

How Hyper-Datafication Impacts the Sustainability Costs in Frontier AI



Sophia N. Wilson¹, Sebastian Mair², Mophat Okinyi³, Erik B. Dam¹, Janin Koch⁴, Raghavendra Selvan¹

¹University of Copenhagen, Copenhagen, Denmark

²Linköping University, Linköping, Sweden

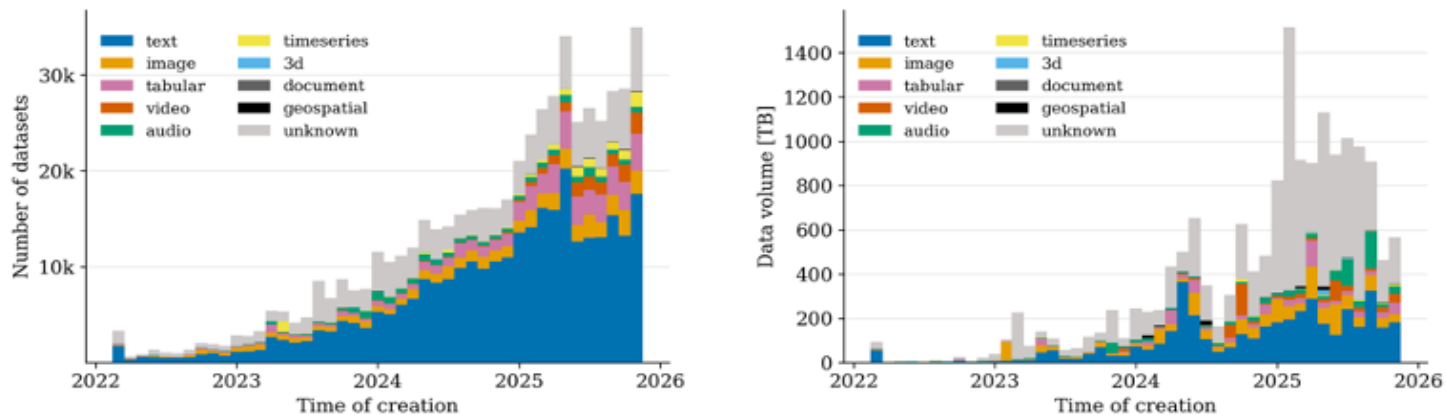
³Techworker Community Africa, Nairobi, Kenya

⁴University of Lille, Lille, France

HYPER-DATAFICATION

Hyper-datafication refers to the industrialised production and accumulation of data for AI model development across three 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.

