

Stop Preaching and Start Practising Data Frugality for Responsible Development of AI

Sophia N. Wilson¹ Guðrún Fjóla Guðmundsdóttir² Andrew Millard³ Raghavendra Selvan¹ Sebastian Mair³

Abstract

This position paper argues that the machine learning community must move from preaching to practising data frugality for responsible artificial intelligence (AI) development. For long, progress has been equated with ever-larger datasets, driving remarkable advances but now yielding increasingly diminishing performance gains alongside rising energy use and carbon emissions. While awareness of data frugal approaches has grown, their adoption has remained rhetorical, and data scaling continues to dominate development practice. We argue that this gap between preach and practice must be closed, as continued data scaling entails substantial and under-accounted environmental impacts. To ground our position, we provide indicative estimates of the energy use and carbon emissions associated with the downstream use of ImageNet-1K. We then present empirical evidence that data frugality is both practical and beneficial, demonstrating that coreset-based subset selection can substantially reduce training energy consumption with little loss in accuracy, while also mitigating dataset bias. Finally, we outline actionable recommendations for moving data frugality from rhetorical preach to concrete practice for responsible development of AI.

1. Introduction

Concerns about the environmental costs, social impacts, and power asymmetries of ever-larger datasets are increasingly voiced within the machine learning (ML) community as part of calls for responsible artificial intelligence (AI) development (Crawford, 2021). Parallel to these concerns, a growing body of work explicitly argues for scaling down rather than continuously scaling up (Goel et al., 2025; Wang

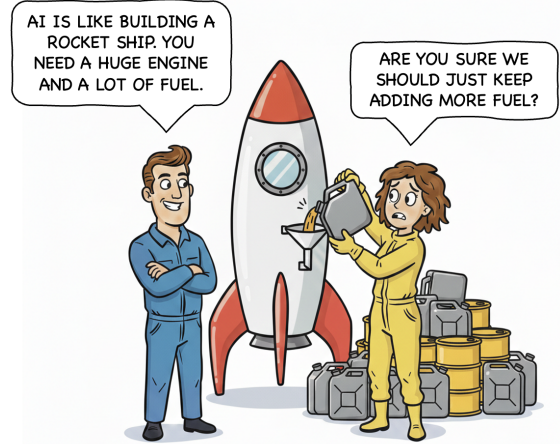


Figure 1. Illustration is inspired by a remark by prominent ML researcher Andrew Ng, comparing an AI model with a rocket ship (quoted in Kelly, 2016). However, progress in rocket science depends less on adding more fuel and more on using it wisely, reflecting the principle of data frugality. Created with Nano Banana (Google, 2026).

et al., 2025). Yet, prevailing research incentives continue to reward scale, with reviewers, benchmarks, and leaderboards incentivising experiments conducted on increasingly large datasets (Ethayarajh & Jurafsky, 2020).

This is particularly evident in frontier AI development, where progress is often equated with ever larger datasets, formalised through empirical scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022). Performance gains are, however, increasingly exhibiting diminishing returns, while computational demand and associated carbon emissions continue to rise (Strubell et al., 2020; Luccioni et al., 2023; Faiz et al., 2024). Moreover, large-scale datasets are known to contain substantial redundancy and noise (Lee et al., 2022; Penedo et al., 2023), raising questions about the efficiency of continued data expansion.

A natural response to these issues is to recognise that not all data are equally informative (Katharopoulos & Fleuret, 2018) and to adopt data frugal approaches that maximise learning efficiency per data sample, reflecting the conceptual message of Figure 1. One prominent example is coreset

¹University of Copenhagen, Denmark ²Technical University of Denmark, Denmark ³Linköping University, Sweden. Correspondence to: Sophia N. Wilson <sophia.wilson@di.ku.dk>.

selection, which constructs smaller yet representative datasets that preserve task performance while reducing training computation (Feldman, 2019). Despite often being preached as a way to reduce energy, many coreset methods do not report the accuracy per unit energy, nor the energy cost of subset construction itself, creating a gap between what is preached and what is practised. Considering a representative sample of ten coreset methods applicable to deep learning¹, the stated motivations for data reduction includes computational efficiency (8/10), energy efficiency (3/10), and reduced storage requirements (2/10). Two papers also mention the democratisation of AI, by enabling participation without access to large-scale computational resources. Despite this, only six papers report time savings and only one evaluates energy savings. The latter, would have been straightforwardly assessed using tools such as Carbon-tracker (Anthony et al., 2020) and CodeCarbon (Courty et al., 2024). These patterns motivate our central position:

The ML community must move from preaching to practising data frugality for responsible development of AI.

To operationalise this shift, we first provide background on the environmental impacts of data in ML, explain how data frugal approaches can limit these impacts, and situate our work within the broader literature (Section 2). We then quantify the aggregate energy use and associated carbon emissions arising from dataset use and storage, making explicit what data frugal approaches aim to reduce (Section 3). Next, we present empirical evidence that data frugality is both practical and beneficial (Section 4). We show that coreset selection can substantially reduce training energy consumption with little loss in accuracy, and that coreset-based dataset curation can mitigate dataset bias. We then critically discuss our position (Section 5) and outline actionable recommendations for how to practise data frugality (Section 6). Finally, we discuss counter-arguments to data frugality (Section 7) and summarise our position (Section 8).

2. Background

We treat data and models as having distinct lifecycles, as illustrated in Figure 2, which motivates a corresponding distinction between data frugality and model frugality, with the former being the focus in this paper.

2.1. Environmental Costs of Data in ML

The environmental costs associated with data have long been overlooked. As a result, data is frequently hoarded (Veritas Technologies LLC, 2018), with more than half of all stored data never accessed or used after its creation (Splunk, 2019).

¹We do not enumerate individual papers, as our goal is to highlight systematic reporting patterns rather than critique specific contributions.

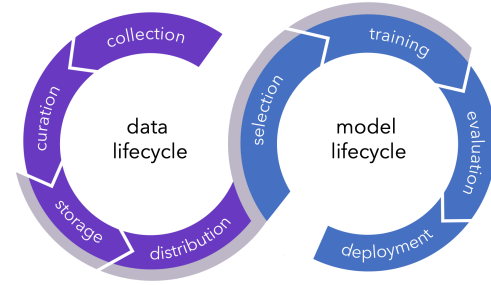


Figure 2. Visualisation of a data lifecycle (purple) and a model lifecycle (blue). Subset selection methods can have positive effects on multiple parts across both lifecycles (grey).

A similar dynamic is evident in ML, where the growing creation and use of large-scale data has led to a form of hyper-datafication in which the associated costs are largely ignored (Wilson et al., 2026).

Within the ML community, concern on the environmental costs has largely centred on model development (Henderson et al., 2020) and deployment (Luccioni et al., 2024) rather than on the data itself. One exception is an analysis of the BLOOM large language model (LLM) by Luccioni et al. (2023), which explicitly accounts for several processes beyond training including data processing, tokenization, and benchmarking. These activities are estimated to contribute 3.3 tonnes of carbon dioxide equivalent (tCO₂e) emissions, corresponding to 5% of total emissions. Another notable study by Common Crawl and Tailpipe² estimates that crawling five billion web pages costs about 326 kgCO₂e, and that relocating the crawl to a lower-carbon grid could reduce emissions by up to 90%. These examples show that data-related processes can represent a non-negligible share of overall system impacts, even when considered only partially.

In practice, environmental costs incur across the full data lifecycle and include both embodied and operational impacts. Embodied impacts arise from the extraction of raw materials, manufacturing, transport, installation, and disposal of hardware and data centre infrastructure. Operational impacts stem from the electricity required to store, transmit, and process data, as well as from associated resource demands such as water use for cooling and land use for data centre construction. The Common Crawl study reports that 2% of their emissions stem from embodied emissions and 98% from operational electricity use. Many of these impacts can, however, be difficult to attribute accurately. For example, network electricity use is often estimated using traffic-proportional averages, even though network energy consumption is largely driven by baseline capacity rather than marginal data transfer, which can lead to substantial overestimation (Mytton et al., 2024).

²<https://tailpipe.ai/measuring-the-carbon-cost-of-crawling-five-billion-web-pages/>

Despite their cumulative scale, data-related energy costs are diffusive and weakly attributed (Mersico et al., 2024), which makes them difficult to quantify. As a result, data is frequently treated as an effectively free input to ML models. Changing this requires making the environmental impacts of data visible across the data lifecycle and explicitly reporting energy metrics. Recent work has therefore proposed attaching the information about these costs directly to datasets, for example as meta-data covering each step of the data lifecycle (Mersy & Krishnan, 2024).

