Position: Will we run out of data? Limits of LLM scaling based on human-generated data

Pablo Villalobos [1] **Anson Ho** [1] **Jaime Sevilla** [1 2] **Tamay Besiroglu** [1 3] **Lennart Heim** [1 4] **Marius Hobbhahn** [1 5]

Abstract

We investigate the potential constraints on LLM scaling posed by the availability of public human-generated text data. We forecast the growing demand for training data based on current trends and estimate the total stock of public human text data. Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained. We explore how progress in language modeling can continue when human-generated text datasets cannot be scaled any further. We argue that synthetic data generation, transfer learning from data-rich domains, and data efficiency improvements might support further progress.

1. Introduction

Recent progress in language modeling has relied heavily on unsupervised training on vast amounts of human-generated text, primarily sourced from the web or curated corpora (Zhao et al., 2023). The largest datasets of *human-generated public text data*, such as RefinedWeb, C4, and RedPajama, contain tens of trillions of words collected from billions of web pages (Penedo et al., 2023; Together.ai, 2023).

The demand for public human text data is likely to continue growing. In order to scale the size of models and training runs efficiently, large language models (LLMs) are typically trained according to neural scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). These relationships imply that increasing the size of training datasets is crucial for efficiently improving the performance of LLMs.

[1]Epoch [2]University of Aberdeen [3]MIT CSAIL [4]Centre for the Governance of AI [5]University of Tübingen. Correspondence to: Pablo Villalobos <pablo@epochai.org>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

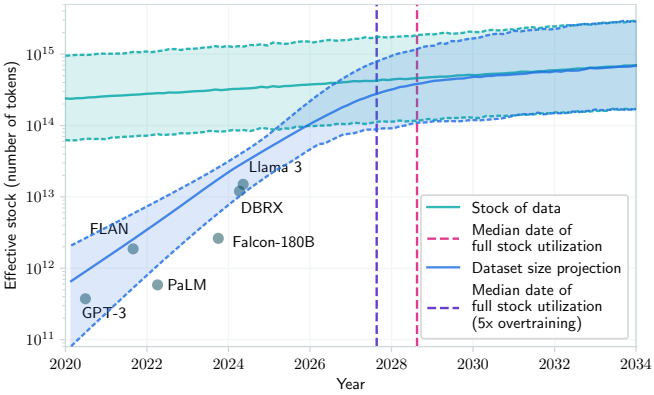

Figure 1. Projections of the effective stock of human-generated public text and dataset sizes used to train notable LLMs. The intersection of the stock and dataset size projection lines indicates the median year (2028) in which the stock is expected to be fully utilized if current LLM development trends continue. At this point, models will be trained on dataset sizes approaching the total effective stock of text in the indexed web: around 4e14 tokens, corresponding to training compute of ∼5e28 FLOP for non-overtrained models. Individual dots represent dataset sizes of specific notable models. The model is explained in Section 2

In this position paper, we argue that human-generated public text data cannot sustain scaling beyond this decade. To support this conclusion, we develop a model of the growing demand for training data and the production of public human text data. We use this model to predict when the trajectory of LLM development will fully exhaust the available stock of public human text data. We then explore a range of potential strategies to circumvent this constraint, such as synthetic data generation, transfer learning from data-rich domains, and the use of non-public data.[1]

1.1. Related work

Stock of internet data Several studies have sought to quantify the internet's size and information content. Murray H. & Moore (2000) estimated the internet's size at approximately 2.1 billion unique web pages containing 21 terabytes of data.

[1]The code used in our analysis can be found at https://epochai.org/code/data-stock.

Coffman & Odlyzko (1998) and Odlyzko (2016) found that public internet traffic experienced a rapid growth rate of approximately 100% per year in the early 1990s, which slowed down to double-digits in the late 2010s, particularly in developed countries.

More recently, Reinsel et al. (2018) estimate the total amount of new data created, captured, or replicated worldwide in any given year to be 33 billion terabytes. Unfortunately, the analysis does not break this down into different data modalities (e.g. images, videos, or text data). Focusing just on Google's index, van den Bosch et al. (2016) estimated the stock from 2006 to 2015, finding that it varied significantly over time but is on the order of tens of billions of web pages.

Data bottlenecks in machine learning Muennighoff et al. (2023) studied several techniques to mitigate data scarcity for training LLMs. In particular, they considered repeating data, adding more code data, and relaxing the quality filters used during data preprocessing. They quantified the loss of performance when using these techniques to compensate for a smaller data budget, finding that both repeating data and including more code data can compensate for a decrease of up to 75% in the text data budget. Xue et al. (2023) also studied multi-epoch training as a solution for data scarcity. Nostalgebraist (2022) argued that high-quality training data would soon become a bottleneck for machine learning.

Leading AI researchers have expressed concerns about data availability limiting the progress of machine learning systems. Dario Amodei, the CEO of Anthropic, estimates a 10% chance that the scaling of AI systems could stagnate due to insufficient data (Roose & Newton, 2023). This underscores the importance of investigating the limitations posed by the finite supply of public human text data.

Estimate	Median	95% CI
Common Crawl	130T	[100T, 260T]
Indexed web	510T	[130T, 2100T]
Whole web	3100T	[1900T, 5200T]
Images	300T	N/A
Video	1350T	N/A

Table 1. Estimates of the stock of data on the web in tokens.[3] In the case of images and video we only have point estimates.

2. A model of data scarcity

The core question we aim to answer is whether the limited availability of public human text data could constrain further LLM scaling. We consider two key variables: the total amount of public human text data available for use ("data stock") and what quantity of this data is actually used in practice during LLM training ("dataset size"). In this section, we develop a model to project both the data stock and dataset sizes.

2.1. Quantifying dataset sizes

Specifying our model requires being explicit about how we quantify "data". To this end, we define the *dataset size* as the number of *tokens* in the training dataset of interest.[4] In large samples of English text, one token usually corresponds to around 0.8 words (see Appendix E).[5]

One limitation of this definition is that the size of a text corpus in tokens depends on how the text is tokenized. That said, in practice, the number of tokens in a corpus does not vary greatly between common tokenizers.[6] Moreover, the two most prominent alternatives – the number of words and the storage size in bytes – can vary significantly between modalities or even be undefined.[7]

2.2. Estimating data stocks

The first main variable of our model is the data stock S. We estimate this by calculating the size of the indexed web and the amount of data that is contained in the average web page, using statistics from Common Crawl.

Since web data contains many low-quality segments of text that do not contribute to model performance (Penedo et al., 2023), we adjust our estimate to account for differences in data quality. We also adjust for the possibility of multi-epoch training. We explain these adjustments in greater detail in Section 2.3. As a further robustness check, we estimate the amount of internet text generated each year based on the world population. Table 1 shows the results of these estimates.

We model our uncertainty about all the observed variables of

[3] Video and image stock estimates are transformed into an equivalent number of text tokens as explained in Appendix D.

[4] Tokenization is the process of encoding text or other types of data using discrete symbols that can be fed into models Zhao et al. (2023). The resulting discrete symbols are known as tokens. The most common choice today is sub-word tokenization, in which each token corresponds to a piece of a word.

[5] The number of tokens in the dataset should not be confused with the number of tokens seen during training, which could be greater than the dataset size if training occurs over multiple epochs.

[6] We make our estimates based on cl100k_base, a byte-pair encoding (BPE) tokenizer from OpenAI (OpenAI, 2024). In Appendix E we show that commonly used tokenizers produce between 0.5 and 0.2 tokens per byte of text, so we expect our results to be similar using other common tokenizers. In the case of image and video we explain the tokenization estimates in Appendix D.

[7] For example, the definition of a "word" can be ambiguous (e.g. in languages which do not use spaces), and in the case of code there are no well-defined words. Meanwhile, storage size can vary by orders of magnitude depending on the choice of compression.

our models as log-normal distributions, and report our 95% confidence intervals (CIs) for each of them. The CIs for the latent variables are obtained by Monte Carlo simulations of the functional relationships that define those variables.

2.2.1. Indexed web

Common Crawl, a regularly updated open-source collection of scraped web data consisting of over 250 billion web pages (Common Crawl, 2024),[8] serves as the basis for most open web datasets, such as RefinedWeb, C4, and RedPajama. As a subset of the indexed web, Common Crawl's maximum size is inherently bounded by the size of the indexed web.[9]

To estimate the size of the indexed web, we use the size of Google's index as a proxy.[10] Applying the methodology proposed by van den Bosch et al. (2016), we estimate that Google's index contains approximately 250 billion web pages, with a 95% confidence interval ranging from 100 billion to 1200 billion web pages (see Appendix B).

Assuming that Common Crawl is a representative sample of the indexed web,[11] we can use it to estimate the average amount of plain text bytes per web page. This number has increased over time, from around 6100 bytes in 2013 to about 8200 bytes in 2021.[12] We estimate the average plain text bytes per web page to be 7000 [95%: 6100, 8200].[13]

Each token corresponds to 4 bytes of plain text [95%: 2, 5] (see Appendix E), so the raw stock of tokens on the indexed web in 2024, calculated according to Equation 1 is around 510 trillion [95%: 130T, 2100T].

Since 2013, the plain-text size of the average Common Crawl web page has been growing by between 2% and 4% each year. However, estimating the growth rate of the total number of web pages is more challenging due to conflicting evidence. The methodology employed by van den Bosch et al. (2016) suggests that the size of Google's index has remained relatively constant over the past decade, which is a counterintuitive result since new web pages are regularly created. Appendix B discusses alternative explanations for this apparent lack of growth in Google's index size.

Indexed Web Projected Growth

$$S_{IW}(y) = N_{IW} \times B_P \times T_B \times (1+g)^{y-y_0} \quad (1)$$

where $S_{IW}(y)$ is the estimate of the current stock of tokens in the indexed web in a given year y, N_{IW} is the number of unique web pages in the indexed web, B_P is the average number of bytes per web page, T_B is the average number of tokens per byte, and g is the estimated rate of growth of the total number of tokens.

To better estimate the growth rate of the indexed web, we consider several proxies: global IP traffic, link rot rates, and the growth in the number of internet users. Global IP traffic was increasing by 24% in 2016 (Cisco, 2017), which can be considered an upper bound on the growth rate of web pages, as the majority of traffic corresponds to consumption rather than creation of text data. Conversely, the number of internet users is growing by approximately 2-4% per year (Section 2.2.2), and estimates of the link rot rate range from 2% to 16% (Appendix B). For Google's index size to remain constant, the link rot rate must be offset by the creation of new web pages or links, suggesting possible growth rates of around 10%.

However, double-digit growth rates would imply that the average internet user is creating significantly more web pages over time, a trend that appears to be contradicted by some observations, such as the roughly constant rate of tweets per user on Twitter (GDELT, 2020) and similar observations for other platforms such as Wikipedia (Wikipedia). Given these considerations, we settle on a confidence interval between 0% and 10% a year.[14]

2.2.2. Internet population

We consider an alternative model of data stocks that explicitly accounts for the process that generates data. This model relies on the observation that much of the internet's text data is user-generated and stored on platforms such as social media, blogs, and forums. While AI-generated text is becoming more prevalent, we exclude it from this model and discuss it in Section 3. In principle, we can estimate the

[8] We use the term "web page" to refer to individual pages within a domain or website, for example a single article in Wikipedia.

