 2021).

*“Instead of focusing on the code, companies should focus on developing systematic engineering practices for improving data in ways that are reliable, efficient, and systematic. In other words, companies need to move from a model-centric approach to a data-centric approach.”* - Andrew Ng[[4]](#footnote-24)

Decisions made when collecting data comprise the design, whether or not they are made consciously and carefully. Work over the past decade has emerged explaining limitations of decisions made in the design of commonly re-used datasets, which include a lack of representativeness (Hullman et al., 2022), measurement quality (Jacobs & Wallach, 2021), accounting for the full range of reasonable interpretations in terms of annotations (Cabitza et al., 2023), and completeness in reporting of the annotation process (Daneshjou et al., 2021; Geiger et al., 2021). Evidence points to a need for a more sophisticated approach to the design of datasets in AI fields, especially frequently re-used benchmarks that become ‘load-bearing’ cornerstones in their niches.

Although improvements to data collection processes have been proposed, they are at best slowly being adopted, and at worst being largely ignored. The increasingly sophisticated systems we build have an accompanying cost not only in terms of sheer training data size, but also in terms of the labor required to curate the data (Kandpal & Raffel, 2025). This is exacerbated by the increasing scale required to train systems using the most advanced methods (Villalobos et al., 2024), and increasingly sophisticated qualities we aim to evaluate systems on, that in turn are challenging to define and measure e.g. fairness (Jacobs & Wallach, 2021), and intelligence (Gignac & Szodorai, 2024). Furthermore, as many systems aim to behave similar to humans, the human-like behavior of the sophisticated systems we are evaluating are a prime target for anthropomorphistically biased mis-interpretations of their outputs (Altmeyer et al., 2024).

The degree to which we invest in researching best practices for the design, collection, and analysis of training and evaluation data will determine the real-world performance of the AI systems we will build. It is thus crucial that we put in place better practices. The emerging field of *AI metrics* focuses on developing best practices for the measurement AI system performance (Gignac & Szodorai, 2024), by treating datasets as measurement instruments to be developed and maintained (Welty et al., 2019). However, improving how AI fields collect and label data requires substantial efforts beyond the already substantial efforts invested. Resources must thus be dedicated to determining what best practices are, which will require dedicated research to solve issues unique to training vs. evaluation data, and to implementing said practices.

### Too high a cost?

Questions remain as to how AI research as it is currently conducted can develop and implement solutions. This conclusiory manuscript thus highlights the crucial challenge for academic study of AI in the coming decade: developing an infrastructure that allows for the study of AI, including the data that are its raw materials, with little - or at the very least, substantially less - harmful bias. It highlights shortcomings in the data used for training, and the accompanying ‘ground truth’ (e.g. labels, annotations etc.), along with challenges in gathering these at scale, as well as the scientific and annotation labor necessary to meet these challenges. It highlights the need for identifiable academic publication venues that gather and disseminate works on the study of data design, collection, analysis, and reporting of data for AI training and evaluation, as well as more modern publication formats that allow for dataset requirements to be studied prior to their collection, and for infrastructure that allows the burden of their collection to be distributed among stakeholders. It concludes that, while works like the case study embedded in this thesis are necessary, the various fields studying topics related to AI are poorly positioned to implement them without substantial modernization.

## Data challenges

### Training Data

Improving the way we design and collect training data, especially data that is likely to be widely re-used, is key to developing trustworthy models (Liang et al., 2022; Welty et al., 2019), but requires investment in deliberate and informed design (Hullman et al., 2022), and maintenance (Liem & Demetriou, 2023). Beyond the challenge of gathering data at the necessary scale, is the challenge of ensuring that it is generally representative of the environment where it is to be deployed (Hullman et al., 2022), has sufficiently diverse coverage of the various scenarios it will encounter (Liang et al., 2022), and is as free as possible of bias (Mehrabi et al., 2021). Such a process requires deliberate design, and cannot be compensated for by increasing the amount of data collected (Hullman et al., 2022).

In some instances, data requirements may be too great for this to be currently possible with data that currently exists. *Data scarcity*, or a lack of sufficient training data, is a common issue with contemporary techniques e.g. deep learning, across machine learning fields e.g. computer vision, healthcare, natural language processing (Bansal et al., 2022). A number of solutions have been proposed for once data has been collected (Alzubaidi et al., 2023; Bansal et al., 2022), but a clearer solution might be to begin with more carefully collected data. Despite the need for solutions in this regard, the overwhelming focus in AI fields remains on algorithmic work (Birhane et al., 2022), and not the creation and curation of load-bearing datasets.

This issue is well illustrated with the most contemporary trends of Generative AI systems. Taking the development of Large Language Models (LLMs) as a use-case: trends show massive increases in the requirements of the training data, with the data for Llama 3 including 15T tokens, up from 2T for Llama 2. Although not all details of the datasets have been shared, and setting aside questions of copyright, licensing, and ethical concerns, any available text that is likely to have some quality is scarce compared to these requirements. For example, among the more refined sources of text available are the 6.9M English Wikipedia articles[[5]](#footnote-28) at an estimated 6.24B tokens[[6]](#footnote-30) - articles across all languages comprise only approximately 10 times that amount. If we extend our reach to other repositories, e.g. the approximately 85-90 million academic papers available on Sci-hub[[7]](#footnote-32) or SCOPUS[[8]](#footnote-34), we gain a rough estimate of 700B tokens. A similar figure might be estimated from libegen and the 7.5M[[9]](#footnote-36) books there. The internet archive has some 44M books and texts, which may yield up to 4.4T tokens if we assume ‘texts’ have approximately the same length as books. Thus, a more likely source for the ever-increasing data requirements are repositories of randomly selected text from the internet, like Common Crawl[[10]](#footnote-38). But this too has limits, and we are projected to have too little human-generated text to continue the increase in model size this decade, even if all of Common Crawl is used (Villalobos et al., 2024).

