Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI

Gaël Varoquaux
gael.varoquaux@inria.fr
Inria, Soda
Saclay, France

Alexandra Sasha Luccioni
Hugging Face
Montreal, Canada

Meredith Whittaker
Signal, University of Western Australia
Paris, France

Abstract

With the growing attention and investment in recent AI approaches such as large language models, the narrative that the larger the AI system the more valuable, powerful and interesting it is is increasingly seen as common sense. But what is this assumption based on, and how are we measuring value, power, and performance? And what are the collateral consequences of this race to ever-increasing scale? Here, we scrutinize the current scaling trends and tradeoffs across multiple axes and refute two common assumptions underlying the 'bigger-is-better' AI paradigm: 1) that performance improvements are driven by increased scale, and 2) that all interesting problems addressed by AI require large-scale models. Rather, we argue that this approach is not only fragile scientifically, but comes with undesirable consequences. First, it is not sustainable, as, despite efficiency improvements, its compute demands increase faster than model performance, leading to unreasonable economic requirements and a disproportionate environmental footprint. Second, it implies focusing on certain problems at the expense of others, leaving aside important applications, e.g. health, education, or the climate. Finally, it exacerbates a concentration of power, which centralizes decision-making in the hands of a few actors while threatening to disempower others in the context of shaping both AI research and its applications throughout society.

ACM Reference Format:
Gaël Varoquaux, Alexandra Sasha Luccioni, and Meredith Whittaker. 2025. Hype, Sustainability, and the Price of the Bigger-is-Better Paradigm in AI. In . ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction: AI is overly focusing scale

Our field is increasingly dominated by research programs driven by a race for scale: bigger models, bigger datasets, more compute. Over the past decade, machine learning (ML) has been used to build systems used by millions, carrying out tasks such as automatic translation, newsfeed raking, and targeted advertising. Pursuing scale has helped improve the performance of ML models on benchmarks for many tasks, with the paradigmatic example being the ability of large language models (LLMs) to encode "general knowledge" when pretrained on large amounts of text data. The perceived success of these approaches has further entrenched the assumption that bigger-is-better in AI. Here, we examine the underpinnings of this assumption about scale and its collateral consequences. We argue not that scale isn't useful, in some cases, but that there is too much focus on scale and not enough value on other research.

From narrative to norm. The famous AlexNet paper [59] has been key in shaping the current era of AI, including the assumption that increased scale is the key to improved performance. While building on and acknowledging decades of work, AlexNet created the recipe for the current bigger-is-better paradigm in AI, combining GPUs, big data (at least for the time), and large-scale neural network-based approaches. The paper also demonstrated that GPUs[1] scale compute much better than CPUs, enabling graduate students to rig a hardware setup that produced a model which outcompeted those trained on tens of thousands of CPUs. The AlexNet authors viewed scale as a key source of their model's benchmark beating performance: "*all of our experiments suggest that our results can be improved simply by waiting for faster GPUs and bigger datasets to become available*" [59, p.2]. This point was further reinforced by Sutton in his "bitter lesson" article, which states that approaches based on more computation win in the long term, as the cost of of compute decreases with time [114]. With this assumption comes both an explosion in investment in large-scale AI models and a concomitant spike in the size of the notable (highly cited) models. Generative AI, whether for images or text, has taken this assumption to a new level, both within the AI research discipline and as a component of the popular 'bigger-is-better' narrative surrounding AI – this, of course, has impacted training and inference compute requirements, which have ballooned in recent years (see Figure 1).

The bigger-is-better norm is also self-reinforcing, shaping the AI research field by informing what kinds of research is incentivized, which questions are asked (or remain unasked), as well as the relationship between industrial and academic actors. This is important because science, in AI as in other disciplines, does not happen in a vacuum – it is built upon relationships and a shared pool of knowledge, in which prior work is studied, incorporated, and extended. Currently, a handful of benchmarks define how "SOTA" (state-of-the-art) is measured and understood, and in pursuit of the goal of improving performance on these benchmarks, scale has become the preferred tool for achieving progress and establishing new records. Reviewers ask for experiments at a large scale, both

[1] GPUs, or Graphical Processing Units, were initially developed for video games, enabling better graphics rendering given their ability to carry out more processes in parallel compared to CPUs, or Central Processing Units, which had been the dominant hardware architecture.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