[9] Although some web pages might be crawled by Common Crawl but not included in any search engine index, possibly due to being considered very low quality, we expect these web pages to be largely eliminated by quality filters in the data pipelines. Therefore, ignoring them should not significantly impact our results.

[10] Google is the most widely used search engine globally and receives a significant fraction of all web traffic (Similarweb, 2024). Consequently, we expect the size of Google's index to approximate the size of the indexed web within a factor of 2-5.

[11] This is the stated intention of the Common Crawl team, and since the crawling procedure is quite similar for Common Crawl and search indices, given the size of the archive it seems unlikely to have any significant bias.

[12] The total size of web pages is 10-20 times larger, since it also includes HTML code, scripts and other non-plain-text data.

[13] We use square brackets to denote 95% confidence intervals.

[14] A doubling of the growth rate from 10% to 20% over a 10-year period would result in an 0.4 OOM increase in the data stock. However, this is not large compared to the historical growth rate in data usage of 0.38 OOM per year. Therefore, our conclusions are not highly sensitive to variations in the growth rate within this range.

amount of public human-generated text data by considering the number of internet users and the average data produced per user, with growth in data generation primarily driven by the increasing number of internet users.

We model the increase in the number of internet users as coming from two contributors: (1) increases in the human population, and (2) increases in "internet penetration," i.e. the percentage of the population that uses the internet. For the former, we turn to standard projections by the United Nations (United Nations, 2022). Since internet penetration has broadly followed an S-curve from ∼0% in 1990 to 50% in 2016 to over 60% today (Ritchie & Roser, 2017), we model this using a sigmoid function, fitting it to the data in Ritchie & Roser (2017).

Finally, the amount of data generated per internet user varies across countries and over time due to differences in culture, demographics, socioeconomic factors, and online services. Quantifying these variations is complex and beyond the scope of this analysis, so we assume that the average data production rate per user remains constant to enable a tractable estimate.

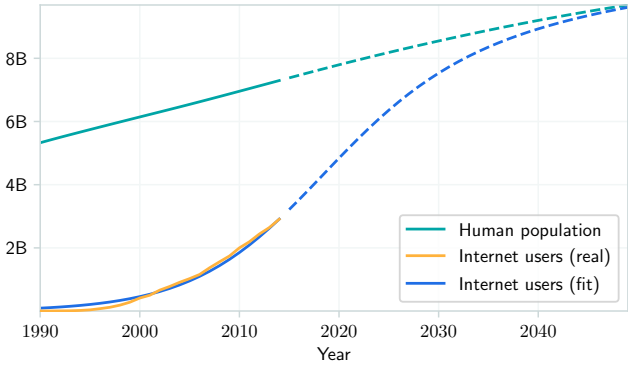

Figure 2. Historical and projected evolution of internet users. Historical data is from Ritchie & Roser (2020).

This model of the number of internet users closely matches the historical data (Figure 2). A more detailed explanation of this model can be found in Appendix A.

Based on reported user statistics for major online platforms (see Appendix C), we estimate the total volume of text data uploaded to the internet in 2024 was between 180T and 500T tokens. To project future data accumulation, we scale this initial 2024 estimate by the projected number of internet users in each subsequent year. This provides the estimated annual data contribution from the global online population. We then cumulatively sum these yearly contributions over time to model the total stock of internet text data. The final estimate is 3100T [95%: 1900T, 5200T] tokens. This estimate includes both data on the indexed web and the deep web, and therefore serves as an upper bound on the size of the indexed web.

2.3. Data quality and multi-epoch training

The preceding subsections outline the core basis of the model that we use in our analysis. However, before performing forecasts, we first need to account for a few additional considerations.

In particular, since our focus is on data constraints in the scaling of *language models*, the literal number of tokens in the training dataset may not be what matters for improving LLM performance. For example, differences in data quality (Li et al., 2023) and the number of training epochs (Muennighoff et al., 2023) can potentially have a substantial effect on final model performance. In this subsection, we analyze the significance of these factors and modify our model accordingly. Our adjustments for data quality and multi-epoch training are illustrated in Figure 3.

2.3.1. Data quality

One way in which only considering the measure of "number of tokens" is too simplistic is that not all public human text data is created equal. Intuitively, we would expect models that are trained primarily on books or Wikipedia to outperform models that are purely trained on YouTube comments. In this way, public human text data from books are "higher quality" than YouTube comments. Such intuitions are in fact supported by some empirical observations. For example, data processing techniques like deduplication (Lee et al., 2022) and data filtering (Gao, 2021) have been shown to improve model performance.

However, building in these effects into our model is non-trivial. For one, there is no standard accepted measure of data quality (Mitchell et al., 2023). Instead, we are forced to rely on a fairly vague working definition: A dataset is of higher quality than another if training on it leads to higher performance, at similar dataset sizes.

Recent findings show that with adequate filtering, data extracted from the web can outperform datasets constructed from human-curated sources (Penedo et al., 2023). In addition, Xie et al. (2023) found that in The Pile, which is a dataset consisting of web data and human-curated sources, increasing the proportion of web data up to 40-70% led to substantially higher performance. These empirical findings suggest that while much of internet public human text data is on average "lower quality" than human-curated sources, one can potentially make up for this through careful data processing.

Given these considerations, we can attempt to determine how much we need to adjust our previous model to account for data quality. We operationalize this in terms of how much "low quality" data is filtered to achieve optimal performance in practice. Penedo et al. (2023) create a 5T-token dataset that outperforms curated corpora by carefully filter-

ing and deduplicating raw data from Common Crawl. The filtering part of this process reduced the size of the web dataset by around 30%. Meanwhile, Marion et al. (2023) found that pruning around 50% of deduplicated data from a subset of Common Crawl using a perplexity measure led to optimal performance.[15] Based on these empirical results, we believe with 95% certainty that between 10% and 40% of deduplicated web data can be used for training without significantly compromising performance.

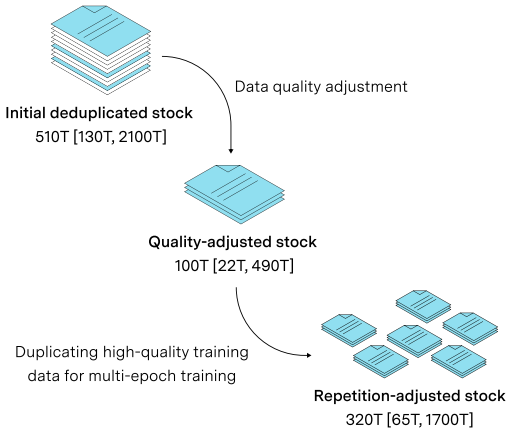

Figure 3. Illustration of the adjustments for quality and repetition and the adjusted stock sizes in number of tokens. First the lower-quality data is filtered out, and then the resulting dataset is duplicated for multi-epoch training.

2.3.2. Multiple epochs

Besides data quality, using the "number of tokens" as a measure does not account for the possibility of multi-epoch training. The degree to which stocks should be adjusted for multiple epochs depends on the effectiveness of training on the same data over multiple epochs, compared to training on new "unique" data.

Muennighoff et al. (2023) investigate this empirically, fitting a scaling law for the performance of a model trained for multiple epochs. Concretely, for a given model trained on multiple epochs, this law gives an estimate of the dataset size that would produce an equally capable model with just one epoch.. This is the "effective dataset size" of a multi-epoch training run. The authors estimate the maximum increase in the effective dataset size that can be gained from multiple epochs at between 3x and 15x, and we anchor to this estimate in adjusting our model. Because additional epochs yield diminishing returns, the upper extreme of 15x would require a very inefficient training procedure with a large number of epochs that does not correspond to common practices.[16] For this reason we reduce it to 5x.

Historical Dataset Size Growth Projection

$$D_H(y) = G_D^{y-y_0} D(y_0) \tag{2}$$

where D_H is the training dataset size, G_D is the factor growth per year, Y_0 is some base year, and Y is the year. Both G_D and $D(y_0)$ are lognormal distributions.

2.4. Projecting growth in dataset sizes

To project the future values of our second key variable, the training dataset size D, we begin by examining historical growth rates and extrapolating them forward.

To estimate historical growth, we use the database of notable machine learning models in Epoch (2022), a comprehensive database that contains annotations of over 300 machine learning models. We filter this data to include only large language models (LLMs) from papers published between 2010 and 2024, resulting in a subset of around 80 data points. We then perform a linear regression on the logarithm of the dataset size against time, as shown in Equation 2. This yields a median estimate of 0.38 orders of magnitude per year (OOM/y), or around 2.4x per year, with a bootstrapped 95% confidence interval of 0.27 to 0.48 OOM/y.

To project this trend forward, we first need to determine the size of the largest datasets used today, which are typically around 10T tokens.[17][18] Naively projecting the historical trend from this baseline suggests that systems could be trained on over one quadrillion tokens by the end of the decade (see Figure 4).

The historical growth rate in dataset sizes cannot continue indefinitely, even if the data stock was unlimited. In the past, the increasing scale of computing power has driven the

[15] First, the Common Crawl raw data is filtered using regular methods down to 20% of the original size, and then the remaining data is further pruned using several quality metrics. Figure 4 in Marion et al. (2023) shows that the best result is obtained when pruning 50% of the deduplicated dataset, for a final size of 10% of the original. Note that there appears to be no benefit from using the rest of the data, instead of repeating the best 10%.

[16] Typical numbers are between 1 and 4 epochs, for example see Taylor et al. (2022) and Touvron et al. (2023).

[17] Skywork-13B (Wei et al., 2023), XVERSE-65B (XVERSE Technology Inc., 2024), and PaLM 2 (Anil et al., 2023) were each estimated to be trained on roughly 4T tokens (Epoch, 2023). DBRX (Mosaic AI, 2024) reportedly used 12T tokens, and Llama 3 (Meta, 2024) used 15T tokens.

[18] Some of the largest recent models, like GPT-4 and Gemini Ultra, do not report the size of their datasets and so we have not included them in the analysis. However, estimates of their training compute are around 5e25 floating point operations (FLOP) (Epoch, 2023), which, assuming they are trained using Chinchilla scaling, would correspond to a dataset size of approximately 13T tokens. Therefore, we expect their datasets to be within the same order of magnitude as the biggest ones on our list.

demand for larger training datasets, consistent with neural scaling laws for dense transformers which suggest that training data size should scale roughly with the square root of training compute (Hoffmann et al., 2022; Dey et al., 2023; Fetterman et al., 2023).

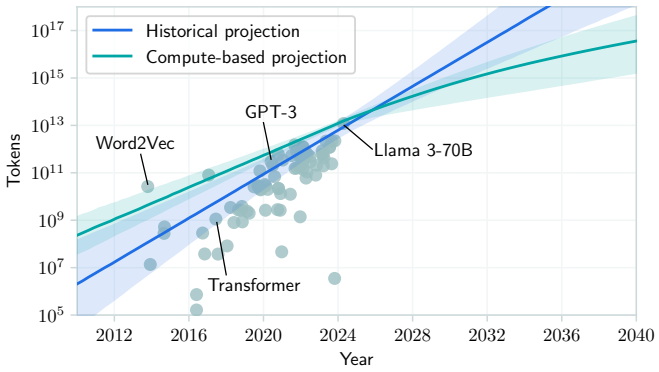

Figure 4. Projections of data usage. Two extrapolations of data usage, one from past trends and one from compute availability estimations plus scaling laws. The shaded areas denote a 90% CI for the extrapolated median. The dots are individual training runs.