Beyond the issue of merely acquiring data that exists, is the issue of the thus-far-ignored cost of training data. GPT-4 cost an estimated 40M USD to train, including human labor, hardware, and energy. Cost for frontier models is projected to increase at a rate of 2.4 times per year, reaching an estimated 1B USD to train by 2027 (Cottier et al., 2024). Notably, these cost calculations ignore the cost of producing the text itself, and though its value is difficult to calculate, estimates range from 10-1000 times more than the total cost of model training, and would exceed annual revenues of the organizations that are training the models, like OpenAI (Kandpal & Raffel, 2025). Thus, the resources spent do not include the costs of producing training data, for which lack of appropriate compensation has given rise to a number of lawsuits (e.g. Authors Guild vs. OpenAI [[11]](#footnote-40)). In other words, the more valuable thing is the data and not the model, and the few organizations that have the resources to train the models may not have been able to afford training if the data weren’t acquired at a near-0 cost.

Thus, the field of AI must wrestle with opposing issues: we want to train models with data that has high quality - e.g. we want the data to be representative of conditions where the model will be deployed, and thus relevant distributions in the training data must reflect the environment in which the models will be deployed (Hullman et al., 2022) - but the data requirements to train them appear thus far to be ever-increasing (Villalobos et al., 2024), and have thus far not included the actual cost of production (Kandpal & Raffel, 2025). We want the data used in training to be ‘good’ because we want the models to be ‘trustworthy’ - in the case of LLMs, we want to have reason to think they will generate text that includes claims that we think are true, and thus likely prefer text that contains information that is likely to be true. Perhaps our closest approximation to ‘what is likely to be true’ is contained in the perspectives presented across all of academic work and official reports. And yet, even if there were little to no barriers to using all of human academic work to train AI systems, this amount of text pales in quantity to the data requirements. And although thus far some models have been made available for academic use (e.g. Meta’s Llama), it is not clear for how long this will be the case, and when, if ever, academia will have the resources to train its own.

### Evaluation Data

There is substantial room for improvement in terms of the practices of data labeling as well, however implementations of such practices are also costly. The field of *AI metrics* distinguishes between a) observable AI system behavior, and b) the targets of measurement. In other words, a given score on a given ‘AI’ benchmark designed to measure ‘intelligence’ is not itself a direct measure of ‘intelligence’. AI metrics treats the indirectly observable phenomena as *computational constructs* - indirectly observable aspects of AI system behavior, a parallel to *constructs* in other fields - indirectly observable phenomena (Gignac & Szodorai, 2024). Thus the benchmark itself is a measurement instrument, aimed at measuring a specific aspect of the AI system.

The process that gives rise to the labels in the data that, which include the annotation interface and task instructions shown to annotators, is also a measurement instrument (Beck et al., 2022). The case study in this thesis contributes to the field of AI metrics by proposing enriching design, collection, analysis, and reporting of training/evaluation data for AI systems, using knowledge from the social sciences Jacobs & Wallach (2021), metrology (measurement science) (Welty et al., 2019), and work in the computational sciences on ‘ground-truthing’ (Cabitza et al., 2023). It requires a higher cost in terms of scientific labor, as it includes an *a-priori* empirical investigation of the data collection process in addition to the data collection process itself. In other words, it requires research on *how to collect the data*. Applying the knowledge to other tasks would require similar research, in principle for every combination of *construct* (i.e. the latent phenomenon of interest being measured), *content* (e.g. text, video, audio etc. and in some cases also subgroups, e.g. tweets vs. podcast transcripts vs. formal speeches etc.), and for relevant *characteristics* of annotators (i.e. ethnicity, political affiliation, etc.). These costs are in addition to the annotation labor.

Although it is clear there is substantial room for improvement along with the repercussive benefits from such improvements, they will require resources. Despite growing recognition of evaluation data problems (Hullman et al., 2022), which in turn have the potential to lead to harms (Mehrabi et al., 2021), actual progress toward better annotation data practices, from collection, to analysis, to reporting remains slow. Current incentives prioritize immediately observable proxies for accuracy (Birhane et al., 2022) rather than careful measurement design (Beck et al., 2022) or long-term dataset stewardship (Liem & Demetriou, 2023). Beyond the challenge of gathering data at the necessary scale, are issues about the quality of the data itself, specifically the quality of the overall process that created the data. This challenge is essentially one of measurement (Welty et al., 2019).

Datasets for evaluation very often contain human input (Daneshjou et al., 2021; Geiger et al., 2020, 2021; Sav et al., 2023). When collecting annotations, labels, or other forms of input from people in order to construct training/evaluation datasets, we are attempting to collect measurements of latent, unobservable *constructs* (Gignac & Szodorai, 2024; Jacobs & Wallach, 2021). In other words, when we examine data created from the observations of multiple people in aggregate, we do not directly observe e.g. the presence or absence of an object in a digital image; rather we observe the probability that a person from a given population will indicate the presence of absence of the object in the image (Welty et al., 2019). In the parlance of psychology, one cannot directly observe an other’s intelligence as one might observe an other’s height (Gignac & Szodorai, 2024). And even though height is observable, our measurements of it are still imperfect: in using a measurement device like a ruler multiple times, should we measure precisely enough, we would observe variance in each measurement with the true score for height imperfectly represented by our collection of imperfect measurements (Welty et al., 2019).

