A Need for a Contemporary Field

The future qualities of AI systems will be shaped by the data and evaluation practices we establish today. AI systems are a direct product of the data used to train and evaluate them. We thus indirectly shape their behavior via the processes we use to design, gather, analyze, and report training and evaluation data (Hullman et al., 2022). The datasets we create have ongoing effects as they are often re-used (Geiger et al., 2021), likely because of the amount of work necessary to create the data. Data professionals spend more time on data preparation and cleansing than they do on model selection, training and deployment [[1]](#footnote-20) [[2]](#footnote-21) [[3]](#footnote-22), yet the focus remains on the latter (Birhane et al., 2022).

*“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-23)

Evidence points to a need for a more sophisticated approach to the collection of datasets in AI fields, especially frequently re-used benchmarks become ‘load-bearing’ cornerstones in their niches. Work over the past decade has emerged explaining limitations 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).

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 the labor required to curate the data (e.g. Large Language Models or LLMs) (Kandpal & Raffel, 2025). This is exacerbated by the 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 better 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 labels 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.

Firstly, improving the way we design and collect of 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 will require investment in deliberate and informed design, distribution and maintenance. Focus must be put into solving sampling problems that are not solved even when the dataset is very large, e.g. sampling data for training that is representative of the environment or desired behavior in deployment conditions (Hullman et al., 2022). In some instances, data requirements may be too great for this to be currently possible. *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). For example, recent generative language models are increasingly being trained with text crawled randomly from the internet over more carefully curated sources that contain information likely to be true, like scientific texts and wikipedia [[5]](#footnote-25), and still this may not be able to keep pace with increasing demands (Villalobos et al., 2024). A number of solutions have been proposed for once data has been collected (Bansal et al., 2022), but a clearer solution might be to begin with more carefully collected data. The challenge is the enormous cost to create large training datasets, with one estimate for LLMs being between up to 1000 times the training costs, which in turn are estimated as being tens of millions USD (Kandpal & Raffel, 2025). Despite the need for solutions in this regard, the overwhelming focus remains on algorithmic work (Birhane et al., 2022), and not the creation and curation of load-bearing training datasets.

Secondly, there is substantial room for improvement in terms of the practices of data labeling as well. Implementations of such practices are also costly. The field of AI metrics distinguishes between a) observable AI system behavior from b) the targets of measurement - e.g. performance on a benchmark from artificial intelligence - by treating them as *computational constructs* - indirectly observable aspects of AI system behavior (**gignac?**). Thus the process that gives rise to the data, 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. Open questions remain for the primary case-study as well as for the field as a whole, all of which anticipate future studies, and by extension further efforts. Although it is clear there is substantial room for improvement along with the repercussive benefits from such improvements, they will require resources.

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 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

Contemporary training methods require increasingly large amounts of training data. This has lead to to the development of techniques aimed at increasing and/or augmenting available data (Alzubaidi et al., 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).

These issues are well illustrated with the most contemporary trends of Generative AI systems. Taking the development of Large Language Models (LLMs) as a use-case: the training dataset for Llama 3 included 15T tokens, up from 2T for Llama 2. Although not all details of the datasets have been shared, and setting aside questions of copywrite, licensing, and ethical concerns, any available text that is likely to have some quality is limited compared to these requirements. Among the refined sources of text available, Wikipedia, which comprises some 6.9M English articles, comprised of approximately 62M pages over all languages, is an estimated 5 billion tokens[[6]](#footnote-26). If we take the approximate 4.4M papers published in 2024 and indexed on Scopus as an indication, academia published an estimated 45B tokens in that year. If we extend our reach to other repositories, e.g. the approximately 85 million academic papers available on Sci-hub[[7]](#footnote-28) or SCOPUS[[8]](#footnote-29) would result in an estimate of 700B. A similar figure might be estimated from libegen and the 7.5M[[9]](#footnote-30) books there. While academic pursuits result in increasing token counts, we have immediate access to a set of approximately 1.5T. Internet archive has some 44M books, which may yield up to 4.4T, although we expect duplicates with the library genesis archive. Thus, a more likely source for the ever-increasing data requirements are repositories like Common Crawl[[10]](#footnote-32). 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 (Villalobos et al., 2024).