However, the growth in compute is also subject to limits, and current fast rates may not be sustained indefintely. Technical limitations, such as the energy efficiency of computing devices (Ho et al., 2023) and limits on the electricity supply to data centers,[19] restrict the feasible amount of compute. Other factors like chip production capacity and economic constraints could slow down the rate at which computing power used in training can scale. Consequently, if the ability to scale computing power is constrained, it will likely lead to a deceleration in the historical trends of dataset size growth.

To introduce this constraint into our model, we need estimates of the maximum compute budget for training that will be available in the future. For this purpose, we use the results from Besiroglu et al. (2022), which performs such a projection based on estimated training compute growth rates in frontier machine learning systems between 2010 and 2022.[20] Following Hoffmann et al. (2022), we further assume that compute-optimality involves training on 20 tokens per parameter, per Equation 3.

[19]In the US, electric grids are already struggling to meet the growing electricity demands from data centers, which are critical infrastructure for AI (Zimmerman et al., 2023). Upgrading transmission capacity to deliver more power to these facilities involves lengthy planning and construction timelines, often spanning many years (ibid.)

[20]Note that this projection has a wide range of uncertainty and includes scenarios in which spending on compute grows orders of magnitude over current levels, up to 1% of GWP.

Compute-based Dataset Size Growth Projection

$$D_C(y) = \sqrt{\frac{20}{6} \cdot C(y)} \tag{3}$$

where $D_C(y)$ is the projected amount of data used in notable training runs and $C(y)$ is the probabilistic projection of largest compute spent on a training run, modeled following Besiroglu et al. (2022).
6 is the number of FLOP per parameter per token and 20 is the approximate number of training tokens per parameter according to Hoffmann et al. (2022).

As illustrated in Figure 4, the resulting model closely matches the historical trend and its projection until around 2030. It then slows down over time.

Our final projection of growth in dataset sizes is an equally-weighted mixture of both the historical and compute-based projections [21] (see Equation 4). It is illustrated in Figure 1.

Mixture Projection of Dataset Size Growth

$$F_{D(y)} = \frac{1}{2}\left(F_{D_H(y)} + F_{D_C(y)}\right) \tag{4}$$

where $D_H(y)$ is the historical projection of dataset sizes, $D_C(y)$ is the compute-based projection, and F_X is the cumulative distribution function of the random variable X.

2.5. When will the stock of public human text data be fully utilized?

Combining our projections of dataset size increases, and our estimate of the stock of data, we can estimate when the full stock will be used in a training run if past trends continue. Figure 5 shows the projected availability and usage of effective data. The intersection between these projections corresponds to public text data being exhausted. The median exhaustion year is 2028, and by 2032 exhaustion becomes very likely. At the point the data stock is fully utilized, models will be using around 5e28 FLOP during training.

An important assumption in our projections is that models are trained compute-optimally[22]. However, many developers might instead decide to "overtrain" models to achieve better efficiency during inference (Sardana & Frankle, 2023),

[21]The historical projection is simpler but seems to contradict reasonable assumptions about compute scaling. A mixture provides a good representation of our uncertainty.

[22]Compute-optimal training refers to selecting the number of parameters and dataset size of the model to produce the maximum possible capabilities for a given level of training compute (Hoffmann et al., 2022).

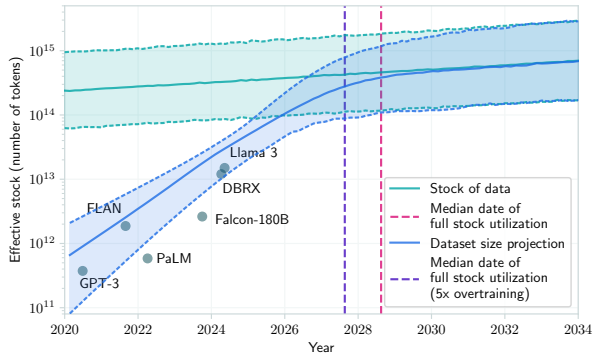

Figure 5. Projection of effective stock of human-generated public text and dataset sizes used to train notable LLMs. The intersection of the stock and dataset size projection lines indicates the median year (2028) in which the stock is expected to become fully utilized if current LLM development trends continue. At this point, models will be trained on dataset sizes approaching the total effective stock of text in the indexed web: around 4e14 tokens, corresponding to training compute of ∼5e28 FLOP for non-overtrained models.

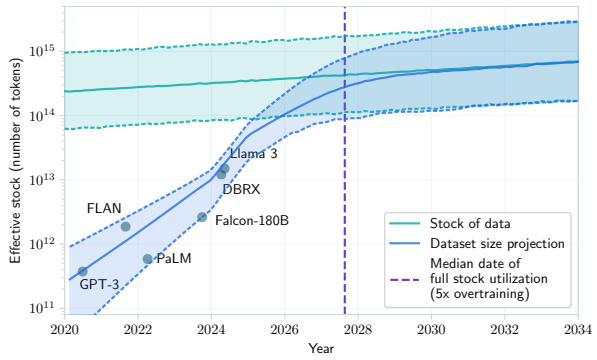

Figure 6. Compute-based data usage projection, assuming that frontier models will be overtrained by 5x starting from 2025. This policy results in the stock of data being fully used earlier than with a compute-optimal scaling policy.

which would require more data. The degree of overtraining that will be chosen by developers depends on a multitude of factors, in particular how many tokens will be generated during inference (Sardana & Frankle, 2023), and is hard to predict in advance. That said, based on our analysis in Appendix F, we consider overtraining by 5x to be a reasonable choice.[23] This would result in a data bottleneck one year earlier than our projections indicate, at a training compute level of ∼6e27 FLOP.[24]

[23]Overtraining by 5x means that the tokens/parameter ratio is 5x higher than that of a compute-optimal model, or equivalently that the model uses $\sqrt{5}$ times more data than a training-compute-optimal model trained using the same amount of compute.

[24]We still decide to focus on compute-optimal training for two reasons. First, we are more interested in frontier capabilities, and at any given compute budget the capabilities of models trained compute-optimally will be higher. Second, normally only smaller models are heavily overtrained. For example, Llama 3 8B is overtrained by close to 100x, while Llama 3 70B is only overtrained by 10x. Models that are closer to the compute frontier tend to be less overtrained due to the high cost of overtraining.

According to our projections, data could become a significant bottleneck for training LLMs this decade, particularly if LLMs continue to be intensively overtrained. This timeline allows for potentially substantial improvements in LLM performance, given the rapid progress in recent years (Ho et al., 2024; Sevilla et al., 2022). However, when considering the near 70-year history of AI, this timeframe is relatively short. While significant advancements can be made in the coming years, the impending data bottleneck presents an urgent challenge for the long-term progress of AI. For AI progress to continue into the 2030s, either new sources of data or less data-hungry techniques must be developed. The following sections of this paper will address some of these possibilities.

3. Beyond public human text data

While the core focus of this paper is on public human text data in particular, understanding the broader implications of our model's predictions requires considering ways in which the model might be wrong or incomplete. Crucially, although the model predicts that public human text data will be fully utilized at around the end of the decade, this does *not* necessarily imply that training data will bottleneck ML scaling at that time. In this section, we briefly survey possible ways of circumventing bottlenecks in public human text data.

For example, our model assumes no substantial change in the underlying process of increasing the public human text data stock. One naive way in which this assumption breaks is if significantly more humans are paid to generate more text. While this might be valuable at small scale for certain types of data, it is unlikely to be an economical way to generate an appreciable increase in text for general-purpose pre-training.[25]

Out of the remaining strategies for circumventing public human text data bottlenecks, we identify three broad categories of techniques that appear particularly promising. These are: a) using models themselves to generate more data, b) multimodality and transfer learning, which involves training language models on other existing datasets (e.g. from different domains), and c) using non-public data.

[25]Appreciably increasing the stock of text data could require hiring millions of people. 10 million people writing 40 words per minute for 8 hours per day would write 70T words in one year, which is the same order of magnitude as the stock of data in Common Crawl, at a cost of hundreds of billions of dollars in wages.

3.1. AI-generated data

OpenAI alone reportedly generates 100B words per day (Griffin, 2024). Within a year, this corresponds to around 36.5T words, not far from our estimates of the total number of high-quality words in Common Crawl. If outputs are accumulated across different models and across time, the growth in the stock of training data could expand dramatically in principle, assuming this approach works.

However, the evidence for the effectiveness of training on generated (synthetic) data is currently mixed. One challenge is that models might lose information about the original human data distribution, such that iteratively training on model outputs results in increasingly homogeneous and unrealistic outputs (Shumailov et al., 2023). More generally, repeatedly training on synthetic data can yield diminishing or even negative returns (Singh et al., 2023), and worse scaling behavior (Fan et al., 2023; Dohmatob et al., 2024). These challenges can be mitigated to some extent by using training data with greater diversity (Fan et al., 2023; OpenAI et al., 2019), or by training on a mixture of human-generated and synthetic data (Gunasekar et al., 2023; Shumailov et al., 2023; Gerstgrasser et al., 2024; Alemohammad et al., 2023).

On the other hand, training on synthetic data has shown much promise in domains where model outputs are relatively easy to verify, such as mathematics, programming, and games (Yang et al., 2023; Liu et al., 2023; Haluptzok et al., 2023).[26] For example, AlphaZero (Silver et al., 2017) was famously trained using self-play, and more recently AlphaGeometry (Trinh et al., 2024) was trained purely using synthetic data from attempts to solve geometry problems. What is less clear is whether the usefulness of synthetic data will generalize to domains where output verification is more challenging, such as natural language.[27]

We consider synthetic data to be one of the most promising avenues for circumventing data bottlenecks because of its potential to produce training data at an massive scale, its demonstrated success in certain domains, and the existence of potential strategies to mitigate the challenges associated with its use.

[26]Verification processes can be used as training signals which guide the data generation and improve performance (Zhang & Parkes, 2023; Huang et al., 2022).

[27]Despite this, researchers have attempted to train models on synthetic feedback, such as using model-generated critiques to prevent certain behaviors (Bai et al., 2022; Burns et al., 2023; Irving et al., 2018). These approaches highlight potential advantages of synthetic data, including avoiding difficulties in generating human feedback at scale (Burns et al., 2023; Khan et al., 2024; Saunders et al., 2022; Michael et al., 2023).

3.2. Multimodal and transfer learning

Another option is to go beyond *text* data, and train models on data from other domains or non-text modalities, like images. Appendix D includes some rough estimates of the stock of data for some of the most prominent modalities, concluding that current video and image stocks are not large enough to prevent a data bottleneck.

But there are other sources that can provide orders of magnitude more data of various types (e.g. financial market data, scientific databases, etc.). For illustration, (Stephens et al., 2015) forecasts growth rates of between 2-40 million terabytes of genomics data every year by 2025.

While it is not clear that leveraging data-rich domains for language modeling is always possible, there is already evidence that this is feasible in some specific cases. For instance, current frontier models like GPT-4V are trained on both image and text data (OpenAI, 2023; Pichai & Hassabis, 2023). Aghajanyan et al. (2023) study this question for several modalities of data and show that these modalities have some synergy with text, when training on an even mix of both. In general, better understanding the feasibility of transfer learning would require further research, such as scaling laws for transfer learning (Hernandez et al., 2021).