2.2. Data Frugality

The frugal mindset aims to maintain model performance while minimising resource consumption (Evchenko et al., 2021; Violos et al., 2025). In this work, we define **data frugality** as the practice of maximising learning efficiency per data sample by identifying and removing redundant or uninformative samples using subset selection methods.

For many learning problems, dataset size can be reduced substantially without noticeable drops in performance (He et al., 2024; Tan et al., 2025). This can be achieved by computing representative subsets, i.e., small subsets of large datasets that contain the most informative data samples (Feldman, 2019). Models trained on such a subset are expected to achieve performance comparable to those trained on full datasets and subsets with approximation guarantees are often called coresets. Apart from improving storage demands and model training, those representative subsets are also beneficial for model selection. Appendix A contains an overview of coreset selection methods.

By prioritising data quality over data quantity, frugal approaches can surpass classical power-law scaling (Sorscher et al., 2022), while reducing computational demand and environmental impact across multiple stages of the data and model lifecycle (see Figure 2) (Violos et al., 2025; Penedo et al., 2023; He et al., 2024; Verdecchia et al., 2022). Datasets that are small from the outset typically incur lower energy use across the entire data lifecycle, as well as model selection and training, compared to a larger dataset. By contrast, coreset methods primarily reduce energy during model training, and can also lower costs associated with storage, distribution, and model selection. Empirical work shows that coreset approaches can deliver substantial training energy reductions with limited performance loss (Killamsetty et al., 2021; Scala et al., 2025).

The benefits of data reduction extend beyond energy and carbon savings. Smaller datasets lower computational barriers improving accessibility (Ahmed & Wahed, 2020; Rather et al., 2024), enhance reproducibility, facilitate representativeness (Clemmensen & Kjærsgaard, 2022), support privacy-preserving and resource-efficient training in distributed settings (Albaseer et al., 2021; Bian et al., 2024),

and can mitigate ethical and representational harms associated with large-scale datasets have (Birhane & Prabhu, 2021; Birhane et al., 2022).

2.3. Related Work

A growing body of recent position papers challenges the assumption that ever-larger datasets are the primary driver for progress in ML. Instead, they highlight diminishing returns, inefficiencies, and rising environmental impacts.

The cost of data is often overlooked. Kandpal & Raffel (2025) argue that while the growing computational, hardware, energy, and engineering demands of LLMs are frequently considered, the human labour underpinning training data is often overlooked. Books, papers, code, and online content created by humans form the foundation of LLM training corpora. They propose assigning monetary value to this labour and argue that it should constitute the most expensive component of LLM training. They estimate that the monetary value of training data already exceeds model training costs by one to three orders of magnitude. A similar perspective is offered by Jia et al. (2025), who analyse the economic data value chain and argue that value systematically accrues to data aggregators and model developers, while data generators remain largely uncompensated. They advocate for equitable data valuation, provenance, and compensation mechanisms. While these works focus on the economic valuation of data, we instead argue for a more frugal usage of it. Furthermore, Santy et al. (2025) argue that current research incentives reward data volume over care, context, and human labour, leading to datasets that are large yet poor in signal. While their focus is on labour, dignity, and governance, their analysis reinforces our claim: treating data as an abundant, low-cost resource reinforces wasteful and misaligned practices. Data frugality can be read as a technical response that partially realigns incentives by rewarding careful selection over wasteful accumulation.

Current LLM scaling proceeds in the wrong direction.

Wang et al. (2025) argue that scaling within AI development usually means scaling up (data, model parameters, compute), whereas future progress should emphasise scaling down or scaling out, including reductions in carbon footprint. They note that large-scale pretraining has already exhausted much of the publicly available high-quality web content and that the remaining content is either low-quality or AI generated. Thus, continued dataset expansion will no longer result in the same performance gains as earlier. As a remedy, they identify data-efficient training and using smaller, high-quality datasets as future directions. This perspective is also shared by Goel et al. (2025), who advocate downscaling datasets and LLMs to reduce resource demands while maintaining performance. One of their recommendations aligns with ours: adopting a more frugal approach to data.

We are measuring the wrong things. McCoy et al. (2025) explicitly challenge “scaling fundamentalism” and propose capability-per-resource as a more appropriate metric for AI progress. They argue for selective data usage and explicit resource-aware reporting as a way forward. Reporting practices are also central to the arguments of Wilder & Zhou (2025). They argue that research incentives shape behaviour and narrow evaluation metrics distort what gets optimised. Current benchmarks and leaderboards often incentivise data accumulation, making evaluation norms part of the sustainability problem. We largely agree with these critiques and provide concrete recommendations especially for research on data reduction techniques.

AI should be more democratic. Upadhyay et al. (2025) propose a regulatory mandate: frontier labs should release small, open “analogue models” distilled from their large proprietary systems. These analogues can enable research on safety, interpretability, and alignment without exposing full-scale models. While their proposal targets models, we make a complementary argument for smaller, highly informative datasets. Such datasets have lower storage cost, are easier to share, and enable training under limited computational resources.

Our contribution. Overall, these position papers converge on the view that continued reliance on scale alone is increasingly inefficient and unsustainable. Our contribution complements this literature by focusing on data frugality and explicitly addressing the gap between preach and practise. Rather than limiting our contribution to critique and high-level recommendations, we make both dataset costs and the gains achievable through data frugality explicit and provide concrete guidance for translating these insights into practice.

3. Estimating the Energy and Carbon Costs

Estimating the energy and carbon costs of a dataset clarifies what data frugal approaches aim to reduce. In this section, we estimate the downstream energy use of ImageNet from model training and storage, the two lifecycle stages most directly affected by coreset selection.

ImageNet is one of the most influential datasets in computer vision (Deng et al., 2009). While the full dataset comprises over 14 million images and about 21,000 classes (ImageNet-21K), we focus on the most widely used subset, ImageNet-1K, which contains 1.28 million training images spanning 1,000 classes and occupying 130 GB of storage. ImageNet-1K served as the benchmark dataset for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) between 2010 and 2017 and remains a canonical training dataset today (Russakovsky et al., 2015).

3.1. Model Training

To estimate aggregate downstream model training, we first approximate the number of independent training runs performed on ImageNet-1K. We base this estimate on an analysis of all accepted ICLR papers from 2017 to 2022 retrieved via the OpenReview API. We focus on ICLR because, with the exception of 2016, all conference editions use OpenReview for peer review, providing consistent and comprehensive coverage over time. For each year, we estimate the fraction of ImageNet-mentioning papers that report training a model from random initialisation, as opposed to using ImageNet for fine-tuning, evaluation or as a reference. We then extrapolate this fraction to 2023–2025 using linear regression. Assuming this fraction is representative beyond ICLR, we apply it to the broader literature by combining it with a keyword-based count of publications mentioning *imagenet* using dimensions.ai³. This procedure yields an estimated 46,179 training runs on ImageNet-1K between 2017 and 2025. The estimates neglect training prior to 2017, however, the usage during the ILSVRC involved a relatively small number of competing teams compared to the scale of subsequent academic use (Russakovsky et al., 2015).

We estimate the energy consumption of a single training run using a standard reference setup for ImageNet training⁴. Specifically, we consider training a ResNet-50 model on a single NVIDIA A100 GPU within a DGX A100 machine⁵ and adopt 300 epochs as a representative training regime (Wightman et al., 2021). We measure energy consumption using Carbontracker⁶ (Anthony et al., 2020). Under this setup, the average energy use is estimated at 0.394 KWh per training epoch. See Appendix B for details.

The total energy consumption associated with ImageNet-1K training therefore amounts to:

$$46,179 \times 0.394 \text{ kWh/epoch} \times 300 \text{ epochs} = 5.46 \text{ GWh.}$$

Using a global average carbon intensity of 445 gCO₂e/kWh (IEA, 2025), this corresponds to approximately 2429 tCO₂e, equivalent to the annual carbon footprint of around 514 people based on a global per-capita average of 4.73 tCO₂e (Ritchie et al., 2023b).

