How Hyper-Datafication Impacts the Sustainability Costs in Frontier AI

ANONYMOUS AUTHOR(S)
SUBMISSION ID: XXX

Large-scale data has fuelled the success of frontier artificial intelligence (AI) models over the past decade. This expansion has relied on sustained efforts by large technology corporations to aggregate and curate internet-scale datasets. In this work, we examine the environmental, social, and economic costs of large-scale data in AI through a sustainability lens. We argue that the field is shifting from *building models from data* to actively *creating data for building models*. We characterise this transition as *hyper-datafication*, which marks a critical juncture for the future of frontier AI and its societal impacts. To quantify and contextualise data-related costs, we analyse approximately 550,000 datasets from the Hugging Face Hub, focusing on dataset growth, storage-related energy consumption and carbon footprint, and societal representation using language data. We complement this analysis with qualitative responses from data workers in Kenya to examine the labour involved, including direct employment by big tech corporations and exposure to graphic content. We further draw on external data sources to substantiate our findings by illustrating the global disparity in data centre infrastructure. Our analyses reveal that hyper-datafication does not merely increase resource consumption but systematically redistributes environmental burdens, labour risks, and representational harms toward the Global South, precarious data workers, and under-represented cultures. Thus, we propose Data PROOFS recommendations spanning provenance, resource awareness, ownership, openness, frugality, and standards to mitigate these costs. Our work aims to make visible the often-overlooked costs of data that underpin frontier AI and to stimulate broader debate within the research community and beyond.[1]

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; • **Hardware** → **Impact on the environment**.

Additional Key Words and Phrases: sustainability, data, artificial intelligence, social impact, carbon footprint

1 Introduction

The recent advancements in artificial intelligence (AI) have enabled automation that was widely considered decades away. AI systems now deliver breakthroughs across scientific discovery, language understanding, and generative modelling, addressing long-standing research challenges while opening new applications [10, 11, 45, 51, 52, 80]. This progress is driven primarily by deep learning methods [54, 75], including large language models (LLMs), multimodal generative models, and reasoning models [11, 29, 32, 69, 87].

While algorithmic and hardware improvements have played an important role, *scale* has acted as the dominant catalyst [36, 47, 78]. Frontier AI development now relies on unprecedented levels of resources: models with hundreds of billions of parameters, hyper-scale data centre infrastructure, electricity consumption comparable to that of entire towns, and financial investments that exceed the gross domestic product of several countries [77].

These investments would be futile without access to another critical resource: *data*. Contemporary frontier AI models train on datasets containing tens of trillions of data points, often scraped at scale from the internet. For example, one of the largest publicly available tokenised datasets `DCLM-Pool`, contains more than 240 trillion tokens [55]. Figure 1 illustrates the rapid growth of datasets and data volume using the datasets available on the Hugging Face Hub[2], a platform for hosting AI models and datasets (see Section 3).

[1]Source code used to extract metadata from datasets hosted on the Hugging Face Hub is available at: https://github.com/saintslab/costs-of-hyperdatafication.
[2]https://huggingface.co/datasets

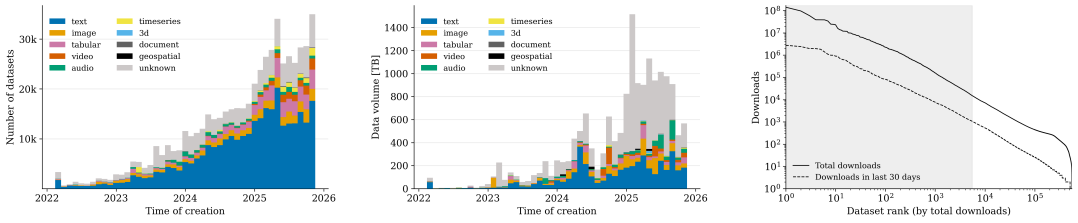

Fig. 1. Growth of datasets and data volume over time and download concentration on the Hugging Face Hub. Left: Monthly counts of newly created datasets (100% of datasets included). Centre: Monthly data volume added (91% of datasets included). The bars are coloured by modality. For multimodal datasets, each modality contributes proportionally to the colouring. Right: Download concentration. The plot shows total downloads (solid line) and downloads in November 2025 – which is 30 days preceding metadata extraction – (dashed line) against dataset rank sorted by total downloads on a logarithmic scale. The shaded region highlights the top 1% of datasets (5,543 repositories).

Growing concern about the sustainability of AI has prompted an expanding body of work [77, 94, 101]. However, existing discussions largely centre on model training and deployment. The sustainability of data itself remains comparatively under-examined, despite the fact that collecting and using internet-scale datasets for frontier AI is highly resource-intensive and can generate substantial environmental, social, and economic costs. As a result, current sustainability discussions overlook data as a central driver of AI's broader impacts.

In this work, we take a comprehensive look at the sustainability of data for AI. Building on existing critiques of the sustainability of AI [94, 100, 101], we focus explicitly on the environmental, social, and economic costs of large-scale data practices. This perspective becomes increasingly urgent as frontier AI research pursues ever more ambitious goals, including artificial general intelligence (AGI), intensifying the demand for data at scale.

We argue that the field is undergoing a structural shift in how data is produced and used. First, frontier AI systems increasingly rely on large-scale synthetic data. For example, the Phi-4 model was pretrained on approximately 10 trillion tokens, of which around 40% were synthetically generated and curated using other AI models [3]. Second, and more fundamentally, AI development is driving the active creation of new data sources designed explicitly for model training. A prominent example is the Ego4D dataset, which consists of 3,670 hours of egocentric video collected from 923 participants across 74 locations, recorded specifically to support AI systems [30]. Such datasets would not exist outside the demands of model development.

We use the term *hyper-datafication* to characterise this convergence, which we define formally below. While the term has appeared sporadically in prior work, it has been used narrowly to describe the large-scale expansion and aggregation of existing data sources [72]. We extend this usage by explicitly incorporating two additional, increasingly central dimensions: the algorithmic generation of synthetic data and the creation of data whose primary purpose is to serve as training material for AI models rather than to support direct human activities.

Hyper-datafication refers to the industrialised production and accumulation of data for AI model development across three coupled processes: (i) the large-scale scaling and recombination of existing data sources, (ii) the use of AI systems to generate synthetic data, and (iii) the creation of purpose-built data whose primary function is to serve as training input for AI systems rather than direct human use.

We analyse the costs of hyper-datafication across the different sustainability dimensions. To support this analysis, we draw on multiple empirical sources. First, we conduct a large-scale metadata study of approximately 550, 000 datasets hosted on the Hugging Face Hub, documenting growth in dataset scale, size, and scope. Using this metadata, we estimate storage-related energy consumption and associated carbon footprints on both the provider and user side, providing indicators of broader environmental trends. We then first examine social costs through a labour-focused analysis. Here, we draw on qualitative evidence from a structured questionnaire administered to data workers in Kenya to illustrate how hyper-datafication shapes the scale, nature, and conditions of contemporary data work. We further examine social costs by analysing patterns of language representation in the Hugging Face Datasets, assessing how supply and demand for datasets are distributed across languages relative to global speaker populations and web presence. This analysis highlights systematic linguistic concentration and under-representation, revealing how hyper-datafication amplifies existing asymmetries in whose languages and cultures are most extensively captured in AI training data. Finally, we incorporate external data sources to contextualise and substantiate our economic analysis and to support claims in the environmental and social sections, drawing on reports, policy documents, and prior empirical studies where direct measurement is not feasible. Based on these analyses, we propose actionable recommendations, formulated as the Data PROOFS guidelines, spanning provenance, resource awareness, ownership, openness, frugality, and standards to mitigate the costs of hyper-datafication.

The following sections briefly discuss existing research on the sustainability impacts of AI systems, and the importance of data in this context (Section 2). We describe our methods for gathering information on the sustainability implications of AI-related data (Section 3), and we analyse our findings in light of the growing trend of hyper-datafication (Section 4). We conclude with a set of recommendations (Data PROOFS) to increase data sustainability and address the actions required to reduce the negative impacts of hyper-datafication (Section 5).

2 Background and Related Work

2.1 Sustainability of AI

Sustainability is broadly understood to have three pillars: environmental, social, and economic [68]. Together, these reflect the need to balance long-term economic viability, ecological limits, and societal well-being when assessing systemic impacts. This holistic approach is essential for addressing the global challenges when pursuing resource-intensive technologies like AI on a planet with a rapidly changing climate [46].

Energy Consumption and Carbon Footprint of AI. The emerging critique on the sustainability of AI has been primarily focused on its environmental impact; particularly, the growing carbon footprint due to the energy consumption of AI [6, 33, 85]. The International Energy Agency (IEA) projects that data centre electricity consumption worldwide will more than double by 2030 (to between 945 and 1300 TWh as depicted in Figure 3), mainly driven by AI workloads [35, 92]. This is evident in recent hyper-scale data centre projects. For instance, Microsoft's USD 106 billion data centre with 3.33 GW capacity in Wisconsin, the United States (US), will consume electricity equivalent to that of 3.3 million households in Wisconsin [23] (see Appendix D.1 for details). Globally, energy production continues to be one of the largest sources of greenhouse gas (GHG) emissions [12, 42]. The growing energy needs of data centres for AI result in a corresponding carbon footprint that has also risen in recent years [6, 27, 59, 85].