Any standardized procedure for comparing two or more individuals is treated like a measurement instrument in the social sciences (Urbina, 2014). The repeatable procedures that we use to gather annotations from humans are similarly measurement instruments (Beck et al., 2022). In addition, because we use the datasets that we collect to benchmark the performance of models, they are also measurement instruments (Welty et al., 2019). Given the complexity of measuring unobservable phenomena, instruments in other fields are subjected to scrutiny prior to being considered usable for their intended purpose, in the form of studies that examine the properties of the measurements produced by the instrument. For example, the process of *construct validation* involves estimating the extent to which an instrument measures an unobservable construct (Wehner et al., 2020). It assumes an unknowable true score, and that all attempts to measure the true score are imperfect. There is no single solution to demonstrating the validity of a construct, but rather an accumulation of evidence, across multiple studies, with observations made using different methods (Smith, 2005).

Thus, the field of AI must wrestle with the opposing pressure of accuracy vs. the cost of developing measurement instruments for practically infinite use-cases: we want measurements that are ‘accurate’ - i.e. we want the data to be as clear a representation of our phenomenon of interest as it could be - but developing an accurate instrument is a costly exercise to add to the cost of collecting the data itself. Further, the topics we wish to measure in the ever-increasingly-complex AI systems as themselves measurement challenges. Topics like ‘fairness’ (Jacobs & Wallach, 2021), and ‘intelligence’ (Gignac & Szodorai, 2024) are not only challenging to develop instruments for, but also to define.

## Labor Challenges

### Scientific Labor

Addressing the aforementioned data challenges will require substantial additional efforts. The human labor responsible for the collection, curation, and eventual annotation of training data in addition to the training of models (including researchers, engineers, and managers, but not data center employees and operations staff) is estimated at 29%-49% of the overall cost, which already amounts to tens of millions for frontier models (Cottier et al., 2024). Thus, a key challenge to implementing better practices involves providing labor both in terms of the scientific work necessary to design data collection and annotation processes, but also to provide the annotations themselves. However our current knowledge gathering apparatus - science as it is now practiced and reported - is dated, and inefficient. Thus, a substantial amount of the labor will likely have to come from an already overburdened work force. A re-consideration of what duties comprise academic work is thus warranted, specifically those that contribute to the body of human knowledge, and whether they can be modernized.

One key example involves academic writing and publication, specifically a primary source of scientific labor in academic settings, the PhD student, and their requirements to produce a dissertation or thesis document in order to progress in their career. PhD students are responsible for as much as a third of all academic output (Larivière, 2012), likely by working long hours, including weekend hours, constantly under time-pressure to produce output (Tienoven et al., 2024). Unsurprisingly, PhD students are a vulnerable population, showing a high prevalence of depression and anxiety - as high as 24%, and 17% respectively in a recent meta-analysis (Satinsky et al., 2021). Like all early career researchers (ECRs, i.e. PhD students, Post Doctoral Researchers, and Assistant Professors), PhD students, contribute substantially to the body of human knowledge via academic publication (Rørstad & Aksnes, 2015).

However, PhD candidates are bogged down by dated requirements to write and defend theses despite the declining relevance of the document, and the contribution it makes to poor outcomes for the student. For example, the trend of thesis citations over time shows decline (Larivière et al., 2008). In the Netherlands where this thesis will be defended, evidence has shown that PhD candidates with fixed duration contracts are exceeding that duration by several months, resorting to completing their thesis on their own time and risking failing at completion (Van de Schoot et al., 2013). Though a key ceremonial moment, a thesis defense is substantially less rigorous than a publication: it is curated, as the ‘peer reviewers’ are chosen by the supervisors of the student - in the ‘real world’ of academia, on the other hand, peer reviewers are selected from far broader networks than those of the PhD candidate’s supervisory staff (Larivière, 2012). It is thus not surprising that publications, and not theses, remain the key factor in the assessment of the value of Academics as scientists (Anderson et al., 2022). Although more and more thesis content is being comprised of academic publications anyway (i.e. *thesis by publication*, Jackson (2013)), the substantial labor required to assemble works into a single document, write additional (introductory / conclusiory) chapters that are themselves complete manuscripts or nearly so, then help organize a formal event for the defense take away from more meaningful labor that this workforce could provide.

A second key example involves a cornerstone of trust in science: the editorial peer-review process dates back several hundred years, and exists as a means to encourage and maintain the quality of scholarly work (Kelly et al., 2014). Early career and experienced researchers alike contribute to peer-review[[12]](#footnote-44), labor conservatively valued at over 1.5B USD in 2020 (Aczel et al., 2021). ECRs perform a substantial amount of the work often without credit[[13]](#footnote-45), and in some cases perform the work completely on their own (McDowell et al., 2019), especially in high-volume conferences like NeurIPS (Shah et al., 2018). This is a crucial note, as the submissions to high-volume conferences show massive increases year over year[[14]](#footnote-46) and struggle to meet the review requirements: e.g. NeurIPS 2021 required outreach to recruit over 1k additional peer-reviewers beyond the over 7k initial volunteers, to produce the 31k reviews needed for the conference[[15]](#footnote-47). As conferences continue to grow, the amount of review work is being divided among a regularly stable labor pool, reducing the attention paid to each individual work, a problem exacerbated by the repeat submission of borderline papers - i.e. papers that were nearly accepted (Zhang et al., 2022). This results in important flaws being missed: e.g. (Kapoor & Narayanan, 2023) show a growing crisis across 17 Machine Learning fields from a lack of trustworthy results due to data leakage.