3.2. Data Storage

We estimate storage-related energy consumption by assuming that each training run corresponds to one retained local copy of ImageNet-1K. Using a dataset size of 130 GB and an annual energy intensity of 60 kWh/TB/year (Selvan, 2025),

³<https://app.dimensions.ai>

⁴<https://github.com/pytorch/examples/tree/main/imagenet>

⁵<https://docs.nvidia.com/dgx/dgxa100-user-guide>

⁶<https://github.com/saintslab/carbontracker>

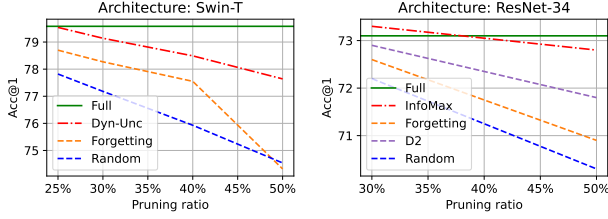


Figure 3. Top-1 accuracy of pruning ImageNet-1K using Dyn-Unc (left), D2 (right), and InfoMax (right). Forgetting, Random, and the performance on the full ImageNet-1K are shown as references. The numbers are taken from He et al. (2024) and Tan et al. (2025). Note that the architectures differ, there is no performance loss for 25%-30% of data pruning when using Dyn-Unc/InfoMax, and that the authors do not compare their methods against each other.

the energy required to store these copies for one year is:

$$46,179 \times 130\text{GB} \times 60 \text{ kWh/TB/year} = 360 \text{ MWh.}$$

This corresponds to approximately 160 tCO₂e, or the annual carbon footprint of around 34 people.

Importantly, the estimates for both model training and storage represent strict lower bounds. As of 1 December 2025, the Hugging Face Hub alone stored 876 versions, copies, and derivatives of ImageNet⁷, as shown in Figure 7 in Appendix B. Of these, 214 explicitly include *imagenet-1k* in their dataset identifier. Summing the downloads of these 214 datasets alone yields over 2.5 million downloads, a factor 55 larger than our estimate of distinct training runs. This indicates, that the true scale of ImageNet-1K usage substantially exceeds what is captured by publication-based counts.

However, even under these conservative assumptions, downstream use of ImageNet-1K alone entails substantial energy consumption and associated carbon emissions, comparable to the annual emissions of hundreds of individuals, underscoring that dataset-level impacts warrant explicit consideration in environmental assessments of AI systems.

4. Gains from Data Frugal Approaches

We now present evidence on how coreset approaches can provide positive effects on time and energy demands of model training as well as on mitigating dataset biases without sacrificing accuracy.

4.1. Pruning Effects on Accuracy

Following our analysis on the energy use and carbon emissions associated with downstream use of ImageNet-1K, we next report state-of-the-art (SOTA) results for coreset selection techniques on this specific dataset. Figure 3 presents results across multiple pruning ratios, evaluated on two ar-

Table 1. Training time (minutes) and energy consumption (kWh) per epoch on full ImageNet-1K and a 25% pruned subset sampled uniformly at random for various architectures using a single NVIDIA A100 GPU. The estimates are averaged over 30 epochs.

Model	Param.	minutes/epoch			kWh/epoch		
		Full	25% pruned	Gain	Full	25% pruned	Gain
ResNet-34	21.8M	35.2	23.8	32%	0.2798	0.1989	29%
ResNet-50	25.6M	40.7	24.3	40%	0.3940	0.2645	33%
Swin-T	28.3M	58.7	44.6	24%	0.7002	0.5300	24%

chitectures: Swin Transformer (Liu et al., 2021) and ResNet-34. The results for Swin-T are from He et al. (2024), who propose the coreset method Dyn-Unc while the results for ResNet-34 are from Tan et al. (2025), who introduce the coreset approach InfoMax. Note that He et al. (2024) and Tan et al. (2025) do not compare their coreset approaches against each other and use different architectures. To put the data pruning results into perspective, we also report on the performance that can be achieved using the full data set (0% pruning) and using a subset sampled uniformly at random. In addition, the performance of the coreset techniques forgetting scores (Toneva et al., 2019) and \mathbb{D}^2 pruning (Maharana et al., 2024) are provided.

As seen in Figure 3, all coreset approaches perform better than a random subset (blue line) as they yield higher accuracies per pruning ratio. Notably, Dyn-Unc (He et al., 2024) (red line, left plot) can prune 25% of both ImageNet-1K and ImageNet-21K (not shown) without any drop in Top-1 accuracy (compare against green line). Similar results can be observed for InfoMax (Tan et al., 2025) (red line, right plot), which can prune up to 35% of ImageNet-1K without a sacrifice in performance. This shows that large datasets such as ImageNet-1K can be pruned with no to little drops in performance.

4.2. Pruning Effects on Time and Energy Demands

Next, we train classifiers based on ResNet-34, ResNet-50, and Swin-T architectures on ImageNet-1K three times for ten epochs using the same experimental setup as in Section 3. Our goal is to estimate the computation time and energy consumption (CPU and GPU) per epoch for the full and a pruned version. Table 1 depicts the results. Pruning 25% of the dataset reduces approximately 32% of the training time and 29% of the energy consumption per epoch when using a ResNet-34 backbone. Note that these substantial reductions are possible without noticeable performance losses using SOTA data reduction techniques (see Figure 3). One caveat is that we neglect the computational burden of constructing the coreset. However, this is a one-time cost. In addition to the benefits of model training, a 25% dataset reduction has positive effects on the energy required for storage.

⁷<https://huggingface.co/datasets?sort=trending&search=imagenet>

4.3. Bias and Representation Effects

Data frugality principles urges us to choose a smaller, representative subset from a larger dataset. As seen above, this can offer a compelling trade-off between performance and energy consumption. Coupling coreset curation with other objectives than dataset coverage can improve other properties of the dataset (Chen & Selvan, 2025). For instance, any model that is trained on a biased dataset, wherein a single majority group is over-represented compared to other minority groups, can also learn to reproduce these biases in its predictions (O’Neil, 2017). Coresets can be selected to mitigate these biases by balancing the samples between groups so that no single group is over-represented at the expense of other groups (Barocas et al., 2023). This is particularly effective if the data collection itself cannot be corrected due to systemic problems and/ or historical reasons (Wilson et al., 2026). Coresets offer an algorithmic remedy to mitigate known biases.

We illustrate this using a toy example based on a coloured version of the MNIST dataset⁸. We associate each digit with a majority colour, while the remaining samples are randomly coloured, as shown in Figure 8 in Appendix C. An ideal classifier should not focus on the colour of the digits, but instead should focus on other relevant features like shape.

Figure 4 shows the influence of different data budgets on classification performance. We report the *aligned* accuracy which is the best accuracy possible if the models do not focus on colours, whereas the *conflicted* accuracy is shown for the case where 99% of samples belong to a majority group and 1% of the samples belong to the minority group. The baseline method is random sampling, which is compared with two other data sampling methods. In the reweighted sampling method, the loss of data points are reweighted inversely proportionally to their groups, i.e., majority group samples are under-weighted. In the balanced sampling, the number of samples between the majority and minority groups are rebalanced so that the bias is removed within the coreset. In budget regimes where the minority samples are fewer than the required budget, we augment the data samples. The trends in Figure 4 capture the usefulness of curating coresets by altering the underlying data distribution which can remove known biases in the dataset. See additional results for different bias strengths in Appendix C.

5. Discussion

Disentangling compute, energy, and carbon efficiency.

A common misconception is that improvements in compute efficiency directly translate to energy efficiency, and

⁸See <https://github.com/kakaoenterprise/Learning-Debiased-Disentangled> and <https://ieee-dataport.org/documents/colored-mnist>

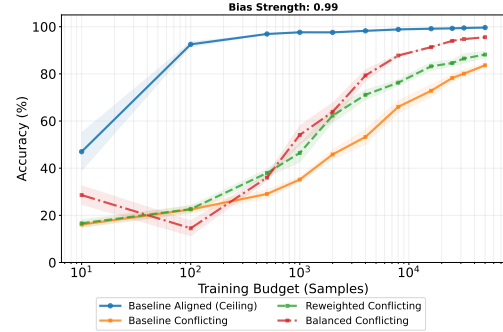


Figure 4. Performance of a classifier trained on the biased Colour-MNIST (99% majority group) dataset is reported for the case with no bias (aligned) and with bias (conflicting). Three coreset methods are shown: random (baseline), reweighted, and balanced.

that gains in energy efficiency in turn imply proportional reductions in carbon emissions. However, this is not necessarily true, see for example Wright et al. (2025). This distinction is also evident in our experimental results (Table 1, Section 4.2). Reducing the size of ImageNet-1K by 25% does not yield a corresponding 25% reduction in time or energy use. Instead, we observe time reductions ranging from 24-40% and energy reductions from 24-33%. These results illustrate that dataset size, training time, and energy consumption cannot be treated as interchangeable proxies. Assuming proportional reductions based solely on data pruning is therefore misleading. Time and energy must be measured and reported explicitly if efficiency claims are to be meaningful.