Broader Environmental Impact of AI. Data centres use vast volumes of fresh water for cooling [56] threatening freshwater resources in vulnerable regions [22]. The manufacture of hardware, use of rare earths, electronic waste generation, and land-use change associated with AI data infrastructure contribute major embodied carbon emissions

and ecological damage [46, 100]. The amount of e-waste has grown from 34 million tons in 2010 to 64 million tons in 2022, and is projected to grow to 82 million tons by 2030. During the same time, the fraction of e-waste recycled decreased from 24% to less than 5% [91].

Social Sustainability of AI. The social cost of AI has been studied using the lens of fairness and bias mitigation [9], differential privacy [2, 21], and security using robustness to adversarial attacks [18]. Recent studies have begun to examine the labour required to build frontier AI models [14, 73], a critical factor that is often overlooked in popular debates regarding the social impact of artificial intelligence [66]. Finally, the democratisation of AI is negatively impacted due to the large-scale resource consumption and concentration of these expensive resources with a few actors [4, 8, 94].

2.2 Costs of Data

The majority of the literature discussed in Section 2.1 focuses on model selection, model training, and model deployment costs. The underlying data costs are subsumed into model development. However, this does not adequately capture the actual resource costs. For example, the model card for the 405 billion parameter open-source LLM, Llama-3.1, reports a carbon footprint of 8,930 tonnes of carbon dioxide equivalent (tCO2eq) due to the energy consumption of 21.5 GWh [5, 77]. This only includes the training cost and does not include any data-related energy consumption. Curating the dataset for training Llama-3.1 with about 15 trillion tokens, generating synthetic data of 25 million tokens, preprocessing and storing them also incur additional environmental costs. When aggregated across the full scale of data used for AI, these costs can be substantial.

Environmental. While there are no comprehensive studies that estimate the environmental cost of AI datasets, there are specialised studies that focus on domain-specific data. For example, Souter et al. [81] study the impact of data preprocessing in medical image analysis, benchmarking the effects of different preprocessing stages. In Wang et al., authors estimate the direct and embodied emissions due to global land cover mapping projects which is a data-intensive task due to the high-resolution satellite images [98]. They estimated the emissions between 2014 and 2024 to be about 3.84 million tCO2eq, and project it to increase 60 times to about 184 million tCO2eq by 2050. Outside of AI, carbon emissions across the data lifecycle have been explored by Mersy and Krishnan [64]; they propose *carbon provenance* to annotate data with carbon emission meta-data, facilitating better carbon accounting of data.

Economic. Research on the economic consequences of AI data highlights the uneven distribution of the burdens and the benefits across regions and actors. Studies show that AI development reproduces structural inequalities between the Minority World and the Majority World [65, 89]. The extreme market concentration reinforces this inequality: a small group of dominant platforms holds disproportionate control over behavioural and operational data, creating data monopolies and high barriers to entry for smaller actors [76, 103]. By analysing the economic data value chain, Jia et al. [44] argue that monetary value systematically accrues to aggregators and model developers, i.e., the dominant platforms, while data generators are largely excluded.

Social. Parallel research examines the social costs borne by data workers who produce, clean, and label data. Studies of annotation centres show that labour is governed through continuous measurement of speed, volume, and accuracy with strict productivity expectations and constant pressure to meet daily targets [17]. Interviews with annotators in India and Madagascar further document the pervasive surveillance and escalating productivity demands [53, 97]. Research on micro-task platforms reports similar patterns, including long and irregular working hours, unpaid work, and significant income volatility [37, 90]. Beyond these conditions, the mental health impacts of data work are increasingly recognised. Studies of content moderation report high levels of stress, secondary trauma, and PTSD-like symptoms among workers

exposed to disturbing material [82–84]. While much of this research has focused on social-media moderation, the documented harms also extend to content moderation within AI data pipelines.

Frontier AI development is relying on large-scale datasets with trillions of data points [5, 78, 95]. There are even speculations that we will "run out of data", prompting the use of synthetic data for AI development [96]. Considering the costs of such hyper-datafication is crucial to address the sustainability of frontier AI comprehensively.

3 Methodology

To address the costs of hyper-datafication holistically, we estimate the energy use of large-scale data storage (environmental), analyse labour conditions using questionnaires administered to data workers and patterns of language representation in AI datasets (social), and draw on public data to examine economic costs associated with AI data (economical). We aim to identify: (i) the environmental impacts of data storage under hyper-datafication; (ii) the actors and labour conditions that sustain hyper-datafication; (iii) the extent to which hyper-datafication represents society using language as a proxy; and (iv) the economic burdens associated with hyper-datafication.

3.1 Metadata of Hugging Face Datasets

To analyse the trends in dataset growth, quantify storage-related energy costs, and examine language representation, we analyse metadata for datasets hosted on the Hugging Face Hub.

Scope. We retrieved metadata on 1 December 2025, when the Hugging Face Hub hosted 570,802 datasets. Metadata extraction failed for 2.89% of repositories, yielding a final sample of 554,300 datasets. We collected repository-level metadata (identifiers, timestamps, downloads, and Hub-side storage), dataset-level metadata (dataset size), and contextual attributes (region, modality, task, and language). Hugging Face has options to query download statistics in two granularities: all time and last 30 days. As we retrieved our data on 1 December 2025, the statistics for *last 30 days* correspond to November 2025.

Data collection. We collected repository-level metadata via the Hugging Face REST API[3], obtained dataset size statistics separately from the Hugging Face datasets-server API[4], and extracted contextual attributes from the Hugging Face Hub Python API[5] by querying the full public dataset catalogue. These attributes are primarily self-declared and therefore exhibit lower coverage. Appendix A.1 summarises all metadata attributes and coverage.

Data analysis. We performed basic preprocessing and used the dataset size as a proxy for local storage footprint, while Hub-side storage was used as a proxy for platform storage. We mapped language tags to ISO-639 codes. For the language analysis, dataset volumes were compared against global speaker distributions and web presence, using Common Crawl page counts as a proxy for online content. Extended descriptive statistics are provided in Appendix A.2.

3.2 Questionnaire for Data Workers

To assess social and labour costs, we conducted an online questionnaire targeting data workers in Kenya.

Participants. The final sample includes 134 respondents located in Kenya. The sample comprises 57 females and 77 males, predominantly aged 20-40 years, with most reporting at least four years of experience. Respondents were informed about the study's purpose and how their data would be used. They were free to participate and stop the survey at any time and were unpaid.

[3]Using the endpoint https://huggingface.co/api/datasets/{id} (one request per dataset ID) and retrieving attributes via the API's expand fields.
[4]See https://github.com/huggingface/dataset-viewer for documentation of the datasets-server infrastructure.
[5]Accessing the public dataset catalogue at https://huggingface.co/datasets

Data collection. We administered an English-language questionnaire via an online form and distributed it via the messaging application Telegram to data workers who had voluntarily joined the Telegram channel organised by a local data workers' collective. Data collection took place in December 2025. The questionnaire comprised ten questions covering demographics, working conditions, exposure to graphic content, and employment relationships with large technology companies. The complete questionnaire is provided in Appendix B.

Data analysis. We computed descriptive statistics and pairwise associations across working hours, experience, salary, exposure to graphic content, and employment type. Some questions allowed free-text responses, which, in a few cases, were used to harmonise categories. Free-text responses that could not be mapped consistently to the catergories were excluded. Detailed preprocessing decisions and complete raw counts are reported in Appendix C.

3.3 External Data Sources

We extract global annual investment in data centre infrastructure from the International Energy Agency (IEA) [41]. To characterise the geographic concentration and expansion of data centres, we use records from the Data Centre Map [1]. We assessed electricity demand associated with data centres using historical and projected estimates from the IEA [38] and focus on electricity consumption as a primary indicator of environmental impact associated with data centre expansion. We source the carbon intensity for electricity generation from Our World in Data [70].

4 Costs of Hyper-Datafication

The costs associated with large-scale data accumulation are multifaceted, and the different dimensions overlap and interact in practice. We separate these categories into environmental, social, and economic for analytical clarity. Moreover, each dimension involves nuances beyond the scope of this study. We therefore focus on aspects most relevant for understanding the structural consequences of sustained data growth.

4.1 Environmental Costs

Hyper-datafication creates a persistent obligation to store data. As datasets grow in number and size, storage requirements accumulate and become environmentally significant. Unlike model training and inference, storage remains largely invisible in environmental sustainability discussions, yet it underpins the continued operation of data-intensive AI systems. The Hugging Face Hub provides a concrete case for examining how these storage demands scale over time.

Figure 1 (Left and Centre) shows a sharp increase in both the number and volume of newly created datasets on the Hugging Face Hub since March 2022. Over this period, the monthly dataset upload rate has increased by more than an order of magnitude. At the same time, the volume of newly added data has increased by nearly three orders of magnitude, from tens of terabytes per month to peak levels exceeding one petabyte per month. While dataset creation has accelerated rapidly, usage remains highly concentrated. Figure 1 (Right) shows that the top 1% of datasets (5,543 repositories) account for 87.3% of total downloads and 81.7% of downloads in November 2025.