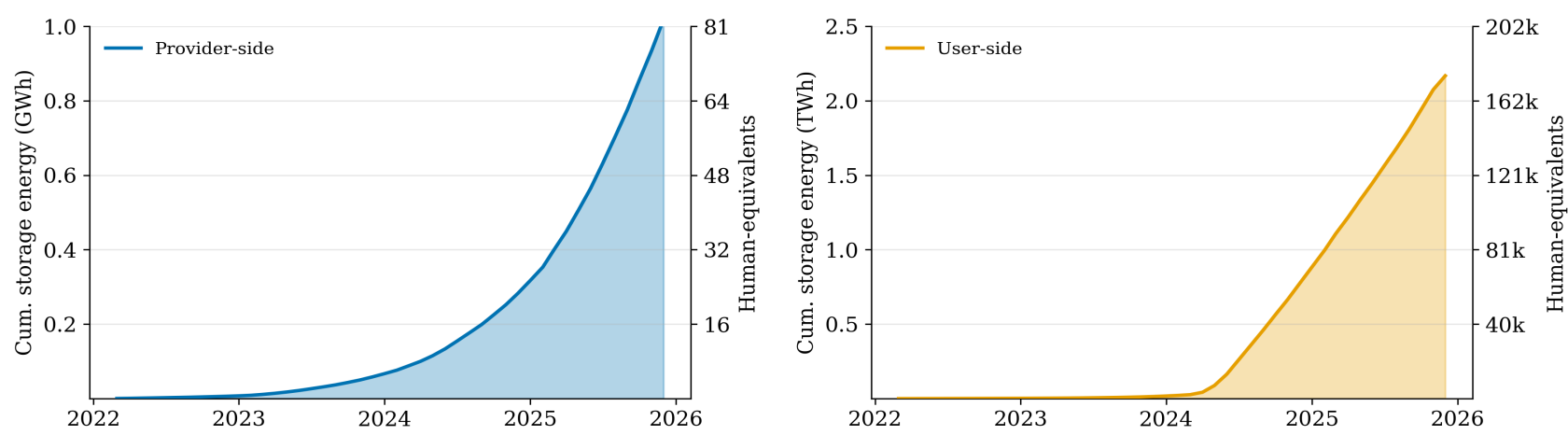
RISE IN DATA VOLUME



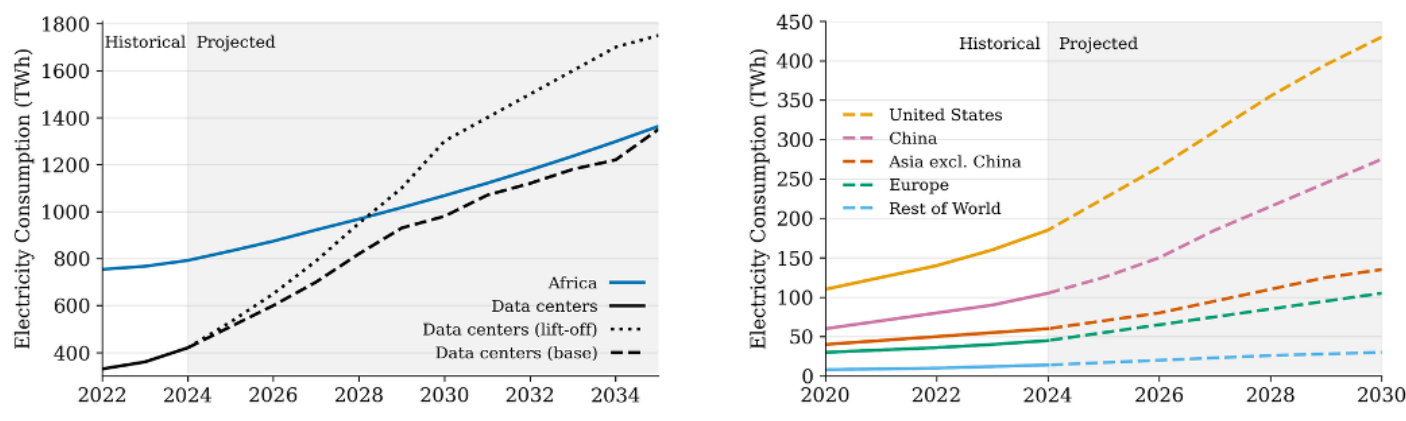
Growth of datasets and data volume on the Hugging Face Hub. The volume of newly added data has increased by nearly three orders of magnitude: from tens of terabytes per month to peak levels exceeding one petabyte per month.

Despite the large volume of data, the top 1% most downloaded datasets (5,543) account for 87% of total downloads.

ENVIRONMENTAL COST: STORAGE-RELATED ENERGY USE



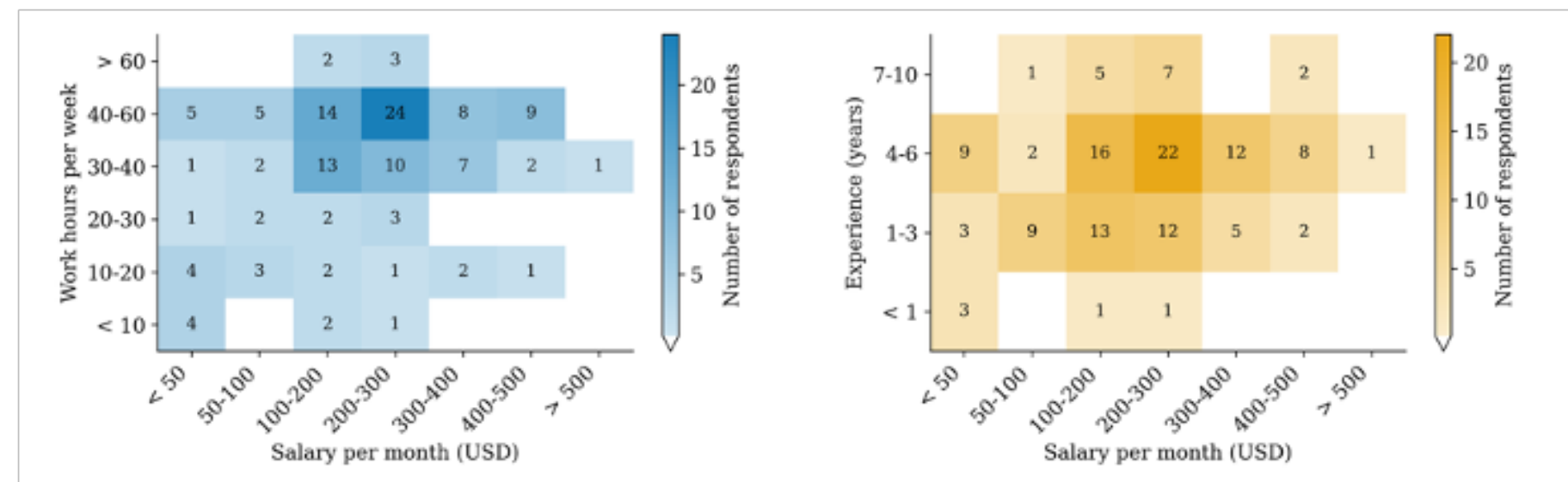
Estimated storage-related energy use of Hugging Face Datasets. For the user-side estimate, it is assumed that 10% of downloads result in three months of local storage. Less than 0.1% of datasets (316) are hosted in the EU, while the rest are hosted in the US. If all datasets were hosted in the EU instead of the US, the associated carbon footprint would be reduced by approximately 38%, due to lower average carbon intensity in the EU (237 gCO₂eq/kWh) compared with the US (384 gCO₂eq/kWh) in 2024 [1].