Alternatives to data frugality. Alternatives to data frugality emphasise efficiency gains at the level of models or hardware rather than datasets. Advances in model architectures, training algorithms, hardware acceleration, and scheduling can mitigate some of environmental impacts without constraining the amount of data used, thereby preserving the benefits of large and diverse datasets. In practice, these interventions can be easier to adopt because they can be implemented unilaterally, whereas using smaller datasets can conflict with comparability and benchmarking.

Closing the gap. Across the literature, efficiency and energy considerations are frequently invoked to motivate data reduction, yet the corresponding metrics are often not reported (as exemplified in Section 1). This gap is not technical, as energy consumption and associated carbon emissions can be straightforwardly tracked using lightweight tools such as Carbontracker (Anthony et al., 2020) and CodeCarbon (Courty et al., 2024). Making such reporting routine would be a first practical step towards aligning stated motivations with actual practice.

Quantifying the full environmental impacts of the full data lifecycle is challenging and requires development of new tools and standards. Dataset creators mention limited re-

sources, including computational constraints, as a challenge during dataset creation (Orr & Crawford, 2024). When estimates of the compute used in dataset creation are available, reporting them would therefore be a natural extension of existing documentation and would improve transparency without imposing substantial additional burden.

Finally, concerns about comparability and reporting burden are not unique to energy and carbon metrics. Similar arguments have long been used in debates around code release and runtime in benchmarking. Acknowledging these challenge is important, but they should not preclude progress. Encouraging incremental transparency through reporting and community norms is a necessary step towards practising responsible AI development.

Costs, limitations and open challenges of coresets. While coreset methods can reduce computational costs, these gains come with trade-offs. Coreset construction introduces additional computational overhead and does not reduce impacts incurred during data collection or curation. This upfront cost is only amortised through reuse, limiting its effectiveness in one-off training scenarios. Moreover, removing data points can have unintended consequences that can be difficult to assess in advance (see Section 7). Coreset methods also face methodological limitations. Most are designed for a specific objectives, such as classification losses or gradient alignment (see Appendix A). This is often sufficient for discriminative learning tasks such as ImageNet classification but less suitable for generative modelling, where preserving fine-grained structure, rare modes, and long-tail variation is critical. As a result, subset that preserves classification accuracy may be inadequate for training generative models such as diffusion models, although, research in that direction exists (Chen & Selvan, 2025). More generally, coreset quality can be sensitive to architectural choices (see Section 4.1), optimisation settings, and data augmentations, and aggressive data pruning may amplify bias.

Limitations of our paper. Our estimates in Section 3 omit unpublished training runs, broader lifecycle impacts beyond electricity use, and ImageNet-related usage beyond the ImageNet-1K subset. The scope and limitations of these estimates are further elaborated in Appendix B. Moreover, all experiments focus on image datasets and may not generalise for tokenised or multimodal datasets used in frontier models.

6. Call to Action

Our position has pointed out the value-action gap between the preaching and practising of data frugality. Such value-action gap is a common occurrence in activism and has been studied extensively in the context of climate action (Blake, 1999). There are several reasons for the value-action gap to persist; primary of these is the *relatively* low effort required

to ideate compared to tangible action which can also be limited due to external constraints. We next present a call to action that aims to *incentivise* relevant stake-holders to pursue data frugality for the responsible development of AI.

People. Individuals have the most immediate agency to challenge the status quo and shift practice away from data scaling by adopting data frugality in their own research choices and workflows.

- *Measure and report data frugality:* Report resource consumption (and savings) due to data frugality to increase awareness of data-related costs, alongside model-related costs.

- *Performance per data-point as metric of success:* Transition from aggregate performance to “Data-Pareto” reporting. This shifts the focus from performance-only to performance relative to dataset size or individual dataset contribution.

- *Share data with care:* Dataset creators should consider curating frugal datasets before publishing them by removing redundant samples. Applying coreset methods at creation time can reduce wasteful storage, transmission, and computation.

- *Measure what you motivate:* When practitioners motivate their approach with something that is measurable, for example energy consumption, it should be evaluated.

Platforms. Data storage and benchmarking platforms, ML communities, and conferences can encourage and reward data frugality by shaping everyday research practice through submission requirements, leaderboards, and recognition mechanisms.

- *Make resource reporting easy and mandatory:* Adopt mechanisms such as CVPR’s mandatory compute reporting form and accompanying awards for recognition⁹.

- *Resource-aware benchmarking:* Maintain canonical benchmark datasets with streamed access and persistent leaderboards to reduce redundant downloads, storage, and repeated baseline training. This should be complemented by metrics such as kWh to generalise across hardware and time.

- *Incentivise data-efficient research:* Host challenges such as Energy Efficient Image Processing¹⁰ or BabyLM¹¹ that seek and reward data-efficient approaches.

- *Reward responsible practice:* Recognise responsible resource, and transparent environmental reporting.

Policies. Institutions, organisations, and governments can institutionalise change by defining standards and reporting requirements and by investing in shared infrastructure.

- *Standardise reporting:* Prescribe clear standards for reporting resource use and environmental impact due to data.

- *Encourage shared dataset storage:* Reduce redundant lo-

⁹<https://cvpr.thecvf.com/Conferences/2026/ComputeReporting>

¹⁰Hosted at MICCAI <https://e2mip.org/>

¹¹<https://babylm.github.io/>

cal copies through shared infrastructure¹², which can host common datasets, discouraging users from storing their own copies, and enforcing efficient usage policies to discourage wasteful computations.

- *Fund and maintain common infrastructure*: Funding agencies should support the development of shared data-related infrastructure, and mandate using them.

- *Data usage and data sunset laws*: Using large-scale data should require justification and approval as is common when using health data¹³, along with the destruction of data after a set period to reduce accumulation of dark data.

7. Alternative Views

Benefits of data scaling. The most common counter-position to data frugality advocates for continued data scaling. Proponents argue that empirical scaling laws demonstrate consistent performance gains with increasing dataset size, making further data expansion a reliable driver of progress (Kaplan et al., 2020; Hoffmann et al., 2022). From this perspective, data reduction may risk slowing innovation or prematurely constraining model capabilities.

Large datasets generally also improve robustness to distribution shift (Hendrycks et al., 2020; Liu et al., 2025) and increase coverage of rare or long-tail events that may be critical for safety and reliability (Van Horn & Perona, 2017). Consequently, data reduction can remove rare but important samples (Dharmasiri et al., 2025). In safety-critical domains, such as autonomous driving, large and diverse datasets are often necessary for capturing edge cases and ensuring robust behaviour (Lei et al., 2025; Kalra & Paddock, 2016).

Rebound effects. Another counter-position holds that efficiency gains do not necessarily yield proportional reductions in overall resource use. Lower costs from more efficient methods may instead incentivise increased scale, frequency of use, or broader deployment, offsetting expected savings through rebound effects (Morand et al., 2025; Wright et al., 2025; Luccioni et al., 2025). In the context of smaller datasets, this may manifest as an increase in the number of experiments conducted rather than a net reduction in energy use. This perspective reinforces the importance of aligning efficiency gains with intentional changes in practice.

When sustainability arguments undermine democratisation. A further concern is that sustainability arguments may be misused to legitimise reduced access or increased centralisation. Zimmermann et al. (2025) warn that calls for restraint can unintentionally reinforce existing power asymmetries if they are formed as reasons to limit participation rather than to redesign systems. Our position explicitly rejects this interpretation: data frugality aims to expand par-

ticipation by enabling meaningful experimentation under constrained resources, not to justify exclusion.

Taken together, these alternative perspectives highlight that data frugality is neither universally applicable, nor sufficient on its own, but must be situated within a broader understanding of task requirements, safety constraints, and systematic incentives.

8. Conclusion

Recent critiques of data scaling have successfully documented its environmental, social, and systemic costs, yet these insights have largely remained at the level of diagnosis. In this position paper, we argue that the persistent gap between *preaching* and *practising* data frugality is no longer justified. By making dataset-level energy and carbon costs explicit, and by demonstrating that representative subset selection can substantially reduce storage, training time, and training energy with little loss in performance, even mitigating bias, we show that data frugality is both technically feasible and practically beneficial today.

At the same time, we acknowledge that data frugality is not a universal remedy. Its effectiveness depends, e.g., on task requirements. Nevertheless, treading data as an abundant input is increasingly problematic. We therefore call for a shift in norms and evaluation practices that make data-related costs visible and reward careful data stewardship. Moving toward responsible AI development requires not only better models, but also better choices about the data we create, store, and (re-)use. Practising data frugality is a concrete step in that direction.