Figure 2 reports estimated provider-side (Left) and user-side storage (Right) energy consumption associated with Hugging Face datasets. These estimations are based on the assumptions and calculations provided in Appendix D.2. Provider-side storage energy consumption remains modest in absolute terms but increases steadily as datasets accumulate, having reached approximately 1 GWh. User-side costs are approximately three orders of magnitude larger (>2 TWh), driven by the repeated number of downloads across a large user base. These estimates account only for actual

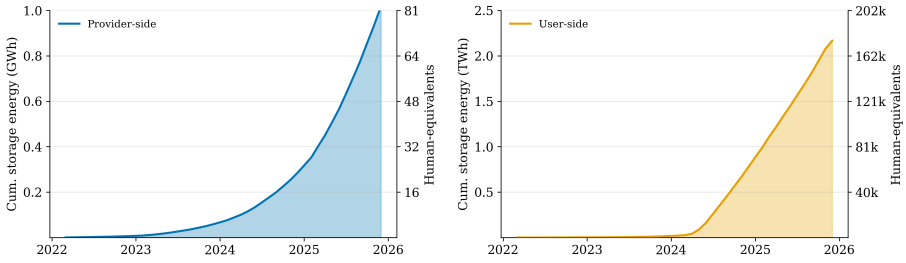

Fig. 2. Left: Estimated provider-side storage energy (GWh). Right: Estimated user-side storage energy (TWh), assuming that 10 percent of downloads result in three months of local storage. Right axes show human-equivalent annual emissions. To provide a sense of scale, we relate cumulative emissions to the global average annual per-capita footprint of 4.73 tCO_2eq [71].

file downloads. Streamed access does not increase download counts[6] and does not generate long-term local storage. Because Hugging Face supports streaming and on-demand access, shared infrastructure likely reduces duplication relative to alternative distribution models.

Since November 2023, Hugging Face has enabled users to specify the geographic storage region for datasets[7]. The available regions are the US and the European Union (EU). Our results show an intense concentration of storage in the US. Of the datasets with region information (99.8%), 552,713 are hosted in the US, compared with only 316 in the EU. This is not surprising, as the majority of Hugging Face's infrastructure is in the US.

Carbon intensity varies across electricity grids due to the differences in the energy mix. Storing all Hugging Face data in the EU rather than the US would reduce the associated carbon footprint by approximately 38%, due to the lower average carbon intensity in the EU (237 gCO2eq/kWh) compared with the US (384 gCO2eq/kWh) in 2024 [70]. Depending on the location of data centre infrastructure, storage-related carbon footprints can differ by up to a factor of 30 across regions[8]. Figure 2 (Right) also shows the user-side carbon footprint of downloading datasets from Hugging Face to be comparable to the annual carbon footprint of about 170,000 people (right vertical axis).

Estimating electricity consumption and carbon emissions for *all AI datasets* is infeasible beyond focused analyses such as the one presented using the Hugging Face Hub. This has to be performed at a coarser scale using data centre-level analysis. Figure 3 (Right) shows the growth projection of data centres, which is highly uneven across geographic regions. The US dominates both historical and projected electricity use, followed by China. Electricity demand in Asia, excluding China, Europe, and the rest of the world, remains substantially lower. Nevertheless, hyper-datafication and increasing pressure to adopt frontier AI drove a 27% increase in global data centre electricity demand between 2022 and 2024 [38]. Projections indicate this consumption will soon rival the total electricity usage of the entire African continent, as shown in Figure 3 (Left).

In addition to the carbon footprint due to the operational electricity consumption of a data centre, there are additional emissions related to data transmission. Operating networking infrastructure, such as wireless network equipment, optical fibre, and other switching equipment, contributes additional emissions [60]. However, when considering transmission costs, the embodied emissions from the manufacturing and construction of underwater optical fibre networks are the

[6] https://huggingface.co/docs/hub/en/datasets-download-stats
[7] https://huggingface.co/docs/hub/storage-regions
[8] This reflects contrasts between high- and low-carbon electricity grids, such as Kosovo (959 gCO2eq/kWh) and Norway (31 gCO2eq/kWh) in 2024 [70].

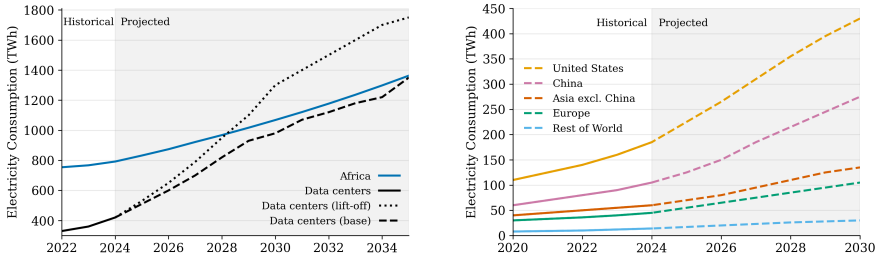

Fig. 3. Left: Historical (2022–2024) and projected (2024–2034) electricity use for all data centres worldwide under two scenarios: a base case reflecting current regulatory conditions and industry projections, and a lift-off case assuming stronger AI adoption enabled by faster data centre deployment as modelled by the IEA [38]. Shown alongside the total electricity consumption in Africa assuming a 5% annual grow [40]. Right: Regional distribution of data centre electricity demand in the base case (2020–2030), showing contributions from the US, China, Asia excluding China, Europe, and the rest of the world [38].

primary factors. It is reported that the embodied emissions of optical fibre are about 2.3 kgCO2eq/km [86] which are amortised over every bit of data that is transferred.

4.2 Social Costs

Data Worker Conditions in Kenya. Kenya has become one of the regional hubs for data work, supported by (relatively) strong mobile infrastructure and a young, technologically capable population [99]. Current estimates suggest that around 1.9 million Kenyans participate in digital labour, including roughly 1.2 million gig workers. This accounts for approximately 3.3% of the population, however, the sector contributes more than 9% of the national GDP [99].

We analyse the labour-related social costs due to hyper-datafication based on the responses from 134 data workers in Kenya. Respondents report engaging in a wide range of tasks, including data labelling (120/134), content moderation (67/134), data cleaning (52/134), and verification of user preferences (33/134), with many performing multiple task types concurrently. A majority of the respondents (79/134) report that they currently work or have previously worked directly for a major technology company, including OpenAI (51/134), Meta (35/134), Google (24/134), Microsoft (13/134), and Amazon (10/134).

Figure 4 shows the relationship between working hours, experience, exposure, and salary. Roughly half of respondents (65/134) report working between 40 and 60 hours per week. The most common experience level is 4-6 years (70/134), while the most common salary range is USD 200-300 per month (42/134). For comparison, the Kenyan national average is USD 540 per month [16].

Exposure to graphic content emerges as a central social cost. A majority of respondents (89/134) report some level of exposure, with 60 reporting daily exposure and 25 reporting exposure every other day. This exposure is especially common among workers engaged in content moderation, which is increasingly rebranded as “Trust and Safety”.

Contrary to common claims, our data suggests that higher exposure does not correspond to higher pay. Figure 4 (Right) shows no consistent association between exposure level and salary. Conditioning on working hours and experience yields the same result.

Among respondents who report having worked directly for a large technology company at some point, half (40/79) report daily exposure to graphic content and a further significant share reports exposure every other day (17/79). Among those who do not report working for large technology companies, exposure rates are roughly one-third (20/55) report

Fig. 4. Left: Distribution of respondents across salary bands by weekly working hours. Centre: Distribution of respondents across salary bands by years of experience. Right: Monthly salary distribution by exposure level, indicating no clear relationship between exposure level and salary.

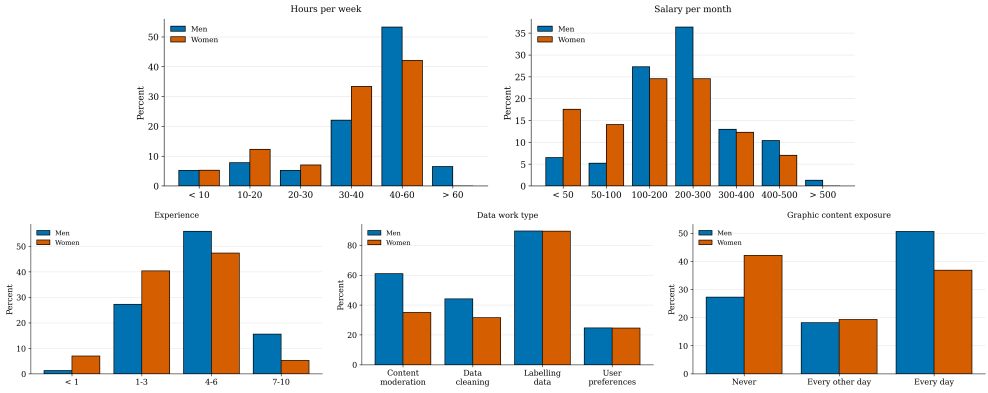

Fig. 5. Gender-disaggregated distributions of weekly working hours, monthly salary, experience, data work types, and exposure to graphic content derived from the anonymous responses to the questionnaire from 134 data workers.

daily exposure, and a smaller share report exposure every other day (8/55). The questionnaire does not distinguish between current and prior employment, and these patterns should therefore not be interpreted as evidence that direct employment with large technology companies causes higher exposure.

Gender-disaggregated results in Figure 5 reveal additional disparities. Women generally report fewer years of experience, shorter work hours, and lower salaries. Men report more frequently engaging in content moderation and greater exposure to graphic material.

Representation in Datasets. While hyper-datafication increases the volume of data available, this does not necessarily translate into better representation of the world. We use the Hugging Face datasets to investigate this further.

Figure 1 (Left and Centre) shows that text data is a large fraction of the diverse data modalities on Hugging Face. Using these text datasets, we first examine the representation of global languages in datasets hosted on the Hugging Face Hub. Among 57,484 datasets with language annotations, we identify 7,934 distinct languages. Most languages (96%) appear only within multilingual datasets and typically in a small number of repositories. The remaining 4% correspond to 306 languages, of which 97 occur only in a single dataset. This indicates that many languages exist in the margins of the dataset ecosystem rather than as sustained data sources.