3.3. Using non-public data

While the indexed web is vast, its size is small relative to the *deep web*: the part of the web that is not accessible by search engines. The largest components of the deep web are closed content platforms like Facebook, Instagram or Twitter. While part of these platforms are indexed, the vast majority is not. Another large reservoir of non-public text data can be found in instant-messaging applications like WhatsApp or Facebook Messenger.

In Appendix C, we estimate that content platforms and instant messaging apps both contain on the order of one quadrillion tokens. Combining this with the similarly-sized upper estimate of the raw stock of text in the indexed web, the total stock could reach 3 quadrillion tokens. This increase would delay a data bottleneck by about a year and a half relative to using only data from the indexed web.

However, the non-public stock seems unlikely to be as useful as indicated by our estimate. First of all, training on this data would be a grave violation of the privacy of the users who submitted the data to platforms without expecting it to be used for training AI models and probably would face legal challenges. Second, the quality of social media content is probably substantially lower than that of web content. Finally, this data is fragmented across several closed platforms that are controlled by different actors, so it is unlikely that all of it can be used in a single training run.

3.4. Data efficiency techniques

According to Ho et al. (2024), training techniques and algorithms for LLMs have been improving at a rate of 0.4 OOM/y [95%: 0.1, 0.8], meaning that roughly 0.4 fewer OOMs of compute are needed each year to achieve the same levels of performance. This is partially due to more efficient data use. Similarly large gains in sample efficiency has been found for reinforcement learning (Dorner, 2021). Although we do not know precisely what fraction of LLM efficiency gains result from "doing more with less data," it is possible that improvements in data efficiency are occurring at a pace that could compensate for the exhaustion of data stocks.

3.5. Other techniques

Another possibility is learning from interactions with the real world, which might include LLMs training on the messages received from users or, if ML models become sophisticated enough to act autonomously, learning from sensory observations or from the results of real-world experiments. This form of learning will probably become necessary at some point if AI models are to surpass human knowledge about the real world.

One additional broad category of techniques is data selection, in which we include techniques like pruning (Marion et al., 2023), domain composition tuning (Xie et al., 2023), and curriculum learning (Campos, 2021). However, we do not find this class of techniques very promising since the gains tend to be modest.[28]

4. Discussion

In this paper, we examine the challenges and opportunities that lie ahead for scaling machine learning systems, particularly in light of the finite nature of public human text data. Our analysis reveals a critical juncture approaching by the end of this decade, where the current reliance on public human text data for training ML models may become unsustainable. Despite this looming bottleneck, we identify transfer learning and self-generated data as viable and promising pathways that could enable the continued growth and evolution of ML systems beyond the constraints of public human text data.

Our conclusions are thus twofold. On the one hand, we expect that the current paradigm based on public human text data will not be able to continue a decade from now. On the other hand, it is likely that alternative sources of data will likely be adopted before then, allowing ML systems to continue scaling.

While our arguments about alternative sources of data are mostly qualitative, a better understanding of data quality could make it possible to make quantitative estimates of the benefits of transfer learning and synthetic data. For example, scaling experiments for transfer learning could be used to quantify the proximity or synergy between different distributions (Hernandez et al., 2021; Aghajanyan et al., 2023) and identify new datasets which can effectively expand the stock of data.

This paper does not explore certain considerations that might be relevant for understanding the future role of data. Firstly, the choice of data should depend on the desired skills or capabilities of the model. Identifying economically or scientifically valuable skills and the datasets needed to teach them could reveal critical data gaps. Secondly, future ML breakthroughs, such as systems capable of autonomous real-world exploration and experimentation, might change the dominant source of information for learning.

5. Conclusion

We have projected the growth trends in both the training dataset sizes used for state-of-the-art language models and the total stock of available human-generated public text data. Our analysis suggests that, if rapid growth in dataset sizes continues, models will utilize the full supply of public human text data at some point between 2026 and 2032, or one or two years earlier if frontier models are overtrained. At this point, the availability of public human text data may become a limiting factor in further scaling of language models.

However, after accounting for steady improvements in data efficiency and the promise of techniques like transfer learning and synthetic data generation, it is likely that we will be able to overcome this bottleneck in the availability of public human text data.

It is important to acknowledge the inherent uncertainty in making long-term projections, especially considering the rapid pace of advancements in the field of AI. Our results highlight the need for further research to quantify data efficiency growth rates and the potential performance gains from emerging methods. Additionally, future work should explore the feasibility and effectiveness of transfer learning from diverse data domains and the impact of synthetic data generation on model performance, among other things.

[28] See, for example, Marion et al. (2023), where pruning provides less of a benefit than scaling the dataset by 3x, or Tirumala et al. (2023), where the benefit of selection is similar to that of scaling datasets by 20%. In general, the benefit of these techniques is likely limited by the fraction of performance-degrading datapoints in the dataset.

Acknowledgments

We thank the ICML reviewers, Nuño Sempere, Eli Lifland, Ege Erdil, Matthew Barnett and Joshua You for their thoughtful comments and contributions to this paper.

Impact Statement

The practice of scraping data from the web and using it for large-scale training of AI systems raises important issues regarding fairness and justice. In particular, there are strong arguments in favor of compensating the creators of the data used to train these systems. While AI has the potential to greatly increase productivity and overall welfare, it is important to factor in these justice-related considerations to ensure that the benefits are distributed equitably.

Our work suggests that data from social media platforms and messaging apps could serve as a significant and valuable resource for training AI systems. However, using this type of data for training raises serious privacy and security concerns. Without proper safeguards in place, sensitive personal information from these platforms could be exposed to users of the AI systems .The risks associated with using non-indexed platform data for training may be substantial enough to outweigh the potential benefits gained from using this data.

References

Abnar, S., Dehghani, M., Neyshabur, B., and Sedghi, H. Exploring the limits of large scale pre-training, 2021. URL https://arxiv.org/abs/2110.02095.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., and Anadkat, S. et al. GPT-4 Technical Report. Technical report, OpenAI, 2023. URL https://arxiv.org/abs/2303.08774.

Aghajanyan, A., Yu, L., Conneau, A., Hsu, W.-N., Hambardzumyan, K., Zhang, S., Roller, S., Goyal, N., Levy, O., and Zettlemoyer, L. Scaling laws for generative mixed-modal language models, 2023. URL https://arxiv.org/abs/2301.03728.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad, 2023.

Alpert, J. and Hajaj, N. We knew the web was big... https://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html, 2008. [Accessed 22-04-2024].

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., and Chu, E. et al. PaLM 2 Technical Report, 2023. URL https://arxiv.org/abs/2305.10403.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., and McKinnon, C. et al. Constitutional AI: Harmlessness from AI Feedback, 2022.

Baumgartner, J., Zannettou, S., Keegan, B., Squire, M., and Blackburn, J. The pushshift reddit dataset, 2020. URL https://arxiv.org/abs/2001.08435.

Besiroglu, T., Heim, L., and Sevilla, J. Projecting compute trends in machine learning, 2022. URL https://epochai.org/blog/projecting-compute-trends. Accessed: 2024-01-29.

Besiroglu, T., Erdil, E., Barnett, M., and You, J. Chinchilla scaling: A replication attempt, 2024.

Bochkarev, V. V., Shevlyakova, A. V., and Solovyev, V. D. Average word length dynamics as indicator of cultural changes in society. *ArXiv*, abs/1208.6109, 2012. URL https://arxiv.org/abs/1208.6109.

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

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., Sutskever, I., and Wu, J. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision, 2023.

Campos, D. Curriculum learning for language modeling. *CoRR*, abs/2108.02170, 2021. URL https://arxiv.org/abs/2108.02170.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., and Gehrmann, S. et al. PaLM: Scaling Language Modeling with Pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Cisco. Cisco Visual Networking Index: Forecast and Methodology, 2016–2021, 2017.

Coffman, K. and Odlyzko, A. The size and growth rate of the internet, 1998. URL https://firstmonday.

org/ojs/index.php/fm/article/download/620/541?inline=1.

Common Crawl. Common crawl, 2024. URL https://commoncrawl.org/.

Cottier, B. Trends in the dollar training cost of machine learning systems, 2023. URL https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems. Accessed: 2024-02-01.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., Hutter, M., and Veness, J. Language modeling is compression, 2023. URL https://arxiv.org/abs/2309.10668.

Dey, N., Gosal, G., Zhiming, Chen, Khachane, H., Marshall, W., Pathria, R., Tom, M., and Hestness, J. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster, 2023. URL https://arxiv.org/abs/2304.03208.

Dohmatob, E., Feng, Y., Yang, P., Charton, F., and Kempe, J. A tale of tails: Model collapse as a change of scaling laws, 2024.

Domo. Data Never Sleeps 10.0. https://www.domo.com/data-never-sleeps, 2022. [Accessed 12-04-2024].

Dorner, F. E. Measuring progress in deep reinforcement learning sample efficiency, 2021. URL https://arxiv.org/abs/2102.04881.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Du, N., Huang, Y., Dai, A. M., Tong, S., Lepikhin, D., Xu, Y., Krikun, M., Zhou, Y., Yu, A. W., Firat, O., Zoph, B., Fedus, L., Bosma, M., Zhou, Z., Wang, T., Wang, Y. E., Webster, K., Pellat, M., Robinson, K., Meier-Hellstern, K., Duke, T., Dixon, L., Zhang, K., Le, Q. V., Wu, Y., Chen, Z., and Cui, C. Glam: Efficient scaling of language models with mixture-of-experts, 2021. URL https://arxiv.org/abs/2112.06905.

Epoch. Parameter, compute and data trends in machine learning, 2022. URL https://epochai.org/data/epochdb/visualization. Accessed: 2024-01-29.

Epoch. Key trends and figures in machine learning, 2023. URL https://epochai.org/trends. Accessed: 2024-01-27.

Erdil, E. and Besiroglu, T. Algorithmic progress in computer vision, 2023. URL https://arxiv.org/abs/2212.05153.

Fan, L., Chen, K., Krishnan, D., Katabi, D., Isola, P., and Tian, Y. Scaling laws of synthetic images for model training ... for now, 2023.

Fetterman, A. J., Kitanidis, E., Albrecht, J., Polizzi, Z., Fogelman, B., Knutins, M., Wróblewski, B., Simon, J. B., and Qiu, K. Tune as you scale: Hyperparameter optimization for compute efficient training, 2023. URL https://arxiv.org/abs/2306.08055.

Friel, N., McKeone, J. P., Oates, C. J., and Pettitt, A. N. Investigation of the widely applicable bayesian information criterion, 2016. URL https://arxiv.org/abs/1501.05447.

Gao, L. An empirical exploration in quality filtering of text data, 2021. URL https://arxiv.org/abs/2109.00698.

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., Phang, J., He, H., Thite, A., Nabeshima, N., Presser, S., and Leahy, C. The pile: An 800gb dataset of diverse text for language modeling, 2021. URL https://arxiv.org/abs/2101.00027.

GDELT. Visualizing Twitter's Evolution 2012-2020 And How Tweeting Is Changing In The COVID-19 Era. https://blog.gdeltproject.org/visualizing-twitters-evolution-2012-2020-and-how-tweeting-is-changing-in-the-covid-19-era/, 2020. [Accessed 12-04-2024].

Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Sleight, H., Hughes, J., Korbak, T., Agrawal, R., Pai, D., Gromov, A., Roberts, D. A., Yang, D., Donoho, D. L., and Koyejo, S. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data, 2024.

Griffin, A. Chatgpt creators openai are generating 100 billion words per day, ceo says, 2024. URL https://www.independent.co.uk/tech/chatgpt-openai-words-sam-altman-b2494900.html.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Giorno, A. D., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks are all you need, 2023.

Haluptzok, P., Bowers, M., and Kalai, A. T. Language models can teach themselves to program better, 2023. URL https://arxiv.org/abs/2207.14502.

Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., Hallacy, C., Mann, B., Radford, A., Ramesh, A., Ryder, N., Ziegler, D. M., Schulman, J., Amodei, D., and McCandlish, S. Scaling laws for autoregressive generative modeling, 2020. URL https://arxiv.org/abs/2010.14701.

Hernandez, D., Kaplan, J., Henighan, T., and McCandlish, S. Scaling laws for transfer, 2021. URL https://arxiv.org/abs/2102.01293.

Ho, A., Erdil, E., and Besiroglu, T. Limits to the energy efficiency of cmos microprocessors. In *2023 IEEE International Conference on Rebooting Computing (ICRC)*, pp. 1–10. IEEE, 2023.

Ho, A., Besiroglu, T., Erdil, E., Owen, D., Rahman, R., Guo, Z. C., Atkinson, D., Thompson, N., and Sevilla, J. Algorithmic progress in language models. 2024. URL https://api.semanticscholar.org/CorpusID:268358466.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. v. d., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Howell, S. and Burtis, A. T. The continued problem of url decay: an updated analysis of health care management journal citations. *Journal of the Medical Library Association : JMLA*, 110:463 – 470, 2022. URL https://api.semanticscholar.org/CorpusID:257746935.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., and Han, J. Large language models can self-improve, 2022. URL https://arxiv.org/abs/2210.11610.

Institute, N. H. G. R. Genomic data science (fact sheet), 2024. URL https://www.genome.gov/about-genomics/fact-sheets/Genomic-Data-Science.

Irving, G., Christiano, P., and Amodei, D. Ai safety via debate, 2018.

jerryspan. GitHub - jerryspan/FacebookR: Facebook Post Reactions dataset. https://github.com/jerryspan/FacebookR/, 2017. [Accessed 12-04-2024].

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Kemp, S. Digital 2023: Global Overview Report — DataReportal. https://datareportal.com/reports/digital-2023-global-overview-report, 2023a. [Accessed 16-04-2024].

Kemp, S. Facebook Messenger Users, Stats, Data, Trends, and More — DataReportal. https://datareportal.com/essential-facebook-messenger-stats, 2023b. [Accessed 15-04-2024].

Kemp, S. Facebook Users, Stats, Data, Trends, and More — DataReportal. https://datareportal.com/essential-facebook-stats, 2023c. [Accessed 12-04-2024].

Khan, A., Hughes, J., Valentine, D., Ruis, L., Sachan, K., Radhakrishnan, A., Grefenstette, E., Bowman, S. R., Rocktäschel, T., and Perez, E. Debating with more persuasive llms leads to more truthful answers, 2024.

Krebs, F., Lubascher, B., Moers, T., Schaap, P., and Spanakis, G. Social emotion mining techniques for facebook posts reaction prediction, 2017. URL https://arxiv.org/abs/1712.03249.

Lee, E. 2021 worldwide image capture forecast: 2020 – 2025, 2021. URL https://riseaboveresearch.com/rar-reports/2021-worldwide-image-capture-forecast-2020-2025/.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better, 2022. URL https://arxiv.org/abs/2107.06499.

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., and Lee, Y. T. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023.

Liu, B., Bubeck, S., Eldan, R., Kulkarni, J., Li, Y., Nguyen, A., Ward, R., and Zhang, Y. Tinygsm: achieving ¿80

Loan, F. A. and Shah, U. Y. The decay and persistence of web references. *Digit. Libr. Perspect.*, 36:157–166, 2020. URL https://api.semanticscholar.org/CorpusID:219404788.

Loan, F. A., Khan, A. M., Andrabi, S. A. A., Sozia, S. R., and Parray, U. Y. Giving life to dead: role of WayBack Machine in recovery of dead URLs. *Data Technologies and Applications*, 2023. URL https://api.semanticscholar.org/CorpusID:259585485.

Lyman, P. and Varian, H. R. How much information, 2003. URL https://groups.ischool.berkeley.edu/archive/how-much-info-2003/.

Mahoney, M. Large Text Compression Benchmark, 2006/2024. URL http://www.mattmahoney.net/dc/text.html. [Accessed 29-01-2024].

Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. When less is more: Investigating data pruning for pretraining llms at scale, 2023. URL https://arxiv.org/abs/2309.04564.

Meta. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/, 2024. [Accessed 24-04-2024].

Michael, J., Mahdi, S., Rein, D., Petty, J., Dirani, J., Padmakumar, V., and Bowman, S. R. Debate helps supervise unreliable experts, 2023.

Mitchell, M., Luccioni, A. S., Lambert, N., Gerchick, M., McMillan-Major, A., Ozoani, E., Rajani, N., Thrush, T., Jernite, Y., and Kiela, D. Measuring data, 2023. URL https://arxiv.org/abs/2212.05129.

Mosaic AI. Introducing DBRX: A New State-of-the-Art Open LLM — Databricks — databricks.com. https://www.databricks.com/blog/introducing-dbrx-new-state-art-open-llm, 2024. [Accessed 12-04-2024].

Mosseri, A. and Chudnovsky, S. Say :wave: to Messenger: Introducing New Messaging Features for Instagram. https://about.fb.com/news/2020/09/new-messaging-features-for-instagram/, 2020. [Accessed 15-04-2024].

Muennighoff, N., Rush, A. M., Barak, B., Scao, T. L., Piktus, A., Tazi, N., Pyysalo, S., Wolf, T., and Raffel, C. Scaling data-constrained language models, 2023. URL https://arxiv.org/abs/2305.16264.

Murray H., B. and Moore, A. Sizing the internet. Technical report, Cyveillance, 7 2000. URL http://www.cs.toronto.edu/~leehyun/papers/Sizing_the_Internet.pdf.

National Electrical Manufacturers Association (NEMA). Comments on innovative advanced transformers. RFI Response DE-FOA-0003021, U.S. Department of Energy, Washington, DC, May 2023. URL https://www.nema.org/docs/default-source/advocacy-document-library/nema-comments-on-innovative-advanced-transformers-rfi-may-5-2023.pdf?sfvrsn=e15ffc3c_3. Response to the Department of Energy's request for information on Innovative Advanced Transformers.

Nguyen, T., Ilharco, G., Wortsman, M., Oh, S., and Schmidt, L. Quality not quantity: On the interaction between dataset design and robustness of clip, 2022. URL https://arxiv.org/abs/2208.05516.

Nostalgebraist. chinchilla's wild implications, 2022. URL https://www.lesswrong.com/posts/6Fpvch8RR29qLEWNH/chinchilla-s-wild-implications.

OBDILCI. Indicators for the Presence of Languages in the Internet. https://www.obdilci.org/projects/main/, 2024. [Accessed 16-04-2024].

Odlyzko, A. The growth rate and the nature of internet traffic. *Transactions on Internet Research*, pp. 39–42, 2016.

OECD. Gross domestic spending on R&D (indicator), 2022. URL https://data.oecd.org/rd/gross-domestic-spending-on-r-d.htm.

OpenAI. Gpt-4v(ision) system card. Technical report, OpenAI, 2023. Accessed: 2024-02-01.

OpenAI. tiktoken, 2024. URL https://github.com/openai/tiktoken.

OpenAI and Pilipiszyn, A. Gpt-3 powers the next generation of apps, 2021. URL https://openai.com/blog/gpt-3-apps/.

OpenAI, Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., Schneider, J., Tezak, N., Tworek, J., Welinder, P., Weng, L., Yuan, Q., Zaremba, W., and Zhang, L. Solving rubik's cube with a robot hand, 2019.

Ott, D. E. Reference hygiene and death on the internet–decay, rot, half-life, deterioration, and corruption. *JSLS: Journal of the Society of Laparoscopic & Robotic Surgeons*, 26(1), 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L. E., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. J. Training language models to follow instructions with human feedback. *ArXiv*, abs/2203.02155, 2022. URL https://arxiv.org/abs/2203.02155.

Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Cappelli, A., Alobeidli, H., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only, 2023. URL https://arxiv.org/abs/2306.01116.

Pichai, S. and Hassabis, D. Introducing gemini: our largest and most capable ai model, 2023. URL https://blog.google/technology/ai/google-gemini-ai/.

Radicati. Email statistics report, 2020-2024. https://www.radicati.com/wp/wp-content/uploads/2020/01/Email_Statistics_Report,_2020-2024_Executive_Summary.pdf. [Accessed 03-05-2024].

Raffel, C., Shazeer, N. M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. URL https://arxiv.org/abs/1910.10683.

Real Time Statistics Project. Internet Live Stats. URL https://www.internetlivestats.com/. [Accessed 02-02-2024].

Reinsel, D., Gantz, J., and Rydning, J. The digitization of the world from edge to core. Technical report, International Data Corporation, 11 2018. URL https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

Ritchie, H. and Roser, M. Technology adoption. *Our World in Data*, 2017. URL https://ourworldindata.org/technology-adoption.

Ritchie, H. and Roser, M. Number of people using the internet. *Our World in Data*, 2020. URL https://ourworldindata.org/grapher/number-of-internet-users.

Roose, K. and Newton, C. Dario amodei on the paradoxes of a.i. safety and netflix's 'deep fake love'. Podcast Transcript, July 2023. URL https://www.nytimes.com/section/technology. Available at The New York Times.

Rosenfeld, A., Sina, S., Sarne, D., Avidov, O., and Kraus, S. WhatsApp usage patterns and prediction of demographic characteristics without access to message content. *Demographic Research*, 39(22):647–670, 2018. doi: 10.4054/DemRes.2018.39.22. URL https://www.demographic-research.org/volumes/vol39/22/.

Sardana, N. and Frankle, J. Beyond chinchilla-optimal: Accounting for inference in language model scaling laws, 2023. URL https://arxiv.org/abs/2401.00448.

Satyanarayana, D. and Damodar, P. Web citation analysis on journal of travel research: A study. *Library Progress (International)*, 42(2):412–420, 2022.

Saunders, W., Yeh, C., Wu, J., Bills, S., Ouyang, L., Ward, J., and Leike, J. Self-critiquing models for assisting human evaluators, 2022.

Scao, T. L., Wang, T., Hesslow, D., Saulnier, L., Bekman, S., Bari, M. S., Biderman, S., Elsahar, H., Muennighoff, N., Phang, J., Press, O., Raffel, C., Sanh, V., Shen, S., Sutawika, L., Tae, J., Yong, Z. X., Launay, J., and Beltagy, I. What language model to train if you have one million gpu hours?, 2022. URL https://arxiv.org/abs/2210.15424.

Schuhmann, C. and Sun, Q. Strategic game datasets for enhancing ai planning: An invitation for collaborative research. *LAION Projects Team Blog*, Oct 2023. URL https://laion.ai/blog/strategic-game-dataset/. Accessed: [insert date here].