Environmental Impact Statement

We track the energy consumption of our experiments using carbontracker¹⁴ (Anthony et al., 2020). The analysis in Section 4.1 does not incur any noteworthy energy consumption as we aggregate numbers taken from He et al. (2024) and Tan et al. (2025). The experiments in Section 4.2 take 113.66 hours and incur a total energy consumption of 71.02 kWh which is about 33.59 kgCO₂e.¹⁵ The results of this experiments are reused for the experiment in Section 3.1. The energy consumption associated with using ChatGPT in the ImageNet analysis (see Appendix B) remains unknown. Moreover, the experiments reported in Section 4.3 incur a total energy consumption of 0.24 kWh, which is about 114 gCO₂e. This yields a total of 33.71 kgCO₂e for this position paper.

¹⁴<https://github.com/saintslab/carbontracker>

¹⁵We use the global average carbon intensity of 473 gCO₂e/kWh for 2024 (Ritchie et al., 2023a). The actual numbers corresponding to the locations of our compute are retracted to facilitate anonymity and will be inserted in the camera-ready version of this paper.

¹²Like the national projects such as Sweden’s Berzelius cluster.

¹³Such as UK Biobank or Human Genome Project.

Acknowledgments

SW and RS acknowledge funding from the Independent Research Fund Denmark (DFF) under grant agreement No. 4307-00143B. RS further acknowledges funding from the European Union’s Horizon Europe Research and Innovation Action programme under grant agreements No. 101070284, No. 101070408 and No. 101189771. SM acknowledges funding from the Danish Data Science Academy (DDSA) through DDSA Visit Grant No. 2025-5673. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

Generative AI Usage Statement

ChatGPT version 5.1 was used to support programming tasks, including the development of scripts for accessing the openreview API and data visualisation. ChatGPT versions 5.1 and 5.2 were used to analyse the ImageNet usage of ICLR papers and to edit and refine language and grammar in selected sections of the manuscript. Google Gemini was used to identify relevant academic work.

References

- Ahmed, N. and Wahed, M. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*, 2020.
- Albaseer, A., Abdallah, M., Al-Fuqaha, A., and Erbad, A. Fine-grained data selection for improved energy efficiency of federated edge learning. *IEEE Transactions on Network Science and Engineering*, 9(5):3258–3271, 2021.
- Anthony, L. F. W., Kanding, B., and Selvan, R. Carbon-tracker: Tracking and predicting the carbon footprint of training deep learning models. ICML Workshop on Challenges in Deploying and Monitoring Machine Learning Systems, July 2020. *arXiv:2007.03051*.
- Bachem, O., Lucic, M., and Krause, A. Scalable k-means clustering via lightweight coresets. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1119–1127. ACM, 2018.
- Barocas, S., Hardt, M., and Narayanan, A. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.
- Bian, J., Wang, L., Ren, S., and Xu, J. Cafe: Carbon-aware federated learning in geographically distributed data centers. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, pp. 347–360, 2024.
- Birhane, A. and Prabhu, V. U. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1536–1546. IEEE, 2021.
- Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 173–184, 2022.
- Blake, J. Overcoming the ‘value-action gap’ in environmental policy: Tensions between national policy and local experience. *Local environment*, 4(3):257–278, 1999.
- Campbell, T. and Beronov, B. Sparse variational inference: Bayesian coresets from scratch. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Chen, T. and Selvan, R. A discrepancy-based perspective on dataset condensation. *arXiv preprint arXiv:2509.10367*, 2025.
- Clemmensen, L. H. and Kjærsgaard, R. D. Data representativity for machine learning and ai systems. *arXiv preprint arXiv:2203.04706*, 2022.
- Coleman, C., Yeh, C., Mussmann, S., Mirzasoleiman, B., Bailis, P., Liang, P., Leskovec, J., and Zaharia, M. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020. URL <https://openreview.net/forum?id=HJg2b0VYDr>.
- Courty, B., Schmidt, V., Luccioni, S., Goyal-Kamal, MarionCoutarel, Feld, B., Lecourt, J., LiamConnell, Saboni, A., Inimaz, supatomic, Léval, M., Blanche, L., Cruveiller, A., ouminasara, Zhao, F., Joshi, A., Bogroff, A., de Lavoreille, H., Laskaris, N., Abati, E., Blank, D., Wang, Z., Catovic, A., Alencon, M., Stęchły, M., Bauer, C., de Araújo, L. O. N., JPW, and MinervaBooks. mlco2/codecarbon: v2.4.1, May 2024. URL <https://doi.org/10.5281/zenodo.11171501>.
- Crawford, K. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.
- Cui, J., Wang, R., Si, S., and Hsieh, C.-J. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pp. 6565–6590. PMLR, 2023.

- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.
- Dharmasiri, A., Yang, W., Kirichenko, P., Liu, L., and Rusakovsky, O. The impact of coreset selection on spurious correlations and group robustness. *arXiv preprint arXiv:2507.11690*, 2025.
- Ethayarajh, K. and Jurafsky, D. Utility is in the eye of the user: A critique of nlp leaderboards. *arXiv preprint arXiv:2009.13888*, 2020.
- Evchenko, M., Vanschoren, J., Hoos, H. H., Schoenauer, M., and Sebag, M. Frugal machine learning. *arXiv preprint arXiv:2111.03731*, 2021.
- Faiz, A., Kaneda, S., Wang, R., Osi, R. C., Sharma, P., Chen, F., and Jiang, L. LLMCarbon: Modeling the end-to-end carbon footprint of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=aIok3ZD9to>.
- Fayyaz, M., Aghazadeh, E., Modarressi, A., Pilehvar, M. T., Yaghoobzadeh, Y., and Kahou, S. E. Bert on a data diet: Finding important examples by gradient-based pruning. *arXiv preprint arXiv:2211.05610*, 2022.
- Feldman, D. Core-sets: Updated survey. In *Sampling Techniques for Supervised or Unsupervised Tasks*, pp. 23–44. Springer, 2019.
- Geifman, Y. and El-Yaniv, R. Deep active learning with a neural architecture search. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- Goel, Y., Sengupta, A., and Chakraborty, T. Position: Enough of scaling LLMs! lets focus on downscaling. In *International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=CYJlJgEzZs>.
- Har-Peled, S. and Mazumdar, S. On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, pp. 291–300, 2004.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.
- He, M., Yang, S., Huang, T., and Zhao, B. Large-scale dataset pruning with dynamic uncertainty. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7713–7722, 2024.
- Henderson, P., Hu, J., Romoff, J., Brunskill, E., Jurafsky, D., and Pineau, J. Towards the systematic reporting of the energy and carbon footprints of machine learning. *Journal of Machine Learning Research*, 2020.
- Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., and Song, D. Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*, 2020.
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., et al. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030, 2022.
- IEA. Electricity 2025. <https://www.iea.org/reports/electricity-2025>, 2025. Accessed: 2025-09-20.
- Jia, R., Oala, L., Xiong, W., Ge, S., Wang, J. T., Kang, F., and Song, D. 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*, 2025. URL <https://openreview.net/forum?id=mdKzkjYldM>.
- Kalra, N. and Paddock, S. M. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation research part A: policy and practice*, 94:182–193, 2016.
- Kandpal, N. and Raffel, C. Position: The most expensive part of an LLM should be its training data. In *International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=L6RpQ1h4Nx>.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Katharopoulos, A. and Fleuret, F. Not all samples are created equal: Deep learning with importance sampling. In *International Conference on Machine Learning*, pp. 2525–2534. PMLR, 2018.
- Kelly, K. *The Inevitable: Understanding the 12 Technological Forces That Will Shape Our Future*. Viking Press, New York, 2016. Quote attributed to Andrew Ng.
- Killamsetty, K., Durga, S., Ramakrishnan, G., De, A., and Iyer, R. Grad-match: Gradient matching based data subset selection for efficient deep model training. In *International Conference on Machine Learning*, pp. 5464–5474. PMLR, 2021.