Figure 6 (Left) shows the representation for the ten largest language groups (in terms of volume share) on the Hugging Face Hub. It compares each group's share of total dataset size on the Hugging Face Hub with its share of Common Crawl

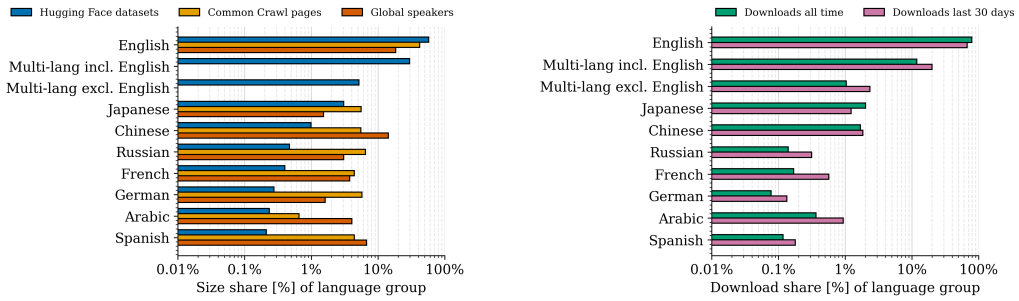

Fig. 6. Representation and demand for the ten largest language groups on the Hugging Face Hub. Left: A depiction of each group's share of dataset volume with its share of Common Crawl pages and global speakers. Right: All-time and recent downloads shares on Hugging Face. Note both horizontal axes are in logarithmic scale.

pages and global speakers. English represents a larger share of the dataset volume (57%) than its share of Common Crawl pages (42%) and more than three times its share of global speakers (18%), as shown in Table 2. Multilingual datasets that contain English add another 30% to this dominance. Several widely spoken languages, including Chinese, Arabic, and Spanish, remain substantially under-represented (see Appendix A.3). Appendix D.3 provides contextual evidence that these imbalances reflect broader asymmetries in global digital participation and traffic.

Figure 6 (Right) shows the distribution of downloads in November 2025 and all-time, where English again accounts for the majority. November 2025 download shares for non-English languages exceed their all-time shares in most cases, but remain small in absolute terms, indicating only limited shifts in demand. English-only datasets account for 79% of all-time downloads and 68% of November 2025 downloads. English-inclusive multilingual datasets account for a further 12% of all-time and 20% of November 2025 downloads. While November 2025 download shares for several non-English languages exceed their historical averages (see Appendix A.3), these increases remain small in absolute terms and do not meaningfully reduce English dominance.

4.3 Economic Costs

Hyper-datafication requires substantial investment to store, process, and transmit continuously growing volumes of data. These investments include specialised hardware, networking equipment, cooling systems, and energy infrastructure [62, 77]. As a result, hyper-datafication has a direct and visible economic footprint in the rapid expansion of data centre infrastructure.

As shown in Figure 7 (Left), global annual investment in data centres has increased more than fivefold over the past decade. These investments are justified by expectations that expanded data capacity will unlock future value through improved AI capabilities. In 2025, global investment in data centres is projected to reach USD 580 billion, exceeding the USD 540 billion invested in global oil supply [39].

This expansion is geographically concentrated, with the US projected to account for more than half of cumulative global data centre investment over the next five years [41]. As shown in Figure 7 (Right), the infrastructure development mirrors this concentration with facilities clustering in Western Europe, China, Australia, Canada, and particularly the US. This spatial pattern only partially aligns with where data is generated and where the social or economic benefits

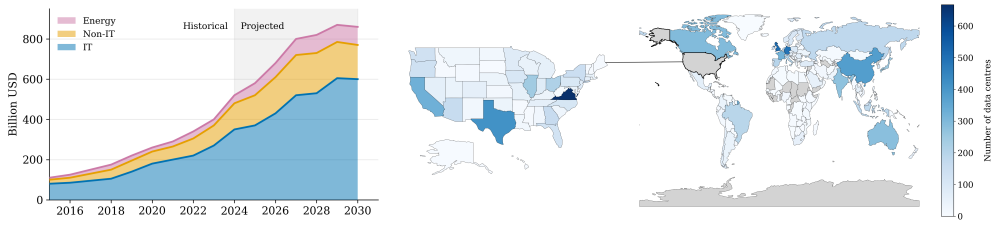

Fig. 7. Left: Historical (2015-2024) and projected (2024-2030) global annual investment in data centres in the base case reflecting current regulatory conditions and industry projections as modelled by the IEA [41]. Right: The world map shows the number of data centres per country. US is highlighted and shown on the left to display state-level variation. Data from Data Center Map [1].

ultimately accrue, as shown by Facebook, whose largest user base is in India with about 384 million users in January 2025, while most of Meta's infrastructure and revenue are concentrated in the US [49].

5 Discussions

Lack of Transparency. The environmental costs of hyper-datafication presented in Section 4.1 are based on a snapshot of 550,000 datasets on Hugging Face Hub. Figure 1 shows how the scale, size, and diversity of datasets are growing in recent years. Figure 2 captures the increase in storage-related operational energy consumption and carbon footprint. However, these estimations do not provide a comprehensive overview of the environmental impact of hyper-datafication. Factors such as the embodied emissions, fresh water to cool data centres, raw materials and minerals used in hardware manufacturing, and the e-waste generated due to hardware disposal are not considered [25, 77]. These broader environmental impacts are difficult to estimate, as acknowledged by works attempting to do this for AI model development [59, 100]. Lack of transparency and openness in reporting environmental costs by vendors, manufacturers, and actors at all stages of the data lifecycle makes comprehensive environmental cost reporting a challenge.

In addition to the environmental costs, social costs of data, such as labour, are seldom reported and discussed. Results from the data worker questionnaire in Section 4.2 highlight these trends where large corporations employ data workers for diverse types of data work. Although longer hours and greater experience are associated with higher earnings, as shown in Figure 4, salaries remain low relative to the Kenyan national average of USD 540 per month [16]. On the one hand, data work creates income opportunities in regions with limited formal employment.[9] On the other hand, these same structural conditions enable exploitation [79]. The availability of surplus labour, combined with limited regulatory oversight, allows data work to be organised through precarious contracts, low wages, and intensive performance monitoring. Economic necessity weakens workers' bargaining power, while task fragmentation and platform-based subcontracting obscure accountability. Transparent reporting of the costs pertaining to data work can give due credit to the data workers who remain unseen in the discourse surrounding frontier AI. Additionally, the due credit and visibility can help them secure better work conditions and bargaining rights.

Material Costs of Data and Resource-Awareness. The narrative of data "infiniteness" behind hyper-datafication is based on the view that data has a negligible cost, and ignores the material footprint of data. In Section 4.1, we presented concrete estimations of the energy consumption and carbon footprint of data using Hugging Face datasets. Extending these estimations to all data currently used for AI development will yield even higher material costs. Measuring the cost

[9] For example, the youth (15-34 years old) in Kenya face an unemployment rate of 67% [26].

of creating, storing, and processing datasets using existing tools like Carbontracker[10] [6] or CodeCarbon[11] and reporting the carbon emissions in standardised datasheets as proposed by Gebru et al. [28] are good starting points. Additionally, attaching environmental costs at each step as meta-data across the data life-cycle, as suggested by Mersy and Krishnan [64] for carbon-provenance-based AI, can clarify the assessment of the environmental sustainability of data.

Data Provenance and Ownership under Data Monopolisation. Another factor that is propelling hyper-datafication is the belief that more data necessarily creates more value, which also drives data concentration. Only a small number of firms, based in a handful of countries, possess the capital and infrastructure required to collect, store, and process data at hyper-scale. Figure 7 shows this spatial disparity, where vast regions of the world have few data centres, whereas the majority of the global data centres are located in the US. Such concentration creates high barriers to entry and reinforces monopolistic dynamics [63].

The concentration of data accumulation enables large-scale value extraction from user-generated and third-party data without corresponding compensation or meaningful control for data producers. Individuals and organisations contribute data under conditions of limited transparency, while platforms and the countries they belong to capture the resulting economic surplus [44]. The outcome is a persistent structural asymmetry in value appropriation [44]. Evaluating how value is created from data requires *comprehensive data provenance*, which is not straightforward with the frontier AI systems [57]. This is also closely tied to the question of *data ownership*, in which big tech platforms exploit data created by users to create value for themselves through frontier AI systems, manifesting extreme forms of appropriation [103]. Protecting people's rights in frontier AI systems, such as generative AI models, relies on data-provenance mechanisms and expanded legal protections, such as granting individuals copyright over their own identity. Recent proposals, including the Danish model of giving the copyright of their identity to people, illustrate how data and identity rights could be formally recognised in this way [13].

Data Frugality to Counter Hyper-Datafication. Hyper-Datafication promotes a *data abundance mindset* in which ever-larger datasets are treated as the primary driver of model progress, based on empirical scaling laws linking model performance to data and compute [36, 47]. Yet, our findings, evidenced by the storage-related energy use (Section 4.1), data labour conditions (Section 4.2), and infrastructure expansion (Section 4.3), show that this logic systematically externalises social and environmental costs: the storage and curation of ever-growing datasets impose increased labour burdens and raise the energy and carbon costs of data storage, curation, and model training. Those costs grow faster than the marginal performance gains they produce [7].