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., and Villalobos, P. Compute trends across three eras of machine learning. In *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2022. doi: 10.1109/ijcnn55064.2022.9891914. URL http://dx.doi.org/10.1109/IJCNN55064.2022.9891914.

Shepard, C. Google's index size revealed: 400 billion docs (& changing), 2024. URL https://zyppy.com/seo/google-index-size/.

Shilov, A. Nvidia to reportedly triple output of compute gpus in 2024: Up to 2 million h100s, Aug 2023. URL https://www.tomshardware.com/news/nvidia-to-reportedly-triple-output-of-compute-gpus-in-2024-up-to-2-million-h100s.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget, 2023. URL https://arxiv.org/abs/2305.17493.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL https://arxiv.org/abs/1712.01815.

Similarweb. Top Websites. https://www.similarweb.com/top-websites/, 2024. [Accessed 16-04-2024].

Singh, A., Co-Reyes, J. D., Agarwal, R., Anand, A., Patil, P., Garcia, X., Liu, P. J., Harrison, J., Lee, J., Xu, K., Parisi, A., Kumar, A., Alemi, A., Rizkowsky, A., Nova, A., Adlam, B., Bohnet, B., Elsayed, G., Sedghi, H., Mordatch, I., Simpson, I., Gur, I., Snoek, J., Pennington, J.,

Hron, J., Kenealy, K., Swersky, K., Mahajan, K., Culp, L., Xiao, L., Bileschi, M. L., Constant, N., Novak, R., Liu, R., Warkentin, T., Qian, Y., Bansal, Y., Dyer, E., Neyshabur, B., Sohl-Dickstein, J., and Fiedel, N. Beyond human data: Scaling self-training for problem-solving with language models, 2023.

Singh, M. WhatsApp is now delivering roughly 100 billion messages a day. https://techcrunch.com/2020/10/29/whatsapp-is-now-delivering-roughly-100-billion-messages-a-day, 2020. [Accessed 15-04-2024].

Size, W. W. W. World wide web size, 2024. URL https://www.worldwidewebsize.com/.

Stephens, Z. D., Lee, S., Faghri, F., Campbell, R. H., Zhai, C., Efron, M., Iyer, R. K., Schatz, M. C., Sinha, S., and Robinson, G. E. Big data: Astronomical or genomical? *PLoS Biology*, 13, 2015. URL https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1002195.

Taniguchi, M., Ueda, Y., Taniguchi, T., and Ohkuma, T. A large-scale corpus of E-mail conversations with standard and two-level dialogue act annotations. In Scott, D., Bel, N., and Zong, C. (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4969–4980, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.436. URL https://aclanthology.org/2020.coling-main.436.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science, 2022. URL https://arxiv.org/abs/2211.09085.

Thorgren, E., Mohammadinodooshan, A., and Carlsson, N. Temporal dynamics of user engagement on instagram: A comparative analysis of album, photo, and video interactions. URL https://api.semanticscholar.org/CorpusID:268241757.

Tirumala, K., Simig, D., Aghajanyan, A., and Morcos, A. S. D4: Improving llm pretraining via document de-duplication and diversification, 2023. URL https://arxiv.org/abs/2308.12284.

Together.ai. Redpajama-data-v2: An open dataset with 30 trillion tokens for training large language models, 2023. URL https://www.together.ai/blog/redpajama-data-v2.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., and Bhosale S. et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023. URL https://arxiv.org/abs/2307.09288.

Trinh, T. H., Wu, Y., Le, Q. V., He, H., and Luong, T. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482, January 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06747-5. URL http://dx.doi.org/10.1038/s41586-023-06747-5.

United Nations. World population prospects 2022, online edition, 2022. URL https://population.un.org/wpp/.

van den Bosch, A., Bogers, T., and de Kunder, M. Estimating search engine index size variability: A 9-year longitudinal study. *Scientometrics*, 107(2):839–856, Feb 2016. doi: 10.1007/s11192-016-1863-z.

Villalobos, P. and Ho, A. Trends in training dataset sizes. https://epochai.org/blog/trends-in-training-dataset-sizes, 2022. Accessed: 2022-09-27.

Wei, T., Zhao, L., Zhang, L., Zhu, B., Wang, L., Yang, H., Li, B., Cheng, C., Lü, W., Hu, R., Li, C., Yang, L., Luo, X., Wu, X., Liu, L., Cheng, W., Cheng, P., Zhang, J., Zhang, X., Lin, L., Wang, X., Ma, Y., Dong, C., Sun, Y., Chen, Y., Peng, Y., Liang, X., Yan, S., Fang, H., and Zhou, Y. Skywork: A more open bilingual foundation model, 2023. URL https://arxiv.org/abs/2310.19341.

White, K. Publications output: U.s. trends and international comparisons, 2019. URL https://ncses.nsf.gov/pubs/nsb20206/.

Wikipedia. Wikipedia:Size of Wikipedia. https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia. [Accessed 22-04-2024].

Xie, S. M., Pham, H., Dong, X., Du, N., Liu, H., Lu, Y., Liang, P., Le, Q. V., Ma, T., and Yu, A. W. Doremi: Optimizing data mixtures speeds up language model pretraining, 2023. URL https://arxiv.org/abs/2305.10429.

Xue, F., Fu, Y., Zhou, W., Zheng, Z., and You, Y. To repeat or not to repeat: Insights from scaling llm under token-crisis, 2023. URL https://arxiv.org/abs/2305.13230.

XVERSE Technology Inc. Xverse-65b: A multilingual large language model, 2024. URL https://github.com/xverse-ai/XVERSE-65B. Apache-2.0 License.

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Yang, K., Swope, A. M., Gu, A., Chalamala, R., Song, P., Yu, S., Godil, S., Prenger, R., and Anandkumar, A. Leandojo: Theorem proving with retrieval-augmented language models, 2023.

Ye, W., Liu, S., Kurutach, T., Abbeel, P., and Gao, Y. Mastering atari games with limited data, 2021. URL https://arxiv.org/abs/2111.00210.

Yole Développement. Status of the CMOS Image Sensor Industry 2021. https://medias.yolegroup.com/uploads/2021/08/YINTR21167-Status-of-the-CMOS-Image-Sensor-Industry-2021-Sample.pdf, 2021. [Accessed 17-04-2024].

YouTube. YouTube for Press. URL https://blog.youtube/press/. [Accessed 02-02-2024].

Zellner, A. Optimal information processing and bayes's theorem. *The American Statistician*, 42(4):278–280, 1988. ISSN 00031305. URL http://www.jstor.org/stable/2685143.

Zhang, H. and Parkes, D. C. Chain-of-thought reasoning is a policy improvement operator, 2023. URL https://arxiv.org/abs/2309.08589.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models, 2023. URL https://arxiv.org/abs/2303.18223.

Zimmerman, Z., Goggin, M., and Gramlich, R. Ready-to-go transmission projects 2023. Technical report, Grid Strategies, September 2023. URL https://cleanenergygrid.org/wp-content/uploads/2023/09/ACEG_Transmission-Projects-Ready-To-Go_September-2023.pdf. Support from Americans for a Clean Energy Grid.

Zittrain, J., Bowers, J., and Stanton, C. The paper of record meets an ephemeral web: An examination of linkrot and content drift within the new york times. *SSRN Electronic Journal*, 2021. URL https://api.semanticscholar.org/CorpusID:236595511.

A. Theoretical growth model of the web

We explain in more detail our theoretical model of data accumulation rates developed in Section 2.2.2 and check it on Reddit submission data. The model is explained in Equation 5.

A purely exponential model cannot reproduce the decrease in the growth rate of Reddit submissions over time, while a purely sigmoidal model plateaus at zero growth. The exponential times sigmoid model is able to better capture the deceleration in submission size growth (see Figure 7).

In our actual model, the slowdown in population growth (which becomes subexponential) leads to additional deceleration, but the time period covered by the Reddit submission dataset seems too short for slowing population growth to be noticeable in the data.

Projection Based on the Number of Internet Users

$$S_{IU}(y) = D_{y_0} \int_{1950}^{y} \frac{H(x)\sigma\left((x - s_0) \times 0.15\right)}{H(y_0)\sigma\left((y_0 - s_0) \times 0.15\right)} dx \tag{5}$$

where D_{Y_0} is the amount of data produced in some reference year Y_0, $H(Y)$ is the projected human population in a certain year, and the sigmoid σ models internet penetration, which is approximately 0% in 1950 (this why we choose it as the initial point for the integral) and 50% in $s_0 = 2016$. 0.15 is a fitted scale parameter. The integral represents the total number of person-years of internet use, normalized by the internet use in the reference year.

B. Estimating the size of the indexed web

The "indexed web" comprises those web pages that are included in the indices of search engines. In particular, since Google is the most popular search engine worldwide, we tried to estimate the number of web pages in Google's index.

We replicate the methodology of van den Bosch et al. (2016). We calculate the frequency of words in a large corpus of clean web documents: the RefinedWeb dataset (Penedo et al., 2023). Then we select a set of words at logarithmically equidistant intervals of frequency, called "pivot words" in van den Bosch et al. (2016). Using the number of results that Google reports when searching each of the pivot words, we can extrapolate the total size of Google's index, assuming that the frequencies of the words are similar in our corpus and in Google's index.

Each pivot word provides a noisy estimate of the total size, so we take the average to arrive at a more robust estimate.

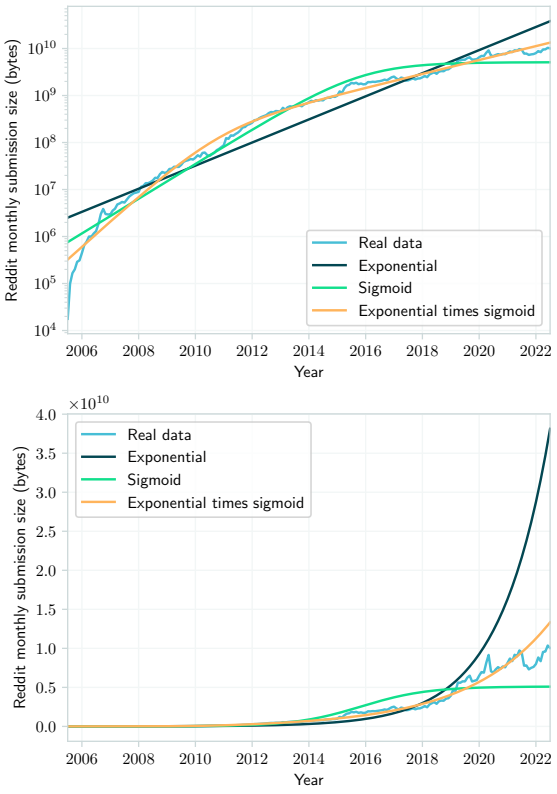

Figure 7. Monthly user submissions to Reddit, in linear scale(down) and log scale(up). While the three functions appear to fit the data reasonably well in the log scale, the linear plot shows that the sigmoid times exponential function predicts much better the recent years.

Using 100 pivot words, the distribution of estimated sizes is approximately log-normal, with a mean of 330B web pages, a median of 250B web pages, and a 95% CI between 100B and 1200B (see 8). This is about 4 times more than Common Crawl, which only has 75B unique urls.

Our results are substantially higher than those obtained by van den Bosch et al. (2016). This is mostly because we retrieved the number of Google results for each word using the Google Custom Search JSON API. However, van den Bosch et al. (2016) used the numbers shown in Google's web interface, which are around half as big for the same search terms.