Global electricity use for data centres. Hyper-datafication and increased adoption of AI drove a 27% increase in global data centre electricity demand between 2022 and 2024 [2].

Projections indicate this consumption will soon rival the total electricity usage of the entire African continent assuming a 5% annual grow [3].

SOCIAL COST: LABOUR WORK



Salary bands by weekly working hours and experience based on responses from 134 Kenyan data workers. Salary is low compared to the Kenyan national average at USD 540 per month [4].

A majority of respondents (89/134) report some level of exposure, with 60 reporting daily exposure and 25 reporting exposure every other day. Our data suggests that higher exposure does not correspond to higher pay.

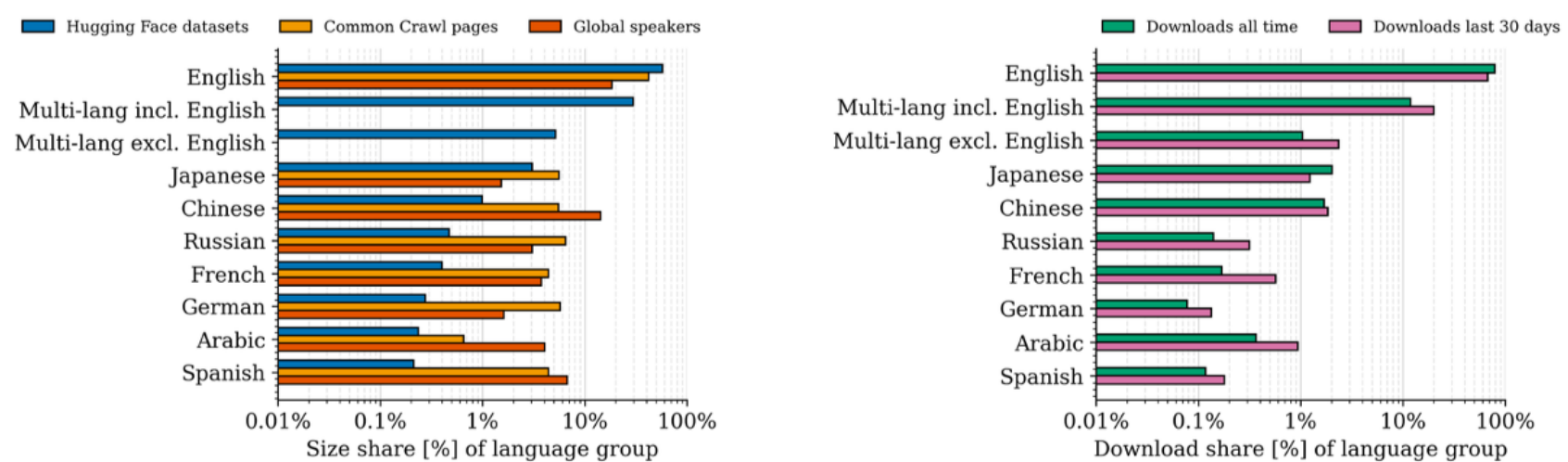
A majority of the respondents (79/134) report that they currently work or have previously worked directly for a major technology company, including OpenAI (51/134), Meta (35/134), Google (24/134), Microsoft (13/134), and Amazon (10/134). Among those who report having worked directly for large technology companies, exposure rates are generally higher.

SOCIAL COST: SOCIETAL REPRESENTATION

Representation and demand for the ten largest language groups on Hugging Face. English and multilingual datasets that include English represent 57% and 30%, respectively, of the total dataset volume.