However, a frugal alternative exists. Techniques such as representative subset selection and coresets exploit the well-established fact that not all data is equally informative [48]. This allows models to achieve comparable performance with orders of magnitude less data and computation [50]. However, under current economic incentives, there is little motivation for companies to adopt such approaches. Data accumulation increases not only technical capacity but also strategic control and market valuation, encouraging data hoarding even when its utility is low.

For this reason, data frugality cannot be left to voluntary optimisation. It must be enforced as a structural constraint on hyper-datafication. A *corporate data tax* could internalise the social and environmental costs of data accumulation, correcting a market failure in which companies benefit from scale while communities and workers bear the burdens. Thus, the tax could discourage wasteful hoarding while generating revenue to compensate data generators and data workers. Frugality, in this sense, is not merely an efficiency principle but a corrective action to the extractive political economy of data [66].

[10] https://carbontracker.info/
[11] https://codecarbon.io/

Recommendations to Mitigate Costs of Hyper-datafication. We have presented several critiques of hyper-datafication in this section substantiated by evidence from Section 4. We synthesise these arguments into a set of recommendations, ranging from concrete technical measures to longer-term visions for the future, aimed at multiple stakeholders.

Data PROOFS Recommendations.

- **Provenance:** Build data provenance into frontier AI systems, for monitoring data quality and attributing due credit to data producers. This requires the development of new tools and regulations.
- **Resource-awareness:** Measure and report the environmental and labour costs of data to counter the narrative about data "infiniteness" and materially ground the bits.
- **Ownership:** Unless explicitly agreed, the data ownership should–by default–belong to users and not platforms. This must be encoded as digital citizen rights.
- **Openness:** Transparent reporting of data costs. Availability of data used for frontier AI development, so that researchers and policy makers can use it for inquiry.
- **Frugality:** Data frugality should be the guiding design principle instead of data abundance. Incentivised as a corporate data tax, which can then be used to compensate data owners and data workers.
- **Standards:** Develop measurement and reporting standards that can facilitate sustainability provenance by embedding environmental, social, and monetary costs as meta-data along the data life-cycle.

In addition to the Data PROOFS recommendations, **collective action** is indispensable to counter the consequences of hyper-datafication and to improve the sustainability of data in frontier AI. Community resistance is growing; in the US, more than 140 activist groups across 24 states are mobilising against new data centre construction and expansion [19]. Over the past two years, these groups have collectively stalled or cancelled proposed projects worth USD 162 billion. Citizen-led protests against hyper-scale data centres are appearing in Mexico, Spain, the Netherlands, India, and other parts of the world [67]. This trend highlights how the environmental, social, and economic costs of data centre development are becoming increasingly visible, and how local communities are contesting the uneven distribution of benefits and harms associated with hyper-scale AI development.

Limitations. We used datasets from Hugging Face Hub as it is one of the few platforms where dataset metadata is available in a standardised format. Extrapolating our arguments to broader data trends in the AI domain is a limitation of our work. However, Hugging Face Hub has evolved into a reliable platform for frontier AI data and models, and serves as a reasonable sample that captures the broader trends and has been used in other works [15, 58, 102]. Thus, our estimates can be seen as lower bounds on true data costs. Particularly, as we have noted in Section 4.2, the language representation on the Hugging Face Hub is highly uneven. It does not reflect the global distribution, nor does it mirror global patterns of digital activity. This should not be interpreted as a failure of a single platform. Rather, it reflects broader structural dynamics in how data is produced and used across the AI ecosystem.

Data work is carried out globally. However, our questionnaire was only sent to data workers in Kenya. We have refrained from making generalised claims and tried to limit the scope of labour discussions to the conditions in Kenya.

Likewise, some of the economic costs are primarily US-focused. This is again due to the availability of data sources. Furthermore, it reflects the dominance of a few regions, such as the US, which host the majority of global infrastructure

for frontier AI development. Additional clarifications about the limitations of using these data sources are provided in Appendix D.4.

6 Conclusions

The push for hyper-datafication is tied to the objective of modelling reality from data. Terms like digital twins are used widely in this context, aiming to mimic consumer behaviour [88], social interactions [34], virtual human twin [24], planetary climate [20, 61], and even human life events like death [74] from large-scale data. We are progressing from an era in which models relied mainly on passively accumulated data to one in which the world is actively reshaped to generate more data. Sensors, platforms, and infrastructures increasingly optimise for continuous data extraction. While hyper-datafication increases the amount of available data, it does not necessarily yield informative and useful data. Furthermore, hyper-datafication does not alleviate existing problems with data-driven models but can aggravate them.

We have presented evidence from multiple sources that bring forth the sustainability costs of hyper-datafication in frontier AI. The results presented in this work invite us to take a step back and reflect on the pursuit of frontier AI on the back of internet-scale data. Hyper-datafication is framed as an inevitable route to more capable and general AI systems. Our findings suggest that this is not a neutral form of progress. Instead, it redistributes benefits and burdens unevenly, while deepening already existing asymmetries in representation and power. We hope the Data PROOFS recommendations serve as inspiration for improving the overall sustainability of data in frontier AI.

Generative AI Usage Statement

GitHub Copilot and ChatGPT version 5.1 were used to support programming tasks, including the development of scripts for Hugging Face metadata extraction and data visualisation. Google Gemini, ChatGPT versions 5.1 and 5.2 were used to identify relevant academic and policy sources, assist with table generation, and to edit and refine language and grammar in selected sections of the manuscript.

References

[1] [n. d.]. DataCenterMap. https://www.datacentermap.com/datacenters/. Accessed: 2025-24-11.

[2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[3] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905* (2024).

[4] Nur Ahmed and Muntasir Wahed. 2020. The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581* (2020).

[5] AI@Meta. 2024. Llama 3 Model Card. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

[6] Lasse F. Wolff Anthony, Benjamin Kanding, and Raghavendra Selvan. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. In *ICML Workshop on Challenges in Deploying and Monitoring Machine Learning Systems*.

[7] Pedram Bakhtiarifard, Christian Igel, and Raghavendra Selvan. 2024. EC-NAS: Energy Consumption Aware Tabular Benchmarks for Neural Architecture Search. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5660–5664. https://doi.org/10.1109/icassp48485.2024.10448303

[8] Pedram Bakhtiarifard, Pınar Tözün, Christian Igel, and Raghavendra Selvan. 2025. Climate And Resource Awareness is Imperative to Achieving Sustainable AI (and Preventing a Global AI Arms Race). *arXiv preprint arXiv:2502.20016* (2025).

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

[10] Martin Brandt, Compton J Tucker, Ankit Kariryaa, Kjeld Rasmussen, Christin Abel, Jennifer Small, Jerome Chave, Laura Vang Rasmussen, Pierre Hiernaux, Abdoul Aziz Diouf, et al. 2020. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature* 587, 7832 (2020), 78–82.

[11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[12] Thomas Bruckner, Igor Alexeyevich Bashmakov, Yacob Mulugetta, Helen Chum, Angel De la Vega Navarro, James Edmonds, Andre Faaij, Bundit Fungtammasan, Amit Garg, Edgar Hertwich, et al. 2014. Energy systems. *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. (2014).

[13] Miranda Bryant. 2025. Denmark to tackle deepfakes by giving people copyright to their own features. https://www.theguardian.com/technology/2025/jun/27/deepfakes-denmark-copyright-law-artificial-intelligence. The Guardian Article..

[14] Callum Cant, James Muldoon, and Mark Graham. 2024. *Feeding the machine: The hidden human labor powering AI.* Bloomsbury Publishing USA.

[15] Joel Castaño, Silverio Martínez-Fernández, Xavier Franch, and Justus Bogner. 2024. Analyzing the evolution and maintenance of ml models on hugging face. In *Proceedings of the 21st International Conference on Mining Software Repositories*. 607–618.

[16] CEIC Data. 2024. Kenya Monthly Earnings. https://www.ceicdata.com/en/indicator/kenya/monthly-earnings Accessed: 2025-12-08.

[17] Srravya Chandhiramowuli, Alex S Taylor, Sara Heitlinger, and Ding Wang. 2024. Making data work count. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–26.

[18] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*. PMLR, 1310–1320.

[19] Data Center Watch. 2025. Data Center Watch Report. https://static1.squarespace.com/static/67819031da098341c45ac84a/t/6849bcfe640a951f79e00715/1749662975141/Data+Center+Watch+Report+.pdf Accessed: 2025-11-30.

[20] Francisco J Doblas-Reyes, Jenni Kontkanen, Irina Sandu, Mario Acosta, Mohammed Hussam Al Turjmam, Ivan Alsina-Ferrer, Miguel Andrés-Martínez, Leo Arriola, Marvin Axness, Marc Batlle Martín, et al. 2025. The Destination Earth digital twin for climate change adaptation. *EGUsphere* 2025 (2025), 1–41.

[21] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*. Springer, 265–284.

[22] Environmental and Energy Study Institute. 2024. Data Centers and Water Consumption. https://www.eesi.org/articles/view/data-centers-and-water-consumption Accessed: 2025-11-13.

[23] Epoch AI. 2025. "Frontier Data Centers". https://epoch.ai/data/data-centers Accessed: 2026-01-10.

[24] European Commission. 2023. European Virtual Human Twins (VHT) Initiative. https://digital-strategy.ec.europa.eu/en/policies/virtual-human-twins. EU Project.