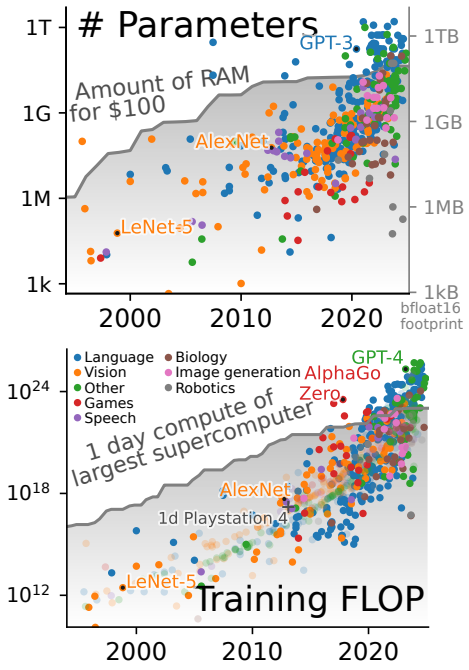

Figure 1: An explosion in model size – Top: The increase in model size means it is more and more expensive to run them in terms of RAM. – Bottom: resources we need are increasing faster than available compute. Data from Epoch [34], specific details in Appendix A.

in the context of new models and in the context of measuring performance against existing models [e.g. 39]; scientific best practices call for running experiments many times *e.g.* for hyperparameter selection [18]. These incentives and norms contribute to pushing computing budgets beyond what is accessible to most university labs – which in turn makes many labs increasingly dependent on close ties with industry in order to secure such access [1, 2, 129]. Taken together, we see the "bigger-is-better" norm in AI creating conditions in which it is increasingly difficult for anyone outside of large industrial labs to develop, test, and deploy SOTA AI systems.

The bigger-is-better assumption is also prevalent beyond the AI research community. It is shaping how AI is used and the understandings and expectations about its capabilities. Popular news reporting assumes larger amounts of compute result in and equate to better results and commercial success [61, 132]. Regulatory determinations and thresholds assume larger means more powerful and more dangerous, *e.g.* the US Executive Order on AI [15] and EU AI Act [91] – this broader assumption shapes markets and policy-making.

Paper outline. We start by discussing why this fixation on scale is misguided – we first examine this assumption in terms of how, and whether, scale leads to improvement in goal setting, finding that scale is not well correlated with better performance in certain contexts, and in fact benefits of scale tend to saturate. (Section 2). Then we look at harmful consequences that result from the growth of large-scale AI. Firstly, large scale development is unsustainable (Section 3), and the drive for more and more data encourages unethical and unauditable data practices (Section 4). Further, due to the expense and scarcity of hardware and talent required to produce large-scale AI, bigger-is-better strategies increasingly concentrate power over AI in the hands of a narrow set players (Section 5). We conclude by outlining how the research community can reclaim the scientific discourse in the AI field, and move away from a singular focus on scale.

2 What problems does scale solve?

2.1 Scale is one of many factors that matter

A staggering increase in scale, and cost. The scale of notable (highly-cited) models has massively increased over the last decade, driven by a super-exponential growth in terms of number of parameters, and amount of compute used (Figure 1). This growth is much more rapid than the increased capacity of hardware to execute necessary processing for model training and tuning. Indeed, while the size of large models, as measured by number of parameters, is currently doubling every 5 months (subsection A.3), the cost of 1 GB of memory has been nearly constant for a decade. This means that the resources required participate in cutting-edge AI research have increased significantly. The compute used to train an AI model went from a single day on a gaming console (in 2013) to surpassing the largest supercomputers on Earth (in 2020). A common response to concerns about the exponential growth of compute requirements is a reference to Moore's law[2] – i.e. that computational power will also increase, and this will help ensure that compute remains accessible [see 114]. However, this is not true in practice, since compute requirements for SOTA models are surpassing improvements in computational power. For instance, Thompson et al. [118] carried out a meta-analysis of 1,527 research papers from different ML sub-domains and extrapolated how much computational power was needed to improve upon common benchmarks like ImageNet and MS COCO. They estimated that an additional 567 times more compute would be needed to achieve a 5% error rate on ImageNet, stating that *"fundamental rearchitecting is needed to lower the computational intensity so that the scaling of these problems becomes less onerous"*.

Empirically, Sutton's "bitter lesson" [114] appears partly incorrect: it is not that, for AI, "*general methods that leverage computation are ultimately the most effective,* [because of] *Moore's law,* [...] *continued exponentially falling cost per unit of computation*", but that increasing resources are spent on AI. This increase in resources is visible in computational costs but is also true of other costs. For instance, building larger AI models require more human labor (Appendix C).

Diminishing returns of scale. The problem with the bigger-is-better approach is not simply that it is inaccessible. It also does not consistently produce model improvements, and, after a given point, even shows diminishing returns. On many tasks, benchmark performance as a function of scale tends to saturate after a certain point (see Figure 2). In addition, there is as much variability in model performance between models within a similar size class as

[2]Moore's law is an observation that states that the number of transistors on a microchip doubles roughly every two years.

there is between models of different sizes [9]. Indeed, many factors beyond scale are important to produce performant AI models. Choosing the right model architecture for the data at hand is crucial. Transformer-based models, widely perceived to be SOTA on most ML benchmarks, are not always the most fitting solution. For instance, when working with tabular data of the kind commonly produced in enterprise environments, tree-based models produce better predictions, *and* are much faster (and thus less expensive and more accessible) compared to neural network approaches (Figure 2d). This is notably due to the fact that their inductive bias is adapted to the specificity of columnar data [44].