A third key example involves the ever-increasing volume of published manuscripts on AI and related topics[[16]](#footnote-48), which makes it impossible to stay abreast of the overall field. 10% of over 4 million publications indexed on the SCOPUS academic database in 2024 had terms related to AI in their title, keywords or abstracts (see Appendix A), up from around 7% in 2022. The number of submitted manuscripts also increases year over year, with popular conferences like NeurIPS receiving upwards of 12K submissions in 2023[[17]](#footnote-49). These figures do not include preprints posted on servers like arXiv, which show over 42K works with AI related terms in the abstract for 2024, more than doubling the about 17.5K posts in 2019. This makes it more and more difficult to monitor the overall quality of the field (Zhang et al., 2022).

These examples highlight the overburden of academic scientific labor, which in turn raises questions about whether research in fields pertaining to AI can sustainably retain an appropriate level of quality, while meeting increasing demands. Predictably, and perhaps understandably, academics appear to be turning to Large Language Models (LLMs) for assistance, as evidence of its use is showing in academic work in both peer reviews (Liang et al., 2024), and in manuscripts (Gray, 2024). This is problematic as LLMs often report information that is not factual (Wang et al., 2024). However, a more appropriate direction might be to question how cornerstone components of academic work are conducted, and whether they can meet the needs of the field in the future.

### Annotation Labor

The labor required to gather annotations is a key component of the evaluation data for machine learning or AI systems, and a substantial contributor to the related costs. In specialist categories, e.g. Figure Eight, which hires annotators for projects related to defense, salaries can range from 41K - 67K USD per year[[18]](#footnote-51). Although commonly used platforms like mTurk[[19]](#footnote-53) list only minimum rates of .01 USD per task, other services, e.g. Prolific[[20]](#footnote-54), hire on a per-task basis and pay at least minimum wage in the UK. Although these costs may seem manageable, they have the potential to balloon at scale. For example, Open AI hired a San Francisco-based firm that sourced annotation labor from Kenya, Uganda, and India to provide the human inputs necessary to fine tune their models [[21]](#footnote-55). Despite the pay rate of 2 USD per hour being far less than the 7.25 USD federal minimum wage in the US[[22]](#footnote-56), OpenAI spent 600K USD in 2021 to label text as being violent, sexual, or hatespeech alone. While the availability of such services has allowed for rapid gathering of evaluation data for AI-related projects, works that highlight data scarcity warn that it may be insufficient to meet the needs of coming models, and that the costs of such labor will be very high.

Work in the case study of this thesis further suggests that *more* annotation work may be required than per project than is current, as it requires a focus on the development of an annotation instrument prior to primary data collection phases, pre-studies to estimate the required number of annotators, and accounting for possible variance in perspectives. For high-stakes use-cases, such as applications in government or healthcare, we may wish to have data independently examined prior to its certification for use, similarly to how some scholars recommend external examination of AI models prior to deployment (Garrett & Rudin, 2024), which may in turn require further iterations of data collection.

## A contemporary field

### Collaboration to address data and annotation challenges

One approach to both gathering the necessary training data at scale as well as the labels or annotations have been shown in collaborations between scientists and the general public. Online platforms, often initiated by academics, host and facilitate the creation of various resources, ranging from media and other forms of data, labels and annotation projects as well as forums for discussion. For example inaturalist.org is an online community with over 8 million users who make contributions in the form of images taken on their smartphones, and/or labels of the species in the images (Van Horn et al., 2018). Zooniverse.org is an online community of over 2.8M users that hosts projects defined by scientists to gather labels from non-scientists (Fortson et al., 2012). A third example is commonvoice.mozilla.org/en, which is a large dataset of speech transcription in 76 languages, provided by approximately 150k participants (Ardila et al., 2020). Of note, it has been shown that people are willing to even share de-identified medical data as long as they are given sufficient transparency and agency over its use (Liang et al., 2022). This approach simultaneously allows for the contribution of diverse media for annotation, labor needed to annotate them - although individual contributions are not incentivised for academics.

A second similarly scalable approach for collaborative dataset creation that may better fit within academia, as it has the potential to produce credit in the form of publications, might be an adaptation of an academic publication format called the *Registered Report* (Chambers, 2013). In a Registered Report, researchers submit a manuscript that includes information relevant to how the study will be conducted, including motivation of the work (i.e. introduction), details of data collection processes, as well as analyses. Typical review stages apply, i.e. suggestions for revisions or rejections, or the manuscript may receive an *in-principle acceptance*, whereby reviewers and editor agree to a publication should the methods used in the manuscript either follow closely the in-principle accepted version, or appropriate justifications be made for any changes that may have occurred. This format was designed to allow for editorial decisions being made based on the strength of the methods, which also are strengthened by a peer-review process prior to data collection.

The Registered Report format is exceptionally well-suited to the collection of data intended for AI training and/or evaluation. Firstly, it allows for peer-review prior to collection, whereby a panel of experts will provide critiques that will either strengthen the eventual design, or reject it in favor of publishing other stronger designs. Given the scope and resources needed to collect AI datasets, this format may thus allow for a more public critique of the design prior to investment of resource cost of collection. Secondly, it further allows for the submissions of responses in the form of data that conforms to the design in the published manuscript in a decentralized fashion, from multiple stakeholders, thus reducing the bias from any single data collection point and allowing for the sharing of financial and other resource burdens. Thirdly, as online platforms like Zooniverse grow and proliferate to focus on different topics and domains, hosted projects may be attached to approved data collection protocols.