[25] Sophia Falk, David Ekchajzer, Thibault Pirson, Etienne Lees-Perasso, Augustin Wattiez, Lisa Biber-Freudenberger, Sasha Luccioni, and Aimee van Wynsberghe. 2025. More than Carbon: Cradle-to-Grave environmental impacts of GenAI training on the Nvidia A100 GPU. *arXiv preprint arXiv:2509.00093* (2025).

[26] Federation of Kenya Employers. [n. d.]. *Youth Employment.* https://www.fke-kenya.org/policy-issues/youth-employment

[27] Charlotte Freitag, Mike Berners-Lee, Kelly Widdicks, Bran Knowles, Gordon S Blair, and Adrian Friday. 2021. The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns* (2021).

[28] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[29] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[30] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18995–19012.

[31] Gianluca Guidi, Francesca Dominici, Jonathan Gilmour, Kevin Butler, Eric Bell, Scott Delaney, and Falco J. Bargagli-Stoffi. 2024. Environmental Burden of United States Data Centers in the Artificial Intelligence Era. arXiv:2411.09786 [cs.CY] https://arxiv.org/abs/2411.09786

[32] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* 645, 8081 (2025), 633–638.

[33] Peter Henderson, Jieru Hu, Joshua Romoff, Emma Brunskill, Dan Jurafsky, and Joelle Pineau. 2020. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research* (2020).

[34] James Hendler and Alice M Mulvehill. 2016. *Social machines: the coming collision of artificial intelligence, social networking, and humanity.* Apress.

[35] Highland Economics Council. 2025. Measuring the Environmental Cost of Artificial Intelligence and Their Data Centers. https://www.hecweb.org/2025/06/28/measuring-the-environmental-cost-of-artificial-intelligence-and-their-data-centers/ Accessed: 2025-11-13.

[36] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 30016–30030.

[37] Lars Hornuf and Daniel Vrankar. 2022. Hourly wages in crowdworking: A meta-analysis. *Business & Information Systems Engineering* 64, 5 (2022), 553–573.

[38] International Energy Agency. 2024. Energy Demand from AI. https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai. Accessed: 2025-11-28.

[39] International Energy Agency. 2024. World Energy Outlook 2024. https://iea.blob.core.windows.net/assets/140a0470-5b90-4922-a0e9-838b3ac6918c/WorldEnergyOutlook2024.pdf Accessed: 2025-11-30.

[40] International Energy Agency. 2025. Electricity 2025. https://www.iea.org/reports/electricity-2025. Accessed: 2025-12-11.

[41] International Energy Agency. 2025. Energy and AI. https://www.iea.org/reports/energy-and-ai Licence: CC BY 4.0, Accessed: 2025-11-13.

[42] International Energy Agency. 2025. Global Energy Review. https://www.iea.org/reports/global-energy-review-2025. Accessed: 2025-06-10.

[43] International Telecommunication Union. 2024. Measuring Digital Development: Facts and Figures 2024. https://digitallibrary.un.org/record/4074377

[44] Ruoxi Jia, Luis Oala, Wenjie Xiong, Suqin Ge, Jiachen T. Wang, Feiyang Kang, and Dawn Song. 2025. A Sustainable AI Economy Needs Data Deals That Work for Generators. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*. https://openreview.net/forum?id=mdKzkjY1dM

[45] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *nature* 596, 7873 (2021), 583–589.

[46] Lynn H Kaack, Priya L Donti, Emma Strubell, George Kamiya, Felix Creutzig, and David Rolnick. 2022. Aligning artificial intelligence with climate change mitigation. *Nature Climate Change* 12, 6 (2022), 518–527.

[47] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361* (2020).

[48] Angelos Katharopoulos and François Fleuret. 2018. Not all samples are created equal: Deep learning with importance sampling. In *International conference on machine learning*. PMLR, 2525–2534.

[49] Simon Kemp. 2025. Digital 2025 Facebook Q1 Report. https://datareportal.com/essential-facebook-stats. Accessed on 2026-01-13.

[50] Krishnateja Killamsetty, Sivasubramanian Durga, Ganesh Ramakrishnan, Abir De, and Rishabh Iyer. 2021. GRAD-MATCH: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*. PMLR, 5464–5474.

[51] Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. 2023. Learning skillful medium-range global weather forecasting. *Science* 382, 6677 (2023), 1416–1421.

[52] Andreas D Lauritzen, Alejandro Rodríguez-Ruiz, My Catarina von Euler-Chelpin, Elsebeth Lynge, Ilse Vejborg, Mads Nielsen, Nico Karssemeijer, and Martin Lillholm. 2022. An artificial intelligence–based mammography screening protocol for breast cancer: outcome and radiologist workload. *Radiology* 304, 1 (2022), 41–49.

[53] Clément Le Ludec, Maxime Cornet, and Antonio A Casilli. 2023. The problem with annotation. Human labour and outsourcing between France and Madagascar. *Big Data & Society* 10, 2 (2023), 20539517231188723.

[54] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[55] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems* 37 (2024), 14200–14282.

[56] Pengfei Li, Jianyi Yang, Mohammad A Islam, and Shaolei Ren. 2025. Making AI less' thirsty': Uncovering and addressing the secret water footprint of AI models. *Commun. ACM* 68, 7 (2025), 54–61.

[57] Shayne Longpre, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, Nathan Khazam, Jad Kabbara, Kartik Perisetla, et al. 2024. A large-scale audit of dataset licensing and attribution in AI. *Nature Machine Intelligence* 6, 8 (2024), 975–987.

[58] Alexandra Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment?. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*. 85–99.

[59] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of bloom, a 176b parameter language model. *Journal of machine learning research* 24, 253 (2023), 1–15.

[60] Jens Malmodin, Dag Lundén, Åsa Moberg, Greger Andersson, and Mikael Nilsson. 2014. Life cycle assessment of ICT: carbon footprint and operational electricity use from the operator, national, and subscriber perspective in Sweden. *Journal of Industrial Ecology* 18, 6 (2014), 829–845.

[61] M Mardani, N Brenowitz, Y Cohen, J Pathak, CY Chen, CC Liu, A Vahdat, K Kashinath, J Kautz, and M Pritchard. 2023. Residual Diffusion Modeling for Km-scale Atmospheric Downscaling. arXiv 2023. *arXiv preprint arXiv:2309.15214* (2023).

[62] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. 2025. The 2025 AI Index Report. Online Resource. https://hai.stanford.edu/ai-index/2025-ai-index-report

[63] Ulises A Mejias and Nick Couldry. 2024. Data grab: The new colonialism of big tech and how to fight back. In *Data Grab*. University of Chicago Press.

[64] Gabriel Mersy and Sanjay Krishnan. 2024. Toward a life cycle assessment for the carbon footprint of data. *ACM SIGENERGY Energy Informatics Review* 4, 5 (2024), 25–33.

[65] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684.

[66] Matteo Pasquinelli. 2023. *The eye of the master: A social history of artificial intelligence*. Verso Books.

[67] Adam Satariano Paul Mozur and Emiliano Rodríguez Mega. 2025. From Mexico to Ireland, Fury Mounts Over a Global A.I. Frenzy. https://www.nytimes.com/2025/10/20/technology/ai-data-center-backlash-mexico-ireland.html. New York Times Article..

[68] Ben Purvis, Yong Mao, and Darren Robinson. 2019. Three pillars of sustainability: in search of conceptual origins. *Sustainability science* 14, 3 (2019), 681–695.

[69] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Online Technical Report.

[70] Hannah Ritchie, Pablo Rosado, and Max Roser. 2023. Data Page: Carbon intensity of electricity generation. https://archive.ourworldindata.org/20251014-145858/grapher/carbon-intensity-electricity.html. Data adapted from Ember and the Energy Institute. Part of Ritchie, Rosado and Roser (2023) *Energy*. Archived on 14 October 2025..

[71] Hannah Ritchie, Pablo Rosado, and Max Roser. 2023. Data Page: CO_2 emissions per capita. https://archive.ourworldindata.org/20251113-170236/grapher/co-emissions-per-capita.html. Part of the publication: *CO_2 and Greenhouse Gas Emissions*. Data adapted from the Global Carbon Project and other sources. Online resource archived on 13 November 2025.

[72] Andrea Rosales and Sara Suárez-Gonzalo. 2024. *Chapter 3 Peter's Problem. An Analysis of the Imaginaries about Automated Futures Portrayed in QualityLand*. De Gruyter, Berlin, Boston, 37–54. https://doi.org/doi:10.1515/9783110792256-003

[73] Jürgen Rudolph. 2025. The hidden labour in AI: Big Tech's dirty secret and the need for critical AI literacy in higher education. *Handbook of AI and higher education. Edward Elgar* (2025).

[74] Germans Savcisens, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. 2024. Using sequences of life-events to predict human lives. *Nature Computational Science* 4, 1 (2024), 43–56.

[75] Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks* 61 (2015), 85–117.

[76] Heike Schweitzer, Jacques Crémer, and Yves-Alexandre de Montjoye. 2019. Competition Policy for the Digital Era. https://www.rewi.hu-berlin.de/de/lf/oe/rdt/pub/working-paper-no-6

[77] Raghavendra Selvan. 2025. *Sustainable AI*. O'Reilly.

[78] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, and Pablo Villalobos. 2022. Compute trends across three eras of machine learning. In *2022 international joint conference on neural networks (IJCNN)*. IEEE, 1–8.