We evaluated whether a change in the relative frequency of words in the web over time might lead to inaccuracies when the same frequencies are used to estimate the size of the index across several years. To do this, we computed word frequencies for webs sampled in different years between 2013 and 2021. The difference in the resulting estimates was smaller than 10%, so we conclude that this is unlikely to add significant noise to the estimate.

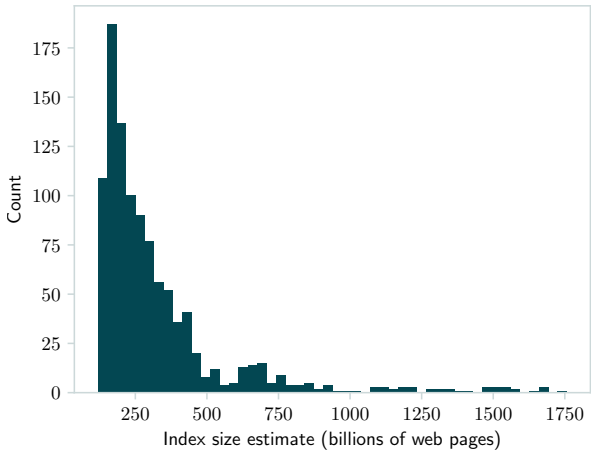

Figure 8. Histogram of the estimates of the size of Google's index from each pivot word.

B.1. Distribution across languages

The exact distribution of languages in the web is hard to estimate. In Common Crawl, around 45% of webpages is in English. This is broadly consistent with the 58.8% reported by the "Digital 2023: Global Overview Report" (Kemp, 2023a). However, the Observatory of Linguistic and Cultural Diversity in the Internet reports that in 2023 around 20% of the content of the web was in English (OBDILCI, 2024).

If Common Crawl does not constitute a representative sample of the web, then our previous method for estimating the size of the web might be biased. However, a reduction in the English share from 45% to 20% would only increase the estimate of the total size by a factor of 2.25, which is not enough to significantly change our conclusions.

B.2. Growth in the size of the indexed web

As documented in van den Bosch et al. (2016), the size of Google's index has varied significantly over the past decade. However, this variation does not seem to follow any monotonic temporal pattern, and instead consists of seemingly random movement around a stable mean. The fact that the temporal mean of the index size has not increased over time is surprising: the number of internet users grows by about 1.8% a year (Kemp, 2023a), the number of pages in Wikipedia grows by 2.6% per year (Wikipedia), and the global size of IP traffic grew by 24% per year in 2016 (Cisco, 2017). Given these facts, it seems unlikely that the overall number of web pages has not grown significantly.

We propose several hypotheses. One is based on link rot, the phenomenon of web links becoming inaccessible over time due to deletion, failure, restructuring of web sites, or other reasons. There have been several estimates of the rate of

link rot over time, but most claim that between 2% and 16% of links break during a year (Howell & Burtis, 2022; Loan & Shah, 2020; Loan et al., 2023; Ott, 2022; Satyanarayana & Damodar, 2022; Zittrain et al., 2021). Some of the broken links will be replaced by new links to the same content (think of a change in domain name, for example, in which the new domain will point to the same documents as the old one), but part of that content will be irreversibly lost. If the rate of growth of the web is similar to the rate of link rot, the two effects might partially cancel each other and lead to a more stable index size.

Another possibility is that Google is keeping the size of their index within a fixed range due to economic or engineering considerations. In this case, perhaps the web pages deemed least valuable or useful are eliminated from the index as more valuable pages appear. It is known that Google does not index all the pages they crawl (Alpert & Hajaj, 2008) due to quality considerations.

A final possibility is that all of our estimation methods are biased towards content in the Anglosphere or the West more generally. It is possible that both Google and Common Crawl are not representative examples of the global web, but only a part of it. Since most of the recent growth in the number of internet users has been in non-Western countries (Kemp, 2023a), it is possible that our estimates are missing this growth.

C. Non-public text data

The term "deep web" refers to that portion of the web that is not accessible by search engines. While this includes many categories of data, in this section we focus on closed[29] content platforms, since they attract a large share of all web traffic.[30]

Given the power-law nature of web pages usage, we can obtain a fairly reliable estimate of the total size of the deep web just by examining a few of the most-visited platforms. In particular, we estimate the amount of data in Facebook, Instagram and Twitter, three of the largest social media platforms, as well as Reddit, an indexed and open social media platform that we use as a sanity check. We then divide the estimate of the stock of each platform by its share of global traffic to arrive at a global estimate.

Facebook: Facebook has about 3B users, who on average make one post and five comments per month (Kemp, 2023c). From a sample of 70,000 posts and 200,000 comments (Krebs et al., 2017) we calculate the average post has 60 tokens, while the average comment has 26 tokens.[31]. This corresponds to 10T tokens being produced each year. Since Facebook has had a roughly similar number of users for the past 10 years or so, its total stock is around 100T tokens.

Instagram: Instagram has 1.6B users (Kemp, 2023a), who in aggregate make about 66,000 posts per minute (Domo, 2022), or about 100 million per year. The average post has about 8-16 tokens (Thorgren et al.), for a total of around 800B tokens generated per year. Over the past 10 years or so, this corresponds to 8T tokens.

Twitter: Twitter has close to 300M monthly active users. Each user posts 1.3 times per day, and the average post has about 7 tokens (GDELT, 2020). This corresponds to about 1.5T tokens per year, and since Twitter has had a roughly constant number of users for the past 12 years, the total stock is around 17T tokens.

Reddit: Reddit is public and indexed, so accurate statistics are available. The number of posts per day was about 60,000 in 2020, while the number of comments was about 500,000 (Baumgartner et al., 2020). These have grown close to linearly since 2011, so the total is equivalent to about 7 years of submissions at that rate. Each post has on average 48 tokens, and each comment has 21.[32] This corresponds to a total of 75B tokens per year, or about 600B tokens in total.

Finally, we divide each estimate by the share of traffic of the corresponding platform to arrive at an estimate of the total stock in the web. We take the average of the estimates of each platform. As shown in Table 2, the results indicate that the size of the deep web is roughly comparable to the size of the indexed web, so using this source of data would only delay a data bottleneck by a couple years.

Platform	Tokens	Traffic share	Total stock estimate
Facebook	100T	3.6%	3000T
Instagram	8T	1.5%	580T
Twitter	17T	1.3%	1370T
Reddit	600B	0.44%	140T
Geom. mean			760T

Table 2. Estimates of the stock of data in the deep web. Estimates of the total stock are produced by dividing the estimate of the number of tokens by the share of traffic. These estimates are rough, and expected to be accurate only up to an order of magnitude.

[29] By closed, we mean that they usually require credentials to access and limit the visibility of their content to unregistered users.

[30] According to Similarweb, Google, YouTube, Facebook, Instagram, Twitter and Baidu account for a third of global web traffic (Similarweb, 2024).

[31] This particular sample might not be representative of the all the content in Facebook, but we think it's unlikely to be too different from the true averages.

[32] These are estimated from 3M posts and 4.5M comments in the Pushshift dataset.

C.1. Instant messaging and email

Instant messaging applications are widely used: Facebook Messenger alone has over a billion users (Kemp, 2023b). They consequently contain a vast amount of non-indexed data. In 2020, the messaging apps owned by Meta (then Facebook) were processing 100B messages per day (Mosseri & Chudnovsky, 2020). In a sample of 6 million WhatsApp messages, Rosenfeld et al. (2018) found that the average message had 5.66 words, or about 4.5 tokens. This corresponds to 165T tokens generated per year, for a total stock over a quadrillion tokens.

Note that the quality of instant messaging data relative to web data is not well documented, so this source of data might be much less useful for training than our calculation of the number of tokens suggests. That said, even if we take the number of tokens at face value, it is comparable to the stock of text in the indexed and deep web, and therefore instant messaging data only increases the stock of text by 50%, which would delay a data bottleneck by less than one year.

Emails could also be a large source of data, comparable to the overall size of the deep web. There are reportedly over 300B emails received per day (Radicati), which amounts to around 100T emails per year. It is unclear how many tokens of unique content this contains. It is likely that most are "mass emails" such as advertisements or newsletters such that most received emails are not unique or have substantial duplicate content. If we conservatively suppose that 10% of emails are unique and contain about 50 words on average (consistent with some existing estimates, see Taniguchi et al. (2020)), this would represent about 625T tokens, which is the same order of magnitude as our estimates of the stock of deep web text.

D. Non-text data

There are significant sources of public data in modalities different than text. Most notably, the web contains large quantities of images and videos which might be used to train multimodal systems. However, we believe that the amount of useful information contained in these additional sources of data is not enough to meaningfully change our conclusions. In this appendix we arrive at some rough estimates for the amount of data available for other common modalities to justify this conclusion.

Images It is hard to know exactly how many pictures are taken globally per year, but a reasonable estimate is on the order of a couple trillion (Lee, 2021). [33] Henighan et al. (2020) estimates that one image has at least as much information as around three tokens of text. Meanwhile, image encoders often use hundreds of tokens per image (Dosovitskiy et al., 2021), but there is probably significant redundancy in this representation. We take the geometric average of these two extremes to arrive at a reasonable middle point of $\sim$30 tokens per image. Assuming that this rate of image capture has been maintained for 10 years, this corresponds to a few hundred trillion tokens, roughly the same scale as the raw size of Common Crawl. For this reason, including images in our model is not likely to produce large changes in our results.

Video In the case of video, YouTube is the most-used video hosting platform worldwide. More than 500 hours of video are uploaded to YouTube every minute (YouTube), which corresponds to 1 trillion seconds of video uploaded per year. Using again the hypothesis that the information in an image corresponds to around 30 tokens of text,[34] assuming each second of video is as valuable as an independent image, and maintaining the same rate of video upload for 10 years, we arrive at an estimate of 100 trillion text-token-equivalents. YouTube represents 7% of the share of internet traffic (Similarweb, 2024), so if we assume that the share of content is similar to the share of traffic,[35] the total stock of video might be on the order of one quadrillion tokens. Since this is similar to the stock of data in the web, including video data in our estimates of the stock would not significantly change our results.

One relevant consideration regarding video and images is that their production is much easier to scale than text. 7B CMOS image sensors were produced in 2020 (Yole Développement, 2021). If all of them were used for recording, in a single year they could produce 2e17 seconds of video, 1000 times more than the current stock in the web.

Exotic modalities Stephens et al. (2015) estimated that astronomy and genomics produce several million terabytes of compressed data per year. At face value, this corresponds to a stock of roughly 1e18 tokens, a thousand times larger than our estimates of the stock for text, images and video.

However, these modalities of data have extremely high redundancy and significant noise. The amount of synergy that these modalities have with text is also an open question. For this reason, we currently cannot evaluate whether these alternative modalities can provide a lasting source of data.

[33] This is consistent with the average human taking about 100 pictures per year.

[34] It is possible that the useful information in videos is overwhelmingly in the spoken words, rather than the images. Since humans speak at 100-160 words per minute, the amount of tokens per second of audio cannot be higher than 4-5, so 30 tokens per second of video is still an upper bound.

[35] While this might not be exactly true, it seems hard to argue that other platforms contain a larger quantity of varied and useful videos.