A third approach integrates instruction with data collection by involving bachelor and masters students. the Collaborative Replications and Education Project (CREP) is a crowdsourced initiative where undergraduate students, under faculty supervision, replicate high-impact psychology studies, thus allowing for direct instruction of students while provided needed replications of pivotal studies [[23]](#footnote-60). These replications are pre-registered, and may be published: for example, a meta-analysis of nine student-led replications of the “red-romance effect” found no significant effect (Wagge et al., 2019), and thus functioned both to instruct students and contribute to the scientific record. Similar approaches can be taken towards annotating data, with students replicating registered data collection protocols, or supplying the annotations themselves.

### Maintaining academic artifacts

A number of publication formats currently exist that can assist in thoroughly summarizing and synthesizing evidence for topics in which there is sufficient volume of publications. *Systematic Reviews*[[24]](#footnote-63) and *Meta-Analyses*[[25]](#footnote-64) attempt to gather all relevant works on a topic, synthesize the evidence, extract insights and highlighting research opportunities on a given topic. In fields where there is an exceptional amount of work on a topic, or set of related topics, *Umbrella Reviews*[[26]](#footnote-65) attempt to synthesize collections of systematic reviews and/or meta-analyses on a given topic. While useful, the accuracy and currency of these formats atrophies in fields that are very rapidly developing. *Living Systematic Reviews* are useful for keeping up-to-date information visible, and are particularly useful in rapidly evolving fields (Elliott et al., 2017). Living Systematic Reviews use the same methods as any form of academic review (e.g. Systematic Reviews, Meta Analyses, etc.) for selecting, reviewing, and synthesizing available evidence, but include an *a-priori* commitment to gathering and updating what is reported. Although searching and selecting items for inclusion can be time consuming, they can be assisted with automation (Thomas et al., 2017), for which many open-source, free tools are available (e.g. asreview[[27]](#footnote-66)).

The practice of science produces a number of artifacts beyond the paper that reports results of the work, including data, models, code / notebooks, configurations and lab notes. These artifacts are a relevant part of the work of science, although they are often obscured from view, and like software artifacts, would benefit from regular maintenance (Liem & Demetriou, 2023). Notably, sites like the Open Science Framework[[28]](#footnote-67) have emerged that allow for the flexible documentation of various components of any scientific workflow as well as updates. Similarly, online git-compatible repositories like GitHub can be leveraged not only as containers for the artifacts, but as easily maintainable documentation of the evolution of various projects (Wattanakriengkrai et al., 2022). Taking these a step further, platforms like Quarto[[29]](#footnote-68) allow for directly reproducible documents to be extracted from code agnostic notebooks, thereby shrinking the distance between published papers and the working environment of scientists. With AI-related artifacts specifically, sites like Replicate[[30]](#footnote-69) and Huggingface[[31]](#footnote-70) have emerged that allow for hosting, documentation, and sharing of artifacts. Notably, all of these currently sit outside of academic peer review.

Open Review allows for peer reviews of works, as well as a trail of the reviews.

### AI Metrics should be a recognized (sub) field

Data used for training and evaluation of models is of central importance. Solutions to problems in both training and evaluation data collection require deliberate study for standards to improve and disseminate. In other words, beyond the implementation of currently thought best practices, is the requirement that the best practices be improved along with demands, and knowledge of them disseminated such that general practice is affected.

As it stands, there is no central field of study for this topic, despite its central importance to all AI-related fields. Other disciplines (psychology, economics, software engineering) have entire fields dedicated to the design, collection, analysis, and reporting of data that involves human behavior (psychometrics, econometrics, software testing). Although datasets that become ‘load bearing’ are frequently cited e.g. Imagenet, AI Metrics itself is not a central focus at the most highly cited academic publication venues, nor is there a focused publication venue or conference. Thus, crucial topics remain under-resourced and under-researched.

Benchmarks and measurement

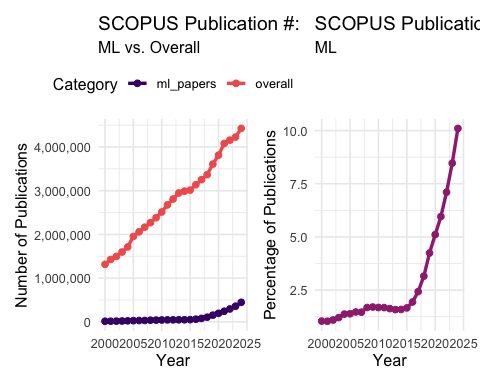
## Concluding Call to Action

Better data design enables better science, more responsible innovation, and safer, more meaningful AI systems, but is costly, even in comparison to the current amount of effort and funding dedicated to AI (and related fields) research. A key priority is that the AI/ML research community, conferences, funding agencies, and publishers must recognize data research, analysis as well as creation, as a research contribution.New formats like data-focused Registered Reports and micropublications should be adopted to 1) distribute efforts, and 2) aggregate knowledge.

Further, we must rethink how we allocate labor. Load-bearing datasets require more effort than the initial collection and subsequent distribution, but also maintenance. This beyond research required to create useful datasets, we also require resources to maintain them.