The largest frontier models cost tens of millions of USD to train, and a third to a half of that cost is human labor. GPT-4 cost an estimated 40M USD for hardware (chips, servers, and networking hardware) and energy, which is projected to increase at a rate of 2.4 X per year, reaching an estimated 1B USD to train by 2027 (Cottier et al., 2024). Notably, this cost exceeds annual revenues in companies training large scale LLMs (Kandpal & Raffel, 2025). Human labor responsible for the collection, curation, and eventual annotation of training data in addition to the training of the model (including researchers, engineers, and managers, but not data center employees and operations staff) is estimated at 29%-49% of the overall cost (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 X more than the total cost of the training of the models (Kandpal & Raffel, 2025). In other words, the more valuable thing is the data and not the model, and the lack of appropriate compensation for its use has given rise to a number of lawsuits (e.g. Authors Guild vs. OpenAI [[11]](#footnote-33)).

Thus, the field of AI must wrestle with two 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). 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.

### Evaluation data

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 rather than careful measurement design or long-term dataset stewardship (Birhane et al., 2022). 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 (Geiger et al., 2020, 2021; **sav?**). 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). 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 as ‘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.

### AI metrics is a 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.

## Labor Challenges

### Scientific Labor

As it stands, our current knowledge gathering apparatus - science as it is now practiced and reported - is overburdened and inefficient. The ever increasing volume of published manuscripts on AI and related topics[[12]](#footnote-38) makes it impossible to stay abreast of the overall field: 10% of over 4 million publications indexed on the SCOPUS academic database in 2024, up from around 7% in 2022, had terms related to AI in their title, keywords or abstracts (see Appendix A). The number of submitted manuscripts also increases year over year, with popular conferences like NeurIPS receiving upwards of 12k submissions in 2023[[13]](#footnote-39). This makes it more and more difficult to find reviewers, and by extension to monitor the overall quality of the field (Zhang et al., 2022). 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.

A primary resource is the labor of academic students, from the bachelor to the PhD level. One study in Canada in 2012 showed that one third of all academic output comes form PhD students (Larivière, 2012). There are clear suggestions to better allocation of this work: e.g. PhD candidates are bogged down by requirements to write and defend theses despite the decreasing trend of thesis citations over time (Larivière et al., 2008), and evidence 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). A thesis defense is a more curated experience as the ‘peer reviewers’ are chosen by the supervisors of the student - on the other hand, peer reviewers in the world of academic publication are far broader than the networks of the PhD candidate’s supervisory staff (Larivière, 2012). Understandably, publications, and not theses, remain the key factor in the assesment of their value as scientists (Anderson et al., 2022). Predictably, academics appear to be turning to 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 overburden raises questions beyond the poor evaluation of the models that underlie ‘AI’ to the ‘AI’ research process itself, as well as to its likelihood of applying improvements.

Collaborative Replications and Education Project (CREP) https://osf.io/wfc6u/

### Annotation Labor

It has been reported that 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 [[14]](#footnote-41). Despite the rate of pay being far less than the federal minimum wage in the US, ranging from to , OpenAI spent 600k USD in 2021 to label text as being violent, sexual, or hatespeech alone. Commonly used platforms like mTurk[[15]](#footnote-42), and Prolific[[16]](#footnote-43) pay at or near minimum wage in the US. 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.

Work in the case study of this thesis further suggests that *more* annotation work may be required than is used for current projects, 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.

### Collaboration to address data requirement 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 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). These collectively allow for the contribution of diverse media for annotation, labor needed to

Another similarly scalable infrastructure for dataset creation might be possible by adapting an academic publication format called the Registered Report (Chambers, 2013). Initially designed to compensate for editorial decisions being made based on the results, rather than the quality of the methods. In many fields, aspects of the data collection design, as well as the design of analysis and prediction of results occur *a-priori*, in principle not to bias interpretation of results. 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. Thus acceptance of publications is 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 datasets intended for AI training and/or evaluation. Firstly, they allow 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 could be adapted such that it is published in its entirety prior to data collection. This may thus allow for a more public critique of the design prior to paying the resource cost of collection, and further allow 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.