E. Tokenization Schemes Across Datasets

This appendix examines the performance of various tokenization schemes across different datasets by analyzing the average number of tokens produced and the average number of characters per token. The data presented in Tables 3, 4, and 5 collectively demonstrate that common tokenizers produce a similar number of tokens across a diverse range of texts.

E.1. Tokenization in RefinedWeb Dataset

Table 3 compares three tokenizers (BERT, GPT2, and XLM-Net) on 1000 random web pages in the first segment of the RefinedWeb dataset, showing that the average number of tokens produced by each tokenizer is similar, ranging from 621 to 653, with a consistent number of characters per token (around 4.2 to 4.4). In this dataset the average length of a word in bytes including whitespace is 5.5, so on average there are 4.4/5.5 = 0.8 words per token.

Tokenizer	Mean Tokens	Chars/Token
BERT	621	4.4
GPT2	645	4.2
XLMNet	653	4.2

Table 3. Tokenization schemes in the first segment of RefinedWeb

E.2. Tokenization Using GPT2 Tokenizer Across Various Datasets

Table 4 demonstrates the performance of the GPT2 tokenizer across various datasets, with the characters per token metric remaining relatively consistent, ranging from 2.22 to 4.15. This suggests that the GPT2 tokenizer produces a similar number of tokens per character across diverse text data, despite differences in the nature and average character length of the datasets.

Dataset	TextsAvg Char Length	Chars/Token
Enron Emails	1010	1618	3.44
FreeLaw	5094	15707	3.53
GitHub	18337	5238	2.53
EuroParl	133	62975	2.5
DM-Mathematics	2007	8194	2.22
ArXiv	2434	47345	3.05
Books3	301	587352	4.15

Table 4. GPT2 Tokenizer performance across multiple datasets

E.3. Performance of Modern Tokenizers on Selected Datasets

Table 5 compares the performance of multiple tokenizers used in modern models (Mixtral-7B, Command-plus-R, and cl100k_base/GPT-4) on selected datasets. The characters per token metric remains fairly consistent for each tokenizer across the datasets and always between 2 and 5, indicating that the choice of tokenizer does not significantly affect the number of tokens produced per character.

Tokenizer	Dataset	Chars/Token
Mixtral-7B	Books3	3.8
	GitHub	2.88
	Chinese Modern Poetry	2.17
Command-plus-R	Books3	4.17
	GitHub	3.29
	Chinese Modern Poetry	3.22
cl100k_base	Books3	4.31
	GitHub	3.78
	RefinedWeb	4.41
	DM-mathematics	2.25
	Chinese Modern Poetry	2.1

Table 5. Performance of various tokenizers across selected datasets

The consistency in the characters per token metric across different tokenizers and datasets supports the conclusion that the conversion from the number of words to the number of tokens is roughly independent of the tokenizer used. This finding has implications for the comparability of tokenization results across studies using different tokenizers and datasets. Further statistical analysis, such as examining the variance or standard deviation of the characters per token metric, could provide additional insights into the consistency of tokenization across datasets and tokenizers.

F. Overtraining in the context of data scarcity

In this appendix we sketch a model of optimal scaling decisions under data scarcity. In particular, we examine how the decision to overtrain models might be affected by data scarcity.

Our starting point is the parametric scaling law of Hoffmann et al. (2022), which predicts the reducible loss of a model L given its number of parameters N and the size of its training dataset D (see Equation 9). Hoffmann et al. (2022) derive from this scaling law a relation between the sizes of the model and the dataset that minimize the reducible loss of their model given a fixed training compute budget. In particular, in compute-optimal models the ratio D/N is

around 20. We call this the Chinchilla scaling law, and we call models that follow it Chinchilla-optimal.

Models for which the ratio D/N is above the Chinchilla-optimal ratio are commonly called overtrained, while models that are below that ratio are called undertrained. At a fixed training compute budget, overtrained models require less compute during inference but more data during training. This is currently attractive for developers, as compute is relatively scarce compared to data. As a consequence, some well-known models, like Llama 3 (Meta, 2024), are overtrained.[36]

Profit maximization problem

$$\text{maximize} \quad I(P - 2N) - 6ND \tag{6}$$

$$\text{s.t.} \quad 6ND + 2NI = C_0 \tag{7}$$

$$\text{where} \quad I = I_0(AN^{-a} + BD^{-b})^{-r}P^{-h} \tag{8}$$

Here N is the number of parameters of the model, D is the size of the training dataset in tokens and P is the price of each inference token in (some multiple of) dollars. C_0 is the total computational budget for training and inference and I is the number of tokens produced during inference. A, a, B and b are fitted parameters of the scaling law, and I_0, r and h are parameters of the inference demand function. All parameters are positive.

We now examine the relationship between overtraining and undertraining in the context of a data bottleneck. To simplify the analysis, here we ignore the cost of gathering data and focus on the computational cost of the model during training and inference. We assume that developers want to achieve the maximum possible profit within their computational budget, and that this computational budget includes both training and inference. In particular, given N and D, as well as a certain number of inference tokens, I, the compute cost C is given by Equation 7.

We assume that inference demand is a function that increases with model quality and decreases with the price of inference. We use the inverse reducible loss L^{-1} as a proxy for quality.[37] The functional form is given by Equation 8.

The optimal scaling policy depends on the values of r and h. If $h = 0$ or $r = 0$, demand for inference is independent of price or capabilities, respectively. If $h = 1$, demand is constant in dollar terms. If $h > 1$, demand is decreasing in dollar terms. Since demand decreases with dollar price for most goods, we try values of h greater than one. Since the loss of a compute-optimal model scales as $\sim C^{-0.15}$, for the profit to increase as a function of training compute (which is what we would expect) r must be higher than $0.15^{-1} \approx 7$.

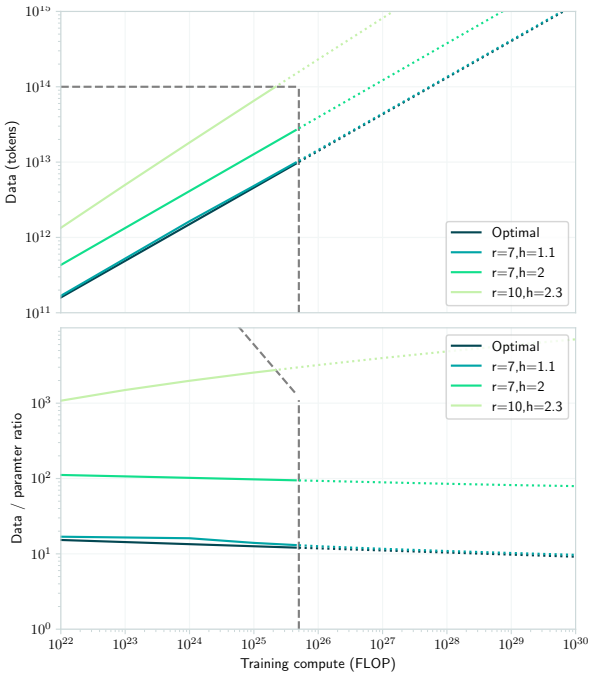

Figure 9. Scaling policies obtained by solving the optimization problem given by Equations 6-8. The dashed lines represent data and compute budgets of 100T tokens and 1e27 FLOP, respectively. Achievable configurations given these budgets are indicated by solid lines, while unachievable ones are indicated by dotted lines. Some scaling policies exhaust the compute budget first, meaning they are compute-constrained, while others exhaust the data budget first, indicating they are data-constrained.

In general, higher values of r and h lead to greater returns to overtraining, due to additional demand and price sensitivity. Figure 9 shows some optimal scaling curves for different values of these parameters of the inference demand function. In particular, the values $r = 7, h = 1.1$ lead to a policy that is very close to Chinchilla-optimal, while the values $r = 7, h = 2$ lead to about 5x overtraining, $r = 10, h = 2.4$ lead to a level of overtraining that increases with training compute, and is around 100x at 5e25 FLOP.

While Chinchilla-optimal scaling is bottlenecked by compute, under some assumptions the optimal scaling policy is instead bottlenecked by data. In any case, more overtraining leads to the stock of data being completely used earlier, as shown in Figure 10.

[36] Llama 3 has 70B parameters and was trained on 15T tokens, so it has 214 tokens per parameter, 11x more than the Chinchilla-optimal ratio.

[37] In general, the performance of a model is inversely correlated with the loss, and this relationship can be approximated by a power law (Henighan et al., 2020).

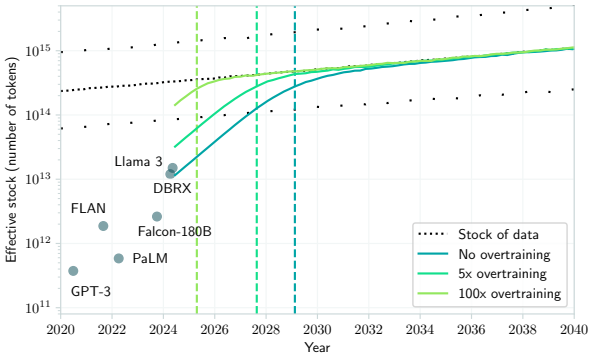

Figure 10. Compute-based projections of data usage and stock utilization years for the three profit-maximizing scaling policies from Figure 9. Only medians are shown for the dataset size projections.

Scaling law for reducible loss

$$L = AN^{-a} + BD^{-b} \quad (9)$$

Here L is the reducible loss, N is the number of parameters in the model, and B is the size of the dataset in tokens. The values of the parameters A, a, B and b are taken from Besiroglu et al. (2024).

G. Limits of undertraining in a data bottleneck

If data becomes scarce relative to compute, researchers might opt to undertrain increasingly large models on the existing stock of data. Using again the parametric scaling law found by Hoffmann et al. (2022), we can predict how much additional performance could be obtained from this approach.

In particular, we assume that the training data is fixed at 300T tokens and calculate the reducible loss predicted by Equation 9 as we increase the training compute by adding more parameters to the model. Figure 11 shows the result from this model: undertraining can provide the equivalent of up to 2 additional orders of magnitude of compute-optimal scaling, but requires 2-3 orders of magnitude more compute. This is enough to sustain a decreasing rate of progress for 3-6 additional years before the final plateau.

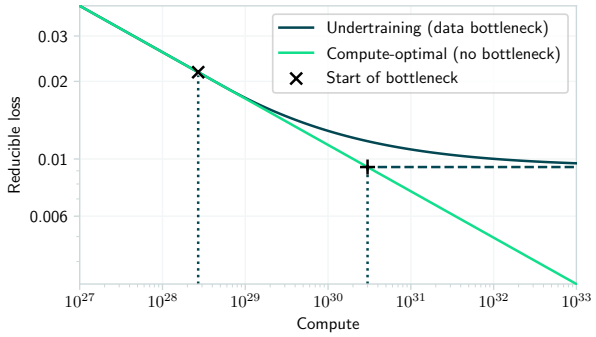

Figure 11. A toy model of undertraining in a data bottleneck scenario in which the stock of data is fixed at 300T tokens. The value of the loss is predicted using the parametric scaling law from Hoffmann et al. (2022), with the revised parameter estimates from Besiroglu et al. (2024). The compute-optimal training compute corresponding to a dataset size of 300T tokens is shown in the left-most black vertical line. We also plot compute-optimal scaling with unlimited data for comparison.