# Appendix

## Appendix A: Citation Trends Plot



## Appendix B: Search terms

### SCOPUS:

for AI related topics: TITLE-ABS-KEY ( ( ( ( machine OR deep OR reinforcement OR supervised OR unsupervised ) AND learning ) OR ( “neural networks” ) OR ( ai OR “artificial intelligence” ) ) ) AND PUBYEAR > 1999 AND PUBYEAR < 2027

for overall publication records: PUBYEAR > 1999 AND PUBYEAR < 2027

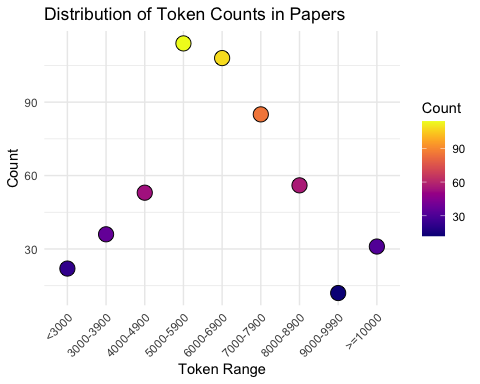
### arXiv:

[Abstract] AI or “artificial intelligence” OR machine AND learning OR supervised AND learning OR reinforcement AND learning OR neural AND networks

17,459 results in 2019 23,923 results in 2020 27,610 results in 2021 29,690 results in 2022 33,419 results in 2023 42,183 in 2024

## Appendix C: Token estimates

Taking a study on the word length requirements of education journals as a proxy, Fairbairn et al. (2009) report the following figures:



# Estimate midpoints for each bin  
midpoints <- c(2500, 3450, 4450, 5450, 6450, 7450, 8450, 9500, 10500)  
  
# Add midpoints to the dataframe  
paper\_lengths$Midpoint <- midpoints  
  
# Estimate mean token count  
mean\_est <- sum(paper\_lengths$Midpoint \* paper\_lengths$Count) / sum(paper\_lengths$Count)  
  
# Calculate weighted variance  
var\_est <- sum(paper\_lengths$Count \* (paper\_lengths$Midpoint - mean\_est)^2) / sum(paper\_lengths$Count)  
  
# Take square root to get standard deviation  
sd\_est <- sqrt(var\_est)  
  
# Print result  
paste()

character(0)

print(paste("SD words: ", sd\_est))

[1] "SD words: 1925.27260683454"

print(paste("Mean words: ", mean\_est))

[1] "Mean words: 6342.166344294"

https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them?utm\_source=chatgpt.com According to Open AI, a token is 3/4 of a word

sd\_tokens <- sd\_est\*1.25  
mean\_tokens <- mean\_est\*1.25   
  
print(paste("SD Tokens: ", round(sd\_tokens, 2)))

[1] "SD Tokens: 2406.59"

print(paste("Mean Tokens: ", round(mean\_tokens, 2)))

[1] "Mean Tokens: 7927.71"

paste( round(((mean\_tokens-sd\_tokens) \* 4500000) / 1000000000, 2),   
 " Billion to",   
 round (((mean\_tokens+sd\_tokens) \* 4500000) / 1000000000, 2),   
 "Billion tokens per year from academia.")

[1] "24.85 Billion to 46.5 Billion tokens per year from academia."

paste( 4.8\*1.25, "Billion tokens from Wikipedia.")

[1] "6 Billion tokens from Wikipedia."

Wikipedia: https://en.wikipedia.org/wiki/Wikipedia%3ASize\_of\_Wikipedia?utm\_source=chatgpt.com

5000000000/2000000000000\*100

[1] 0.25

Sci Hub: https://www.sci-hub.mk/ 88343822 documents as of 9 May 2025 13:47.

(7927.71 \* 88343822) / 1000000000

[1] 700.3642

https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/ LibGen

7.5 million books and 81 million research papers.

# at about 100k words per book  
(100000\*7500000) / 1000000000

[1] 750

(100000\*44000000) / 1000000000

[1] 4400

Llama 2 was 2T tokens, Llama 3 was 15T. I didn’t really see where they got their data from. So I made some guesses.

Looked at Wiki. The whole thing is like 5B.

Sci-Hub, 88.3M papers, which at 8kish tokens is an upper boundary of 700B or so.

7.5 million books on libgen you get about another 700B or so, at 100k-ish per words.

Internet archive has some 44 **million** books. 4.4T.

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers’ time spent on peer review. *Research Integrity and Peer Review*, *6*(1), 1–8.

Altmeyer, P., Demetriou, A. M., Bartlett, A., & Liem, C. (2024). Position: Stop making unscientific AGI performance claims. *arXiv Preprint arXiv:2402.03962*.

Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamarı́a, J., Albahri, A. S., Al-Dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., et al. (2023). A survey on deep learning tools dealing with data scarcity: Definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, *10*(1), 46.

Anderson, C. G., McQuaid, R. W., & Wood, A. M. (2022). The effect of journal metrics on academic resume assessment. *Studies in Higher Education*, *47*(11), 2310–2322.

Ardila, R., Branson, M., Davis, K., Kohler, M., Meyer, J., Henretty, M., Morais, R., Saunders, L., Tyers, F., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the twelfth language resources and evaluation conference* (pp. 4218–4222). European Language Resources Association. <https://aclanthology.org/2020.lrec-1.520/>

Bansal, M. A., Sharma, D. R., & Kathuria, D. M. (2022). A systematic review on data scarcity problem in deep learning: Solution and applications. *ACM Computing Surveys (Csur)*, *54*(10s), 1–29.