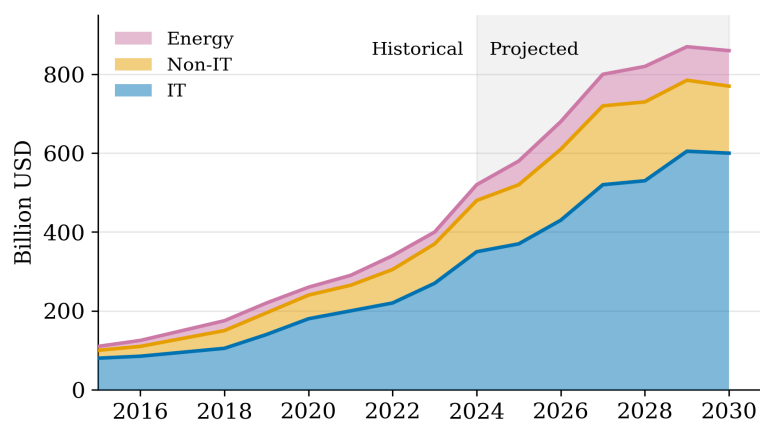
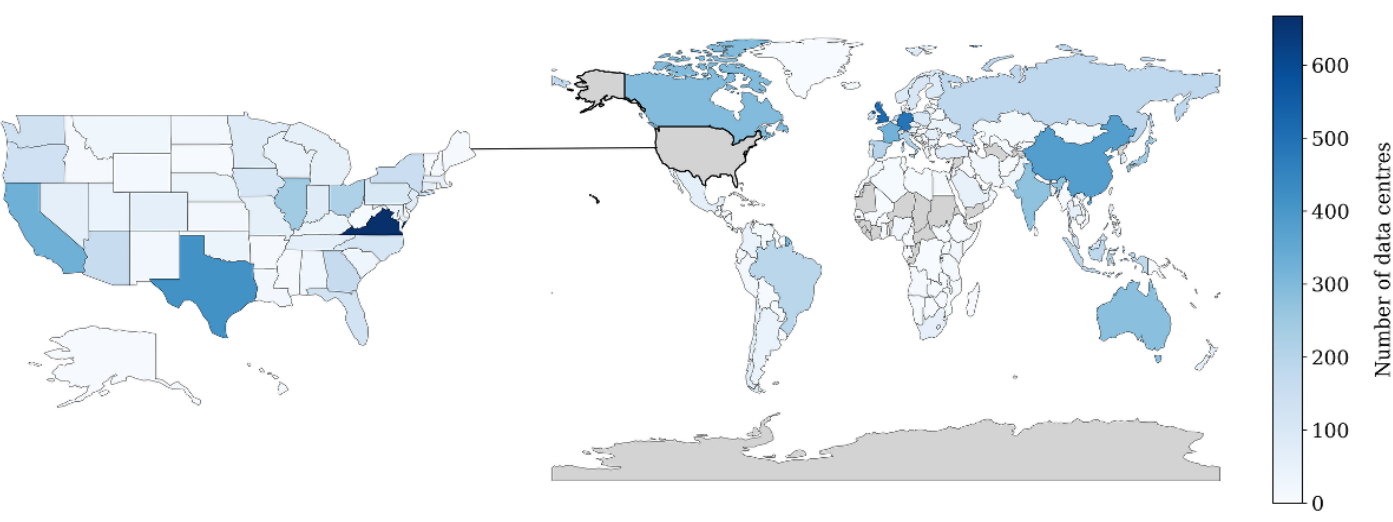
English-only datasets represents a larger share of the dataset volume (57%) than its share of Common Crawl pages (42%) and more than three times its share of global speakers (18%).

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.



ECONOMIC COST: INFRASTRUCTURE INVESTMENTS

Global data centre concentration. Infrastructure development mirrors the investment concentration with facilities clustering in Western Europe, China, Australia, Canada, and particularly the US [5].



Global investment in data centres. In 2025, global investment in data centres was projected to reach USD 580 billion, exceeding the USD 540 billion invested in global oil supply [6].

The US is projected to account for more than half of cumulative global data centre investment over the next five years [7].

DATA PROOFS RECOMMENDATIONS

- **Provenance:** Build data provenance into frontier AI systems, for monitoring data quality and attributing due credit to data producers.
- **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.

REFERENCES

1. Ritchie et al. 2023. Data Page: Carbon intensity of electricity generation. <https://archive.ourworldindata.org/20251014-145858/grapher/carbon-intensity-electricity.html>.
2. International Energy Agency. 2024. Energy Demand from AI. <https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai>.
3. International Energy Agency. 2025. Electricity 2025. <https://www.iea.org/reports/electricity-2025>.
4. CEIC Data. 2024. Kenya Monthly Earnings. <https://www.ceicdata.com/en/indicator/kenya/monthly-earnings>.
5. DataCenterMap. <https://www.datacentermap.com/datacenters/>.
6. International Energy Agency. 2024. World Energy Outlook 2024. <https://iea.blob.core.windows.net/assets/140a0470-5b90-4922-a0e9-838b3ac6918c/WorldEnergyOutlook2024.pdf>.
7. International Energy Agency. 2025. Energy and AI. <https://www.iea.org/reports/energy-and-ai>.