[79] Wilington Shitawa. 2024. Click Captives: The Unseen Struggle of Data Workers. Data Workers' Inquiry. https://data-workers.org/wilington/

[80] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[81] Nicholas E Souter, Chris Racey, Nikhil Bhagwat, Reese Wilkinson, Niall W Duncan, Gabrielle Samuel, Loïc Lannelongue, Raghavendra Selvan, and Charlotte L Rae. 2025. Comparing the carbon footprint of fMRI data processing and analysis approaches. *Imaging Neuroscience* 3 (2025), IMAG–a.

[82] Ruth Spence, Antonia Bifulco, Paula Bradbury, Elena Martellozzo, and Jeffrey DeMarco. 2023. The psychological impacts of content moderation on content moderators: A qualitative study. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace* 17, 4 (2023).

[83] Ruth Spence and Jeffrey DeMarco. 2025. Content moderator mental health and associations with coping styles: replication and extension of previous studies. *Behavioral Sciences* 15, 4 (2025), 487.

[84] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[85] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3645–3650.

[86] Aislin Sullivan, Pushkar Tandon, Roshene McCool, Amalia Diaz, and Constantin Herrmann. 2023. A sustainable future with optical fiber. https://www.corning.com/media/worldwide/coc/documents/Fiber/white-paper/WP1000.pdf.

[87] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).

[88] Olivier Toubia, George Z Gui, Tianyi Peng, Daniel J Merlau, Ang Li, and Haozhe Chen. 2025. Database report: Twin-2k-500: A data set for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science* 44, 6 (2025), 1446–1455.

[89] Mohamed Ali Trabelsi. 2024. The impact of artificial intelligence on economic development. *Journal of Electronic Business & Digital Economics* 3, 2 (2024), 142–155.

[90] Paola Tubaro, Marion Coville, Clément Le Ludec, and Antonio A Casilli. 2022. Hidden inequalities: the gendered labour of women on micro-tasking platforms. *Internet Policy Review* 11, 1 (2022), 1–26.

[91] UNITAR, ITU, and UNU. 2024. *The Global E-waste Monitor 2024*. Technical Report. United Nations Institute for Training and Research; International Telecommunication Union; United Nations University. https://ewastemonitor.info/wp-content/uploads/2024/12/GEM_2024_EN_11_NOV-web.pdf

[92] United Nations Environment Programme. 2024. AI Has an Environmental Problem. Here's What the World Can Do About It. https://www.unep.org/news-and-stories/story/ai-has-environmental-problem-heres-what-world-can-do-about Accessed: 2025-11-13.

[93] U.S. Energy Information Administration. 2024. "Wisconsin: State Profile and Energy Estimates". https://www.eia.gov/state/?sid=WI Accessed: 2026-01-10.

[94] Aimee Van Wynsberghe. 2021. Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics* 1, 3 (2021), 213–218.

[95] Pablo Villalobos and Anson Ho. 2022. Trends in training dataset sizes. https://epoch.ai/publications/trends-in-training-dataset-sizes Accessed: 2025-11-20.

[96] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? Limits of LLM scaling based on human-generated data. In *International Conference on Machine Learning*. https://openreview.net/forum?id=ViZcgDQjyG

[97] Ding Wang, Shantanu Prabhat, and Nithya Sambasivan. 2022. Whose AI Dream? In search of the aspiration in data annotation.. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–16.

[98] Hengbin Wang, Yu Yao, Yuanyuan Zhao, Shaoming Li, Zhe Liu, and Xiaodong Zhang. 2025. Carbon dioxide emissions from global land cover mapping are projected to increase by 2050. *Communications Earth & Environment* 6, 1 (2025), 1018.

[99] World-Economic-Forum and Geneva-Graduate-Institute. 2025. *Trade and Labour: Pathways for Decent Work in Kenya's Digital Economy*. Technical Report. World Economic Forum. White Paper.

[100] Dustin Wright, Christian Igel, Gabrielle Samuel, and Raghavendra Selvan. 2023. Efficiency is not enough: A critical perspective of environmentally sustainable AI. *Commun. ACM* (2023).

[101] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of machine learning and systems* 4 (2022), 795–813.

[102] Xinyu Yang, Weixin Liang, and James Zou. 2024. Navigating dataset documentations in AI: A large-scale analysis of dataset cards on hugging face. *arXiv preprint arXiv:2401.13822* (2024).

[103] Shoshana Zuboff. 2023. The age of surveillance capitalism. In *Social theory re-wired*. Routledge, 203–213.

Appendix

A Hugging Face Datasets Metadata

A.1 Metadata Coverage

Table 1 provides an overview of all metadata attributes and coverage.

Table 1. Overview of metadata attributes. The final sample includes 554,300 datasets.

Attribute	Description	Coverage [%]
`id`	Unique dataset identifier on the Hub	100.0
`created_at`	Timestamp of repository creation	100.0
`last_modified`	Timestamp of the most recent commit	100.0
`downloads_all_time`	Total number of downloads since creation	100.0
`downloads_30d`	Number of downloads in the last 30 days	100.0
`used_storage`	Estimated storage used by the dataset on the Hub (bytes)	91.0
`dataset_size`	Local storage footprint of the dataset (Apache Arrow, bytes)	74.7
`region`	Reported hosting region (US or EU)	99.8
`modality`*	Data modality	78.3
`task`*	High-level task category	12.9
`sub-task`*	Specific task	12.9
`language`*	Languages represented in the dataset (ISO-based codes)	10.2

* indicates that the attribute allows multiple labels.

A.2 Modalities and Tasks

Modalities include text, image, tabular, video, audio, time-series, 3d, document, and geospatial. The tasks include audio-speech (AS), computer vision (CV), multimodal (MM), natural language processing (NLP), reinforcement learning (RL), and tabular (TAB).

Datasets without a declared modality account for 22% of all datasets but represent 51% of total storage volume, even though for 41% of them the dataset size could not be extracted. The dominant grey area in Figure 1 (Centre) therefore corresponds to only 14% of all datasets.

Figure 8 shows the distribution of modality and task combinations, with text datasets dominating the modality landscape. Figure 9 shows the distribution of dataset sizes and total downloads across modalities and tasks.

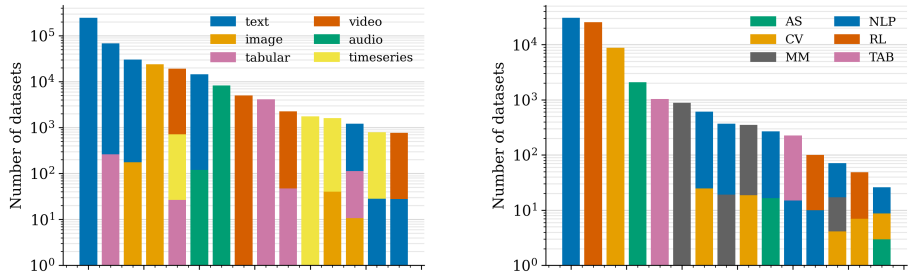

Fig. 8. Distribution of dataset modalities and task categories on the Hugging Face Hub. Left: The fifteen most common dataset modality combinations on a logarithmic scale. Right: The fifteen most common dataset task combinations on a logarithmic scale.

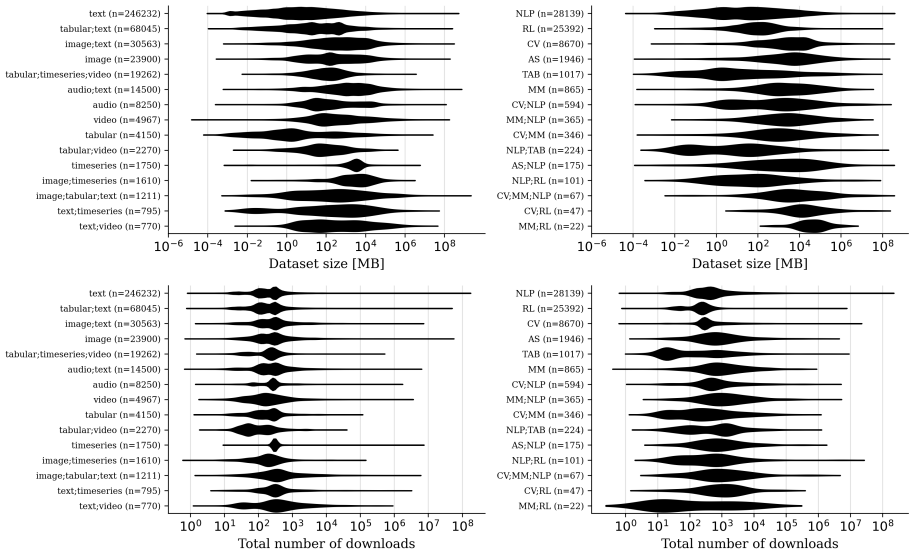

Fig. 9. Distribution of dataset sizes and downloads on the Hugging Face Hub by modality and task. Left: Violin plots of dataset sizes (top) and total downloads (bottom) for the fifteen most common modality combinations. Right: Violin plots of dataset sizes (top) and total downloads (bottom) for the fifteen most common task combinations.

A.3 Language and Download Distribution for Top Ten Languages

To support Figure 6, Table 2 reports language shares on the Hugging Face Hub by dataset volume, compared with Common Crawl page shares and global speaker populations. Table 3 reports the corresponding shares of total downloads and downloads in November 2025 corresponding to the 30 days preceding data collection.

Table 2. Language distribution across Hugging Face (HF) datasets, Common Crawl (CC) pages, and global speaker populations for the top ten language groups by volume.