## Reporting challenges

### Volume and synthesis

Sheer volume of academic publications makes it difficult to stay current. This doesn’t encapsulate all of the scholarly output produced. Receiving declining attention are theses. Components currently largely absent are peer-reviews (although open review tries to address this) and rejected manuscripts - many of which get resubmitted [citation]. Partially absent are teaching materials, conference materials, e.g. posters, slides, and recordings of talks etc. Importantly, code scripts and/or notebooks, various forms of data, parameters and settings, and the actual model(s) produced are often absent. This also doesn’t include replications, which have been shown to be useful in other fields. Only recently are there readily available guidelines to more completely report key information: data/model | sheets / cards, (Geiger et al., 2021).

Many of the outputs of academic work would benefit from some sort of maintenance. i.e. we must treat academic insights like software artifacts, with management and possibly updates.

Reviewer quality, and the difficulty in finding reviewers.

### Contemporary Approaches to Academic publication

OSF.io and github have become repositories for materials. Open Review allows for peer reviews of works, as well as a trail of the reviews.

Living systematic reviews are useful for rapidly evolving fields. Notably, works are appearing that aim to develop AI tools to assist in the writing of such reviews by assisting with screening [].

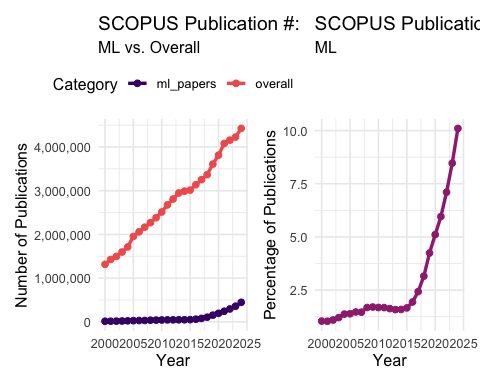
## Concluding Call to Action

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

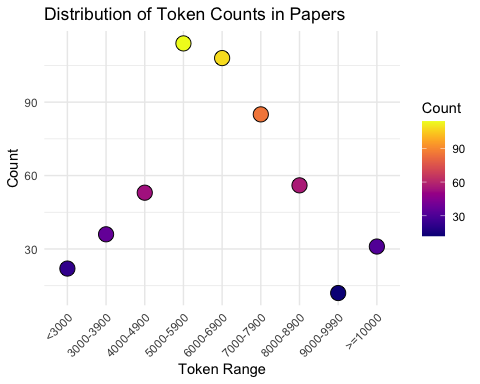
### arXiv:

[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.

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.

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.

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*.

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.

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.

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

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*.

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-20)
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-21)
3. Anaconda data science report 2024: https://www.anaconda.com/wp-content/uploads/2025/01/Anaconda\_SODS.pdf [↑](#footnote-ref-22)
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-23)
5. For example, LLama 2 required 2T tokens for training, whereas Llama 3 required 15T tokens. All of Wikipedia and all scientific texts likely collectively amount to less than 1T tokens. [↑](#footnote-ref-25)
6. According to [wikipedia](https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia#:~:text=Monthly%20statistics-,Number%20of%20pages,pages%20are%20created%20than%20articles.) at 9 May, 2025. [↑](#footnote-ref-26)
7. https://www.sci-hub.se/about [↑](#footnote-ref-28)
8. https://blog.scopus.com/posts/scopus-roadmap-whats-new-in-2022#:~:text=There%20are%20currently%2087%2B%20million,new%20articles%20per%20day%20indexed. [↑](#footnote-ref-29)
9. According to [The Atlantic](https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/LibGen). [↑](#footnote-ref-30)
10. https://commoncrawl.org/ [↑](#footnote-ref-32)
11. https://authorsguild.org/app/uploads/2023/12/Authors-Guild-OpenAI-Microsoft-Class-Action-Complaint-Dec-2023.pdf [↑](#footnote-ref-33)
12. terms included AI, artificial intelligence, machine learning, and neural networks. See Appendix for specific search strings. [↑](#footnote-ref-38)
13. https://github.com/tranhungnghiep/AI-Conference-Info?utm\_source=chatgpt.com [↑](#footnote-ref-39)
14. 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-41)
15. https://en.wikipedia.org/wiki/Amazon\_Mechanical\_Turk [↑](#footnote-ref-42)
16. https://www.prolific.com/participants [↑](#footnote-ref-43)