Beck, J., Eckman, S., Chew, R., & Kreuter, F. (2022). Improving labeling through social science insights: Results and research agenda. *International Conference on Human-Computer Interaction*, 245–261.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The values encoded in machine learning research. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Cabitza, F., Campagner, A., & Basile, V. (2023). Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*, 6860–6868.

Chambers, C. D. (2013). Registered reports. *A New Publishing Initiative at Cortex. Cortex*, *49*(3), 609–610.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., Besiroglu, T., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv Preprint arXiv:2405.21015*.

Daneshjou, R., Smith, M. P., Sun, M. D., Rotemberg, V., & Zou, J. (2021). Lack of transparency and potential bias in artificial intelligence data sets and algorithms: A scoping review. *JAMA Dermatology*, *157*(11), 1362–1369.

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., et al. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, *91*, 23–30.

Fairbairn, H., Holbrook, A., Bourke, S., Preston, G., Cantwell, R., & Scevak, J. (2009). A profile of education journals. *AARE 2008 International Educational Research Conference*, 1–20.

Fortson, L., Masters, K., Nichol, R., Edmondson, E., Lintott, C., Raddick, J., & Wallin, J. (2012). Galaxy zoo. *Advances in Machine Learning and Data Mining for Astronomy*, *2012*, 213–236.

Garrett, B. L., & Rudin, C. (2024). Testing AI. *Available at SSRN 4948789*. <https://papers-ssrn-com.tudelft.idm.oclc.org/sol3/papers.cfm?abstract_id=4948789>

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

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336.

Gignac, G. E., & Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. *Intelligence*, *104*, 101832.

Gray, A. (2024). ChatGPT" contamination": Estimating the prevalence of LLMs in the scholarly literature. *arXiv Preprint arXiv:2403.16887*.

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348.

Jackson, D. (2013). Completing a PhD by publication: A review of australian policy and implications for practice. *Higher Education Research & Development*, *32*(3), 355–368.

Jacobs, A. Z., & Wallach, H. (2021). Measurement and fairness. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

Kandpal, N., & Raffel, C. (2025). Position: The most expensive part of an LLM should be its training data. *arXiv Preprint arXiv:2504.12427*.

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, *4*(9).

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer review in scientific publications: Benefits, critiques, & a survival guide. *Ejifcc*, *25*(3), 227.

Larivière, V. (2012). On the shoulders of students? The contribution of PhD students to the advancement of knowledge. *Scientometrics*, *90*(2), 463–481.

Larivière, V., Zuccala, A., & Archambault, É. (2008). The declining scientific impact of theses: Implications for electronic thesis and dissertation repositories and graduate studies. *Scientometrics*, *74*(1), 109–121.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., et al. (2024). Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv Preprint arXiv:2403.07183*.

Liang, W., Tadesse, G. A., Ho, D., Fei-Fei, L., Zaharia, M., Zhang, C., & Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. *Nature Machine Intelligence*, *4*(8), 669–677.

Liem, C. C., & Demetriou, A. M. (2023). Treat societally impactful scientific insights as open-source software artifacts. *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, 150–156.

McDowell, G. S., Knutsen, J. D., Graham, J. M., Oelker, S. K., & Lijek, R. S. (2019). Co-reviewing and ghostwriting by early-career researchers in the peer review of manuscripts. *Elife*, *8*, e48425.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35.

Rørstad, K., & Aksnes, D. W. (2015). Publication rate expressed by age, gender and academic position–a large-scale analysis of norwegian academic staff. *Journal of Informetrics*, *9*(2), 317–333.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–15.

Satinsky, E. N., Kimura, T., Kiang, M. V., Abebe, R., Cunningham, S., Lee, H., Lin, X., Liu, C. H., Rudan, I., Sen, S., et al. (2021). Systematic review and meta-analysis of depression, anxiety, and suicidal ideation among ph. D. students. *Scientific Reports*, *11*(1), 14370.

Sav, A.-G., Demetriou, A. M., & Liem, C. C. (2023). Annotation practices in societally impactful machine learning applications: What are popular recommender systems models actually trained on? *Perspectives@ RecSys*.

Shah, N. B., Tabibian, B., Muandet, K., Guyon, I., & Von Luxburg, U. (2018). Design and analysis of the NIPS 2016 review process. *Journal of Machine Learning Research*, *19*(49), 1–34.

Smith, G. T. (2005). On construct validity: Issues of method and measurement. *Psychological Assessment*, *17*(4), 396.

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., et al. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology*, *91*, 31–37.

Tienoven, T. P. van, Glorieux, A., Minnen, J., & Spruyt, B. (2024). Caught between academic calling and academic pressure? Working time characteristics, time pressure and time sovereignty predict PhD students’ research engagement. *Higher Education*, *87*(6), 1885–1904.

Urbina, S. (2014). *Essentials of psychological testing*. John Wiley & Sons.

Van de Schoot, R., Yerkes, M. A., Mouw, J. M., & Sonneveld, H. (2013). What took them so long? Explaining PhD delays among doctoral candidates. *PloS One*, *8*(7), e68839.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8769–8778.

Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2024). Position: Will we run out of data? Limits of LLM scaling based on human-generated data. *Forty-First International Conference on Machine Learning*.

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., IJzerman, H., Legate, N., & Grahe, J. (2019). A demonstration of the collaborative replication and education project: Replication attempts of the red-romance effect. *Collabra: Psychology*, *5*(1).