Language group	HF dataset size (%)	CC pages (%)	Speakers (%)
English	57.4	42.1	18.5
Multi-lang incl. English	29.8	–	–
Multi-lang excl. English	5.2	–	–
Japanese	3.1	5.6	1.5
Chinese	1.0	5.6	14.3
Russian	0.5	6.5	3.1
French	0.4	4.4	3.8
German	0.3	5.7	1.6
Arabic	0.2	0.7	4.1
Spanish	0.2	4.4	6.8

Table 3. Download distribution across Hugging Face datasets for the top ten language groups by volume.

Language group	Downloads, all time (%)	Downloads, last 30 days (%)
English	79.4	67.6
Multi-lang incl. English	11.8	20.1
Multi-lang excl. English	1.0	2.4
Japanese	2.0	1.2
Chinese	1.7	1.9
Russian	0.1	0.3
French	0.2	0.6
German	0.1	0.1
Arabic	0.4	0.9
Spanish	0.1	0.2

B Questionnaire for Data Workers

The questionnaire consisted of ten questions covering demographics, working conditions, and exposure to graphic content. All questions were administered in English. The questionnaire can be seen below.

Employment and Work Characteristics.

- **What type of data work do you typically do? (multiple selections allowed)**
  - Labelling data (images, text, etc.)
  - Content moderation
  - Verification of user preferences (e.g. in chatbots)
  - Data cleaning
  - None of the above
  - Other (free-text)
- **How often are you exposed to graphic content (violence, self-harm, hate speech, sexual abuse etc.) as part of your data work?**
  - Never
  - 1–2 times a day
  - Several times each day

  - Every other day
  - Other (free-text)
- **Have you directly worked (not via intermediate BPOs or platforms) for any of the following big tech companies? (multiple selections allowed)**
  - Google
  - Microsoft
  - Facebook/Meta
  - OpenAI
  - Amazon
  - None of the above
  - Other (free-text)
- **How long have you worked as a data worker?**
  - Less than one year
  - 1–3 years
  - 4–6 years
  - 7–10 years
  - More than 10 years
- **How many hours of data work do you typically do per week?**
  - Less than 10 hours
  - 10–20 hours
  - 20–30 hours
  - 30–40 hours
  - 40–60 hours
  - More than 60 hours
- **What is your average gross monthly salary (before taxes) for doing the data work?**
  - Less than 50 USD
  - 50–100 USD
  - 100–200 USD
  - 200–300 USD
  - 300–400 USD
  - 400–500 USD
  - More than 500 USD

Demographics.

- **Gender:**
  - Female
  - Male
  - Prefer not to say
  - Other (free-text)
- **Age bracket:**
  - Under 20

  - 20–30
  - 30–40
  - Older than 40
- **Which country are you based in?**
  - Free-text

Additional Comments.

- **Any additional points you would like to add? (optional)**
  - Free-text

C Questionnaire Summary Statistics

Responses were preprocessed prior to analysis to ensure consistency across categories. For exposure to graphic content, the questionnaire included four fixed-response options (Never, 1–2 times a day, Several times each day, Every other day). The two daily exposure categories (1–2 times a day and Several times each day) were collapsed into a single category labelled Every day. One free-text response reported "Everyday" which was also mapped to Every day. A small number of free-text responses (e.g. "sometimes", "once in a while") were grouped into a single Sometimes category and excluded from further analysis due to low frequency (4/134).

Free-text responses for type of data work were not included in the analysis. These responses were all unique single mentions (e.g. "AI training", "customer service", "3D").

For employment relationships with large technology companies, free-text responses were used to clarify indirect work arrangements (e.g. work conducted via business process outsourcing (BPO) firms). All such responses were mapped to the predefined category None of the above, reflecting indirect rather than direct employment.

The preprocessed counts are summarised in Table 4.

Table 4. Raw counts for all categorical variables and data work types.

Category	Count	Category	Count	Category	Count	Category	Count
Gender		**Experience (years)**		**Salary per month (USD)**		**Data work type**	
Female	57	<1	5	<50	15	Verification of user preferences	33
Male	77	1–3	44	50–100	12	Data cleaning	52
		4–6	70	100–200	35	Content moderation	67
		7–10	15	200–300	42	Labelling data	120
				300–400	17		
				400–500	12		
				>500	1		
Age		**Hours per week**		**Exposure to graphic content**		**Big tech companies**	
<20	1	<10	7	Never	45	Google	24
20–30	87	10–20	13	Sometimes	4	Microsoft	13
30–40	45	20–30	8	Every other day	25	Facebook/Meta	35
>40	1	30–40	36	Every day	60	OpenAI	51
		40–60	65			Amazon	10
		>60	5				

D Additional Analyses

D.1 Household Electricity Equivalents compared to Data Centres

The projected household electricity equivalent for the Microsoft Fairwater data centre in Wisconsin with 3.33 GW capacity is approximately 3.3 million homes. This figure is derived by calculating the annual energy output of the facility and dividing it by the average yearly electricity consumption of a Wisconsin residence. A 3.33 GW facility operating at a standard 90% load factor which is typical for high-density hyper-scale AI infrastructure, generates roughly $3.33 \times 365 \times 24 \times 0.9 = 26.25$ billion kWh of electricity annually.

Given that the average Wisconsin household consumes approximately 660 kWh per month [93], or roughly 7,920 kWh per year, the total electrical demand of this single data centre is equivalent to the consumption of approximately 3,315,000 households.

Studies also show that data centres in the US emitted over 105 million tCO2eq in 2024, with carbon intensity ≈48% higher than the national average, as many data centres are situated in areas with fossil-heavy grids [31]. This also mirrors the global trend, as the concentration of data centres is higher in fossil-heavy grids such as those in the US and China (see Figure 7).

D.2 Model for Storage-related Emissions

To approximate storage-related emissions, we use a simple lower-bound calculation. Provider-side energy use for dataset i is estimated as

$$E_{\text{prov},i} = \epsilon \times T_i \times S_{\text{prov},i}, \tag{1}$$

where $\epsilon = 60$ kWh TB^{-1} year^{-1} is the assumed storage energy intensity [77], T_i is the time from creation to 1 December 2025 (years), and $S_{\text{prov},i}$ is the used storage size (TB) assumed to be constant. This provides a conservative estimate, as it omits internal replication by cloud providers.

We apply a parallel calculation for user-side storage by assuming that a fraction of downloads result in local copies that persist for some period. We estimate user-side energy for dataset i as

$$E_{\text{user},i} = \epsilon \times f_{\text{stored}} \times T_{\text{user}} \times D_i \times S_{\text{user},i}, \tag{2}$$

with $f_{\text{stored}} = 0.1$ denoting the assumed share of downloads that remain stored, $T_{\text{user}} = 0.25$ years the assumed average retention time, D_i the total number of downloads, which we treat as uniformly distributed over the dataset's lifetime, and $S_{\text{user},i}$ is the local dataset size. These assumptions reflect limited visibility into actual user-side behaviour and should be interpreted as indicative rather than exact. Based on the available attributes in Table 1, 91% of datasets are included for the provider-side estimate and 75% datasets included for the user-side estimate. Emissions are derived using the average US grid intensity of 384 gCO_2eq/kWh [70], equivalent to roughly 23 $kgCO_2eq/TB$.

D.3 Internet Traffic and Data Curation

To contextualise global data flows and digital presence, we draw on regionally disaggregated mobile and fixed broadband traffic statistics from the International Telecommunication Union (ITU) [43]. We use this data to illustrate structural asymmetries in global data generation and circulation that underpin large-scale AI systems.

Digital participation is uneven well before data is curated for AI. Although mobile broadband networks now cover 96% of the global population, around one third of people still do not use the internet [43]. This disparity is reflected in global traffic patterns.

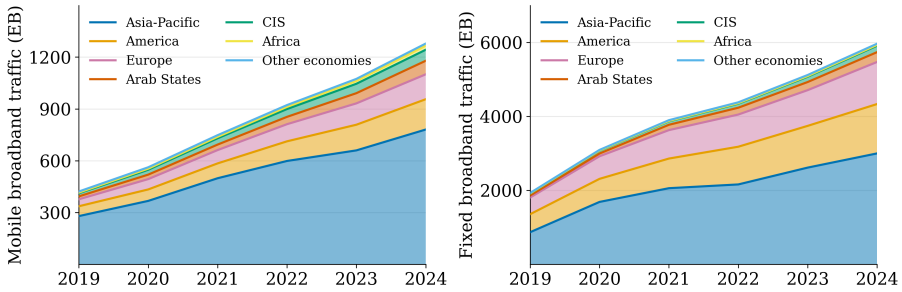

Fig. 10. Mobile and fixed broadband traffic for the Asia-Pacific region, America, Europe, the Arab States, and the Commonwealth of Independent States (CIS) from 2019 to 2024. Data from International Telecommunication Union [43].

As Figure 10 shows, Asia-Pacific, America, and Europe dominate both mobile and fixed broadband traffic, while Africa, the Arab States, and the Commonwealth of Independent States (CIS) generate comparatively little.

Even though, traffic volume is not a direct measure of AI data contribution, it provides a meaningful indication of whose digital activity is more likely to be captured, indexed, or incorporated into large-scale datasets.

D.4 External Data Source Limitations

All external sources are subject to their respective limitations. Financial estimates rely on investigative reporting rather than audited disclosures, and economic relationships between AI developers and cloud providers are complex and evolving. The geographic concentration of data centres captures presence but not size or utilisation. Accordingly, we use these sources to characterise structural dynamics and support qualitative interpretation, rather than to provide exhaustive or definitive quantification.