- Killamsetty, K., Abhishek, G. S., Lnu, A., Ramakrishnan, G., Evfimievski, A. V., Popa, L., and Iyer, R. K. AUTOMATA: Gradient based data subset selection for compute-efficient hyper-parameter tuning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=ajH17-Pb43A>.
- Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL <https://aclanthology.org/2022.acl-long.577/>.
- Lei, Y., Shan, C., and Ge, W. Edgecasednet: An enhanced detection architecture for edge case perception in autonomous driving. *Plos one*, 20(12):e0338638, 2025.
- Liu, Y.-A., Zhang, R., Guo, J., de Rijke, M., Fan, Y., and Cheng, X. On the scaling of robustness and effectiveness in dense retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2351–2360, 2025.
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.
- Luccioni, A. S., Viguier, S., and Ligozat, A.-L. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24 (253):1–15, 2023.
- Luccioni, A. S., Strubell, E., and Crawford, K. From efficiency gains to rebound effects: The problem of jevons’ paradox in ai’s polarized environmental debate. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 76–88, 2025.
- Luccioni, S. and Crawford, K. The nine lives of imagenet: A sociotechnical retrospective of a foundation dataset and the limits of automated essentialism. *Journal of Data-centric Machine Learning Research*, 2024.
- Luccioni, S., Jernite, Y., and Strubell, E. Power hungry processing: Watts driving the cost of ai deployment? In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 85–99, 2024.
- Lucic, M., Bachem, O., and Krause, A. Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. In *Artificial Intelligence and Statistics*, 2016.
- Maharana, A., Yadav, P., and Bansal, M. \mathbb{D}^2 pruning: Message passing for balancing diversity & difficulty in data pruning. In *The Twelfth International Conference on Learning Representations*, 2024. URL <https://openreview.net/forum?id=thbtoAkCe9>.
- Mair, S. and Brefeld, U. Coresets for archetypal analysis. In *Advances in Neural Information Processing Systems*, pp. 7247–7255, 2019.
- Mair, S., Boubekki, A., and Brefeld, U. Frame-based data factorizations. In *International Conference on Machine Learning*, pp. 2305–2313. PMLR, 2017.
- Marion, M., Üstün, A., Pozzobon, L., Wang, A., Fadaee, M., and Hooker, S. When less is more: Investigating data pruning for pretraining LLMs at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023. URL <https://openreview.net/forum?id=XUIYn3jo5T>.
- McCoy, D., Wu, Y., and Butzin-Dozier, Z. AI progress should be measured by capability-per-resource, not scale alone: A framework for gradient-guided resource allocation in LLMs. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL <https://openreview.net/forum?id=6plSmhBI33>.
- Mersico, L., Abroshan, H., Sanchez-Velazquez, E., Saheer, L. B., Simandjuntak, S., Dhar-Bhattacharjee, S., Al-Haddad, R., Saeed, N., and Saxena, A. Challenges and solutions for sustainable ict: The role of file storage. *Sustainability*, 16(18):8043, 2024.
- Mersy, G. and Krishnan, S. Toward a life cycle assessment for the carbon footprint of data. *ACM SIGENERGY Energy Informatics Review*, 4(5):25–33, 2024.
- Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.
- Morand, C., Ligozat, A.-L., and Névéol, A. The environmental impacts of machine learning training keep rising evidencing rebound effect. *arXiv preprint arXiv:2510.09022*, 2025.
- Moser, B. B., Shanbhag, A. S., Frolov, S., Raue, F., Folz, J., and Dengel, A. A coreset selection of coreset selection literature: Introduction and recent advances. *arXiv preprint arXiv:2505.17799*, 2025.

- Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. On coresets for logistic regression. In *Advances in Neural Information Processing Systems*, pp. 6561–6570, 2018.
- Mytton, D., Lundén, D., and Malmmodin, J. Network energy use not directly proportional to data volume: The power model approach for more reliable network energy consumption calculations. *Journal of Industrial Ecology*, 28 (4):966–980, 2024.
- Nguyen, T., Novak, R., Xiao, L., and Lee, J. Dataset distillation with infinitely wide convolutional networks. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=hXWpPJedrVP>.
- O’Neil, C. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.
- Orr, W. and Crawford, K. The social construction of datasets: On the practices, processes, and challenges of dataset creation for machine learning. *New Media & Society*, 26 (9):4955–4972, 2024.
- Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. In *Advances in Neural Information Processing Systems*, volume 34, pp. 20596–20607, 2021.
- Penedo, G., Malartic, Q., Hesslow, D., Cojocaru, R., Alobeidli, H., Cappelli, A., Pannier, B., Almazrouei, E., and Launay, J. The refinedweb dataset for falcon LLM: Outperforming curated corpora with web data only. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. URL <https://openreview.net/forum?id=kM5eGcdCzq>.
- Rather, I. H., Kumar, S., and Gandomi, A. H. Breaking the data barrier: a review of deep learning techniques for democratizing ai with small datasets. *Artificial Intelligence Review*, 57(9):226, 2024.
- Ritchie, H., Rosado, P., and Roser, M. Data page: Carbon intensity of electricity generation. <https://archive.ourworldindata.org/20251014-145858/grapher/carbon-intensity-electricity.html>, 2023a. Data adapted from Ember and the Energy Institute. Part of Ritchie, Rosado and Roser (2023) *Energy*. Archived on 14 October 2025.
- Ritchie, H., Rosado, P., and Roser, M. Data page: CO₂ emissions per capita. <https://archive.ourworldindata.org/20251113-170236/grapher/co-emissions-per-capita.html>, 2023b. Part of the publication: *CO₂ and Greenhouse Gas Emissions*. Data adapted from the Global Carbon Project and other sources. Online resource archived on 13 November 2025.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.
- Santy, S., Bhattacharya, P., Ribeiro, M. H., Allen, K. R., and Oh, S. Position: When incentives backfire, data stops being human. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=4UhTWPwVke>.
- Scala, F., Flesca, S., and Pontieri, L. An efficient model training framework for green ai. *Machine Learning*, 114 (12):275, 2025.
- Selvan, R. *Sustainable AI: Tools for Moving toward Green AI*. O’Reilly Media, 2025. ISBN: 9781098155513.
- Shim, J.-h., Kong, K., and Kang, S.-J. Core-set sampling for efficient neural architecture search. *arXiv preprint arXiv:2107.06869*, 2021.
- Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. S. Beyond neural scaling laws: beating power law scaling via data pruning. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=UmvSlP-PyV>.
- Splunk. The state of dark data. https://www.splunk.com/en_us/form/the-state-of-dark-data.html, 2019. Online; accessed 24 January 2026.
- Strubell, E., Ganesh, A., and McCallum, A. Energy and policy considerations for modern deep learning research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 13693–13696, 2020.
- Tan, H., Wu, S., Huang, W., Zhao, S., and Qi, X. Data pruning by information maximization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=93XT0lKOct>.
- Toneva, M., Sordoni, A., des Combes, R. T., Trischler, A., Bengio, Y., and Gordon, G. J. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=BJlxm30cKm>.

- Upadhyay, S. K., Quirke, P., Oozeer, N. F., and Bandi, C. Position: Require frontier AI labs to release small “analog” models. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL <https://openreview.net/forum?id=xcdlSMYXxD>.
- Van Horn, G. and Perona, P. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017.
- Verdecchia, R., Cruz, L., Sallou, J., Lin, M., Wickenden, J., and Hotellier, E. Data-centric green ai an exploratory empirical study. In *2022 International Conference on ICT for Sustainability (ICT4S)*, pp. 35–45. IEEE, 2022.
- Veritas Technologies LLC. The data hoarding report, 2018. URL <https://www.veritas.com/content/dam/Veritas/docs/reports/Veritas-Data-Hoarders-Report-US.pdf>. Accessed: 27 January 2026.
- Violos, J., Diamanti, K.-C., Kompatsiaris, I., and Papadopoulos, S. Frugal machine learning for energy-efficient, and resource-aware artificial intelligence. *arXiv preprint arXiv:2506.01869*, 2025.
- Wang, Y., Li, Y., and Xu, C. Position: AI scaling: From up to down and out. In *International Conference on Machine Learning Position Paper Track*, 2025. URL <https://openreview.net/forum?id=UTxi86wmas>.
- Wightman, R., Touvron, H., and Jégou, H. Resnet strikes back: An improved training procedure in timm. *arXiv preprint arXiv:2110.00476*, 2021.
- Wilder, B. and Zhou, A. Fostering the ecosystem of AI for social impact requires expanding and strengthening evaluation standards. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL <https://openreview.net/forum?id=dl5pvd5IgW>.
- Wilson, S. N., Mair, S., Okinyi, M., Dam, E. B., Koch, J., and Selvan, R. How Hyper-Datafication Impacts the Sustainability Costs in Frontier AI. *arXiv preprint*, 2026.
- Wright, D., Igel, C., Samuel, G., and Selvan, R. Efficiency is not enough: A critical perspective on environmentally sustainable ai. *Communications of the ACM*, 68(7):62–69, 2025.
- Yang, Y., Kang, H., and Mirzasoleiman, B. Towards sustainable learning: Coresets for data-efficient deep learning. In *International Conference on Machine Learning*, pp. 39314–39330. PMLR, 2023.
- Yin, Z., Xing, E., and Shen, Z. Squeeze, recover and re-label: Dataset condensation at imagenet scale from a new perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=5Fgdk3hZpb>.
- Zhao, B., Mopuri, K. R., and Bilen, H. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=mSAKhLYLSsl>.
- Zimmermann, A., Zeppa, A., Pandey, S., and Diao, K. Don’t give up on democratizing AI for the wrong reasons. In *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track*, 2025. URL <https://openreview.net/forum?id=FjxyAotxtT>.