Wang, Y., Wang, M., Manzoor, M. A., Liu, F., Georgiev, G. N., Das, R. J., & Nakov, P. (2024). Factuality of large language models: A survey. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 19519–19529). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1088>

Wattanakriengkrai, S., Chinthanet, B., Hata, H., Kula, R. G., Treude, C., Guo, J., & Matsumoto, K. (2022). GitHub repositories with links to academic papers: Public access, traceability, and evolution. *Journal of Systems and Software*, *183*, 111117.

Wehner, C., Roemer, L., & Ziegler, M. (2020). Construct validity. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 875–878). Springer International Publishing. <https://doi.org/10.1007/978-3-319-24612-3_1288>

Welty, C., Paritosh, P., & Aroyo, L. (2019). Metrology for AI: From benchmarks to instruments. *arXiv Preprint arXiv:1911.01875*.

Zhang, Y., Yu, F.-Y., Schoenebeck, G., & Kempe, D. (2022). A system-level analysis of conference peer review. *Proceedings of the 23rd ACM Conference on Economics and Computation*, 1041–1080. <https://doi.org/10.1145/3490486.3538235>

1. Anaconda data science report 2022: https://www.anaconda.com/resources/whitepaper/state-of-data-science-report-2022 [↑](#footnote-ref-21)
2. Anaconda data science report 2023: https://know.anaconda.com/rs/387-XNW-688/images/Anaconda%202023%20State%20of%20Data%20Science%20Report.pdf [↑](#footnote-ref-22)
3. Anaconda data science report 2024: https://www.anaconda.com/wp-content/uploads/2025/01/Anaconda\_SODS.pdf [↑](#footnote-ref-23)
4. [landing.ai](https://landing.ai/data-centric-ai#:~:text=%E2%80%9CInstead%20of%20focusing%20on%20the,a%20data%2Dcentric%20approach.%E2%80%9D) [↑](#footnote-ref-24)
5. According to [wikipedia](https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia#:~:text=Monthly%20statistics-,Number%20of%20pages,pages%20are%20created%20than%20articles.) on 9 May, 2025. [↑](#footnote-ref-28)
6. Using the conversion method of words to tokens recommended by [Open AI](https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them). [↑](#footnote-ref-30)
7. According to [sci-hub](https://www.sci-hub.se/about) on 9 May, 2025. [↑](#footnote-ref-32)
8. According to [SCOPUS](https://blog.scopus.com/posts/scopus-roadmap-whats-new-in-2022#:~:text=There%20are%20currently%2087%2B%20million,new%20articles%20per%20day%20indexed.) on 9 May, 2025. [↑](#footnote-ref-34)
9. According to [The Atlantic](https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/LibGen). [↑](#footnote-ref-36)
10. According to [Common Crawl](https://commoncrawl.org/) on 9 May, 2025. [↑](#footnote-ref-38)
11. https://authorsguild.org/app/uploads/2023/12/Authors-Guild-OpenAI-Microsoft-Class-Action-Complaint-Dec-2023.pdf [↑](#footnote-ref-40)
12. https://www.elsevier.com/connect/more-early-career-researchers-are-stepping-up-to-peer-review [↑](#footnote-ref-44)
13. https://elifesciences.org/inside-elife/982053f4/early-career-researchers-views-on-peer-review [↑](#footnote-ref-45)
14. https://github.com/tranhungnghiep/AI-Conference-Info?utm\_source=chatgpt.com [↑](#footnote-ref-46)
15. https://neuripsconf.medium.com/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f [↑](#footnote-ref-47)
16. Search terms included AI, artificial intelligence, machine learning, and neural networks. See Appendix B for details. [↑](#footnote-ref-48)
17. https://github.com/tranhungnghiep/AI-Conference-Info?utm\_source=chatgpt.com [↑](#footnote-ref-49)
18. according to [glassdoor](https://www.glassdoor.com/Salary/Figure-Eight-Federal-data-annotation-specialist-Salaries-E4588136_D_KO21,47.htm) on May 28, 2025 [↑](#footnote-ref-51)
19. https://www.mturk.com/pricing#:~:text=MTurk%20Fee,per%20assignment%20or%20bonus%20payment. [↑](#footnote-ref-53)
20. https://www.prolific.com/participants [↑](#footnote-ref-54)
21. https://www.business-humanrights.org/pt/%C3%BAltimas-not%C3%ADcias/openai-and-sama-hired-underpaid-workers-in-kenia-to-filter-toxic-content-for-chatgpt/ [↑](#footnote-ref-55)
22. [usa.gov](https://www.usa.gov/minimum-wage#:~:text=The%20federal%20minimum%20wage%20is,tips%20is%20%242.13%20per%20hour.) [↑](#footnote-ref-56)
23. [Collaborative Replications and Education Project](https://osf.io/wfc6u/) [↑](#footnote-ref-60)
24. https://en.wikipedia.org/wiki/Systematic\_review [↑](#footnote-ref-63)
25. https://en.wikipedia.org/wiki/Meta-analysis [↑](#footnote-ref-64)
26. https://en.wikipedia.org/wiki/Umbrella\_review [↑](#footnote-ref-65)
27. https://asreview.nl/ [↑](#footnote-ref-66)
28. https://osf.io/ [↑](#footnote-ref-67)
29. https://quarto.org/ [↑](#footnote-ref-68)
30. https://replicate.com/ [↑](#footnote-ref-69)
31. https://huggingface.co/ [↑](#footnote-ref-70)