A. Overview of Data Reduction Techniques

Coreset selection methods. The goal of a *representative subset* is to approximate a sum over a large number of terms with fewer terms, i.e., $f(\theta) := \sum_{i=1}^n w_i f_i(\theta) \approx \sum_{i \in \mathcal{I}} \tilde{w}_i f_i(\theta) =: \tilde{f}(\theta)$, where \mathcal{I} is a subset of the indices of much smaller size, i.e., $|\mathcal{I}| =: \tilde{n} \ll n$, and \tilde{w}_i is a potentially adjusted weight. Typically, the large sum iterates over the training data and the functions f_i correspond to per-point loss functions and $\tilde{f}(\theta)$ is often an unbiased approximation of the objective function $f(\theta)$. However, instead of matching the objective function, other quantities like the gradient can be matched. Without guarantees on the approximation error, the subset merely serves as a *heuristic*. When theoretical guarantees on the approximation error can be derived, we call the representative subset a *coreset*.

Numerous constructions exist for representative subsets and coresets across learning tasks, see Moser et al. (2025) for an overview. For example, constructions for k -means clustering (Har-Peled & Mazumdar, 2004; Lucic et al., 2016; Bachem et al., 2018) and matrix factorization (Mair et al., 2017; Mair & Brefeld, 2019) often rely on geometric principles and sensitivity scores that estimate each data sample’s influence on the learning objective. Similar approaches have also been applied to logistic regression (Munteanu et al., 2018; Campbell & Beronov, 2019).

Other strategies derive informative scores from early training dynamics, such as the expected L2 norm (EL2N) or the gradient norm (GraNd) of the loss (Paul et al., 2021). These methods identify samples most critical to optimisation. For instance, GraNd enables up to 25% pruning on CIFAR-100 and 50% pruning on CIFAR-10 with negligible performance loss. This principle has also been applied to train language models using as little as 30% of the original dataset (Marion et al., 2023). Furthermore, Fayyaz et al. (2022) adapt GraNd and EL2N to NLP tasks.

Other gradient-based coreset methods aim to preserve the optimisation dynamics of training on the full dataset while enabling substantial reductions in dataset size and training time. For instance, CRAIG (Mirzasoleiman et al., 2020) selects subsets such that the gradient of the loss on the subset closely matches that of the full dataset, achieving speed-ups of up to $6\times$ for logistic regression and $3\times$ for deep neural networks. Grad-Match (Killamsetty et al., 2021) uses a similar approach, reducing dataset size to 30% for vision tasks with negligible accuracy loss. Similarly, CREST (Yang et al., 2023) demonstrates speed-ups of up to $2.5\times$ on various vision and NLP benchmarks with minimal performance loss. A combination of prediction uncertainty and training dynamics when scoring data points is approached by Dyn-Unc (He et al., 2024) while InfoMax (Tan et al., 2025) selects subsets that maximize representative information content. Both approaches demonstrate that ImageNet-1k can be pruned up to 25% with negligible to no performance degradation. More generally, training-dynamics-based approaches also identify important examples by explicitly monitoring learning behaviour, such as forgetting events, i.e., examples that repeatedly transition from correct to incorrect classification (Toneva et al., 2019), or early-training proxy signals that predict final importance (Coleman et al., 2020). More recently, \mathbb{D}^2 pruning (Maharana et al., 2024) uses a message-passing formulation that jointly balances sample interactions between data points. It improves over methods based solely on training dynamics or geometric coverage.

While all these methods reduce dataset size and accelerate training with a small drop in performance, their potential to reduce energy consumption and carbon emissions remains largely unexplored. Notable exceptions include Killamsetty et al. (2021), who quantify wall-clock speed-ups from gradient-matched subset selection as a proxy to reduce compute, and Scala et al. (2025), who explicitly evaluate data reduction strategies in terms of energy efficiency during training. Both works go beyond reporting accuracy and runtime and explicitly quantify energy savings from training on reduced datasets.

Coreset approaches for model selection. Several works leverage coreset and subset selection techniques to reduce the cost of model selection, including hyperparameter optimisation and neural architecture search (NAS). For example, Grad-Match (Killamsetty et al., 2021) uses gradient-matching subsets to preserve optimisation dynamics, while AUTOMATA (Killamsetty et al., 2022) explicitly integrates subsets into the hyperparameter optimisation loop. Similar ideas have been applied to NAS, where coresets (Shim et al., 2021) and active learning approaches (Geifman & El-Yaniv, 2019) have been used to substantially reduce model selection cost.

Data condensation. Dataset condensation seeks to replace large training sets with a small number of synthetic data points that induce similar training dynamics. Theoretical insights into this idea have been developed in the infinite-width regime of CNNs by Nguyen et al. (2021), while practical methods based on gradient matching made condensation effective in practice (Zhao et al., 2021). Other work has addressed scalability, enabling dataset condensation at ImageNet scale (Cui et al., 2023; Yin et al., 2023). Unlike coreset methods, which select representative subsets of data, dataset condensation produces synthetic datasets, offering often better compression at the cost of interpretability, data provenance, and additional optimisation overhead.

B. ImageNet Usage Analysis

ICLR-based identification of ImageNet usage. The following section describes our analysis of ICLR papers from 2017–2022 that used ImageNet to train models from random initialisation, and how we projected these trends to subsequent years.

Using the OpenReview API, we retrieved all available papers from the version 1 (v1) endpoint. We focus on ICLR because all years except 2016 used OpenReview for peer review. For each year, we analysed accepted papers for mentions of ImageNet and recorded the corresponding forum IDs. For matched papers, we processed the full PDF using an LLM to determine the context in which ImageNet was used (e.g., training from random initialisation, fine-tuning, or citation only) and to identify the variant of ImageNet used. Our primary analysis focuses on papers that trained models on ImageNet from random initialisation.

In many cases, papers do not explicitly describe how ImageNet was used. We therefore allow the model to infer the training scenario from the surrounding text, supported by quotes to justify the classification. Similarly, the dataset variant is not always specified; for papers trained from scratch, we assume the most commonly used version is ImageNet-1K with 1.28 million training images.

As we were unable to access the 2023–2025 papers, we estimated ImageNet-related statistics for these years using ordinary least squares (OLS) linear regression fitted to the 2017–2022 trends. Table 2 reports both observed and projected values, with Figure 5 providing a visual summary. Figure 5 shows that the fractions of ICLR papers mentioning and training on ImageNet exhibit a slight decline over time; however, the total number of accepted papers has increased substantially, resulting in continued growth in the absolute number of ImageNet-using papers.

Table 2. Observed (2017-2022) and projected (2023-2025) ImageNet usage and training counts. Totals for 2023-2025 are real acceptance counts; ImageNet-related values for 2023-2025 are projected via linear trends fit on 2017-2022.

Metric	Observed						Projected		
	2017	2018	2019	2020	2021	2022	2023	2024	2025
Total accepted (real)	196	336	502	687	858	1094	1574	2260	3704
ImageNet papers	75	129	183	238	319	395	554	785	1270
ImageNet ratio	0.383	0.384	0.365	0.346	0.372	0.361	0.352	0.348	0.343
Trained on ImageNet	28	43	55	80	105	110	154	207	315
Trained on ImageNet / total	0.143	0.128	0.110	0.116	0.122	0.101	0.098	0.091	0.085

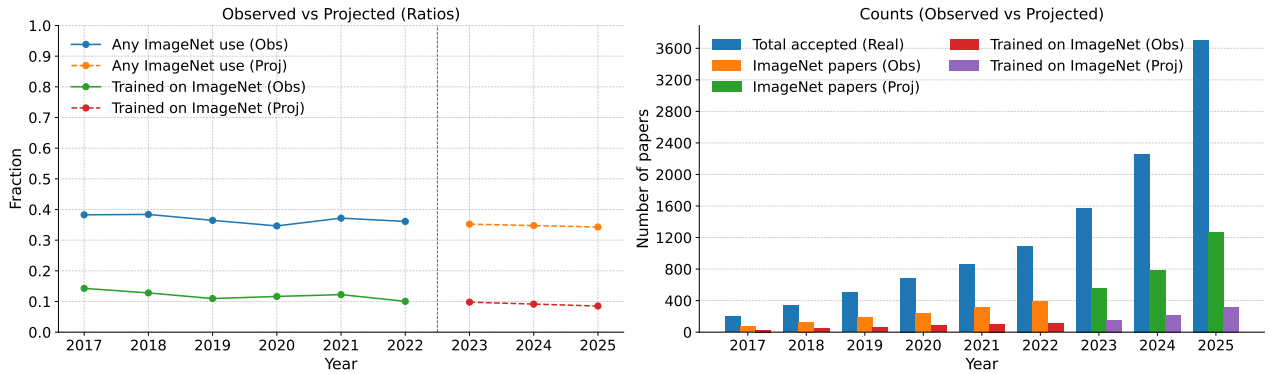


Figure 5. Left: Observed and projected fraction of papers at ICLR with ImageNet use. Right: Observed and projected growth of papers using ImageNet to train models from random initialization.

Extrapolating ImageNet usage beyond ICLR. To estimate total ImageNet usage beyond ICLR, we assume that the fraction of papers training on ImageNet from scratch at ICLR is representative of the broader ML literature. We therefore retrieve the total number of papers mentioning “ImageNet” between 2017 and 2025 via a keyword search on dimensions.ai¹⁶ and apply the estimated fraction to this corpus. Table 3 summarises these estimates.

¹⁶app.dimensions.ai

Table 3. Yearly number of publications mentioning ImageNet based on dimensions.ai and the inferred number of models trained from random initialisation, obtained by multiplying the ICLR-derived training ratio by the total number of ImageNet-mentioning papers.

	2017	2018	2019	2020	2021	2022	2023	2024	2025
Total published	12,998	21,389	31,538	38,374	48,294	53,825	59,847	63,516	65,124
Trained on ImageNet	1,104	1,946	3,090	3,875	5,891	6,243	6,583	8,130	9,312

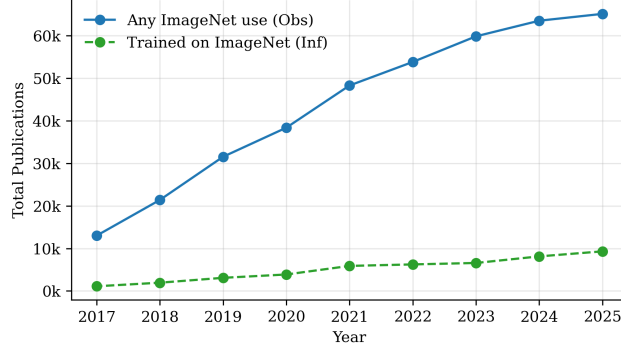


Figure 6. Total number of publications mentioning ImageNet (blue) retrieved from dimensions.ai and the total number of models trained on ImageNet from random initialisation (green), obtained by applying the annual ICLR-derived training ratios to the total publication counts.

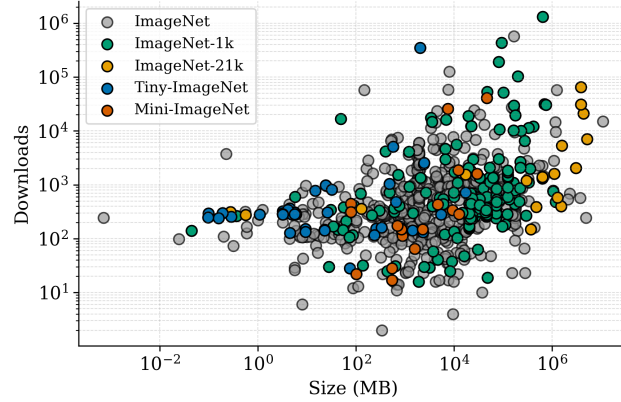


Figure 7. Total downloads versus dataset size (MB) for the 876 ImageNet-related datasets hosted on the Hugging Face Hub as of 1st December 2025. Among these, 214 datasets explicitly include *imagenet-1k* in their dataset identifier.

Figure 6 shows the total number of publications mentioning ImageNet and the corresponding estimated number of models trained on ImageNet obtained by applying the annual ICLR-derived ratios. This procedure results in an estimated total of 46,179 ImageNet training runs over the period 2017-2025.

Per-epoch energy measurement. To estimate the average energy consumption per training epoch, we trained a ResNet-50 model (He et al., 2016) three times for ten epochs and tracked the energy use with carbontracker¹⁷ (Anthony et al., 2020).

Versions and copies stored on the Hugging Face Hub. Since the release of ImageNet, numerous derivative versions and subsets have been released by both the original authors and the broader research community (Luccioni & Crawford, 2024) Figure 7 shows the 876 datasets hosted on the Hugging Face Hub¹⁸ which includes *imagenet* in their dataset identifier.

¹⁷<https://github.com/saintslab/carbontracker>

¹⁸Retrieved on 1 December 2025.

Among these 214 explicitly include *imagenet-1k* in their dataset identifier.

Scope and limitations. The estimates presented in Section 3 are strict lower bounds and reflect a deliberately narrow scope. We restrict our analysis to ImageNet-1K and to two stages of the data and model lifecycle: dataset storage and model training. Other stages, including data collection, curation, and distribution as well as model selection, evaluation, and deployment are out of scope. We further limit our accounting to energy use, specifically electricity consumption, and associated carbon emissions, excluding broader environmental impacts such as embodied hardware emissions and end-of-life effects.

The estimated number of training runs is derived from accepted papers and therefore does not capture the many training runs that do not result in a paper. Finally, we note that ImageNet-1K is a curated subset of the larger ImageNet-21K dataset, constructed to support more tractable benchmarking while retaining broad visual coverage. Our estimates therefore capture only a subset of all ImageNet-related usage.

C. Additional Material for Bias Experiment

Figure 8 depicts a visualisation of the data used in the experiment described in Section 4.3. Figure 9 shows the results for additional bias strengths, i.e., the fraction of samples belonging to the majority group $\{0.0, 0.75, 0.95\}$ for the following data budgets: $\{100, 500, 1000, 10000, 25000, 50000\}$. Unsurprisingly, stronger bias strength lead to larger gains from rebalancing the dataset to mitigate bias, as noted in Figure 4.

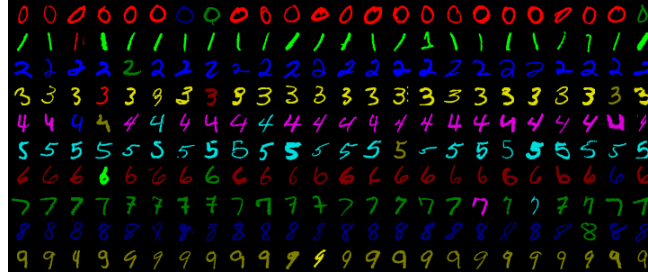


Figure 8. Visualisation of a batch of samples, generated with a bias probability of 0.1. Each digit has a majority group (colour) and a smaller minority of instances from other colours.

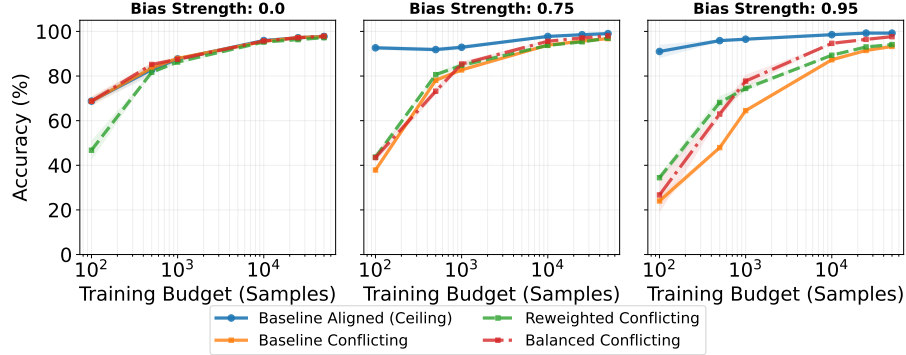


Figure 9. Performance of a classifier trained on the biased Colour-MNIST dataset is reported for the case with no bias (aligned) and with bias (conflicting) for four bias strengths (fraction of samples in the majority group): 0.0%, 75%, 95%, and 99%. Three coreset methods are shown: random (baseline), reweighted, and balanced.