Even for neural networks, or attention-based models, progress has been driven by much more than scale. Training or fine-tuning strategies play a very important role [26]. For instance, in text embeddings across models of the same size, and often even for the same type of pre-trained model, fine-tuning to produce useful similarities can markedly improve the resulting embeddings on domain-specific tasks [82, and Figure 2e]. We see this exemplified in the E5 model, which leverages both contrastive learning and curated data to achieve remarkable results [126]. Considering text-based models, encoder-based models often lead to representations that facilitate learning or language processing much better than decoder-based models, and do so in a much less compute intensive way [43, 127].

2.2 In many applications, utility does not requires scale

Different tasks can also call for models of different sizes, as seen on Figure 2. Using these benchmarks as a reference for comparison, we see that e.g. a 1 GB model performs well on medical image segmentation, even though the images themselves are relatively large. In computer vision, a 0.7 GB model can achieve good performance on a task such as object detection, even though scene parsing could require up to 3 GB of memory. For text processing, a natural language understanding task could be addressed with an LLM with hundreds of billions of parameters requiring more than 20 GB of memory (and multiple GPUs), even though a 1.3 GB model can also provide a good semantic embedding, which can be leveraged for solving the same task. Particularly for more focused tasks, smaller models often suffice: object detection (2b) versus scene parsing (2c, or semantics (2e) versus 'understanding' (2f).

Given this varied landscape, and the particularities in size and performance across tasks, where should we direct our research efforts? There is no single answer to this question, but it is instructive that a proposal made by Wagstaff over a decade ago called for ML with meaningful applications, each of which calls for approaches adapted to its constraints [125], a philosophy that has become decidedly unpopular in recent years in the pursuit of 'general-purpose' AI models. Let us survey published applications of ML, to sheds some light on corresponding tradeoffs. We are slowly starting to see an opening towards ML applications driven by real-world needs of end users – for instance, in the call for papers for the main track of the 2025 ICML conference, which explicitly calls for "Application-Driven Machine Learning" – a direction typically relegated to less prestigious (and less selective) workshops in other ML conferences.

If we consider health applications, where we would like machine learning to positively contribute to society, AI models are often built in data-scarce contexts on which large models can more easily overfit [123]. For example, to predict seizure outcome after epilepsy, Eriksson et al. [35] find no prediction-performance difference between logistic regression, boosted trees, or deep learning, even as these approaches differ dramatically in terms of resources required. The largest source of medical data are probably electronic health records – Rajkomar et al. [98] reported performance of deep learning, though their appendices show that logistic regression achieves the same performance. Later, Yèche et al. [136] find deep-learning under-performing in a thorough electronic health records benchmark. For education, one promise of machine learning is that it could produce customized pedagogy, shaped to the student based on which topics or skills they have learned, and which they are striving the understand. For this purpose, a random embedding of the student's learning sequence outperforms more resource intensive optimized deep or Bayesian models [30]. For robotics, Fu et al. [38] present an impressively useful system that can autonomously complete many daily-life tasks such as serving dishes whose key is co-training by a human. The learning systems used are rather modest, such as a ResNet18 architecture [49], dating from 2016 with 11 million parameters, which only take up 22 MB of memory with a precision of bfloat16–which is dwarfed by many of today's SOTA systems, whose size is often measured in GB (Figure 2).

If we inspect applications such as those mentioned above, we can see that for applied ML research what often matters most – even if we focus on simple purpose-specific metrics – is to address well-defined and application-specific goals while respecting the compute constraints at hand. And doing so often benefits from smaller, purpose-specific approaches as opposed to larger, more generic models. The benefit of purpose-specific models is also apparent in business applications: specificities of the business model lead to well-defined goals (efficiency, profit, etc.) [13]. Industry surveys indeed show that the most frequently used ML models are linear models and tree-based ones [60].

Happily for the AI research community, a focus on smaller models also unearths myriad unresolved scientific questions, which remain in need of scrutiny and innovative thinking independent of scale. For example, for many applications, bottlenecks lie in decision-making, calibration, and uncertainty quantification [121], not more compute or data. Similarly, causal machine learning and distributional shift are very active areas of research that are important irrespective of scale [56]. And interpretability is still an unrealized goal for most neural-network-based approaches, even as the ability to understand and audit such systems remains centrally important in the context of meaningful real life applications, especially in high-stakes scenarios like health [41, 83].

2.3 Unrepresentative benchmarks show outsize benefits from scale

Progress in AI is measured by an ever-evolving set of benchmarks which are meant to enable comparing the performance of various models in a standardized way. In fact, one of the main reasons ML practitioners are drawn to scale is that large scale models have, over the past decade or more, beaten smaller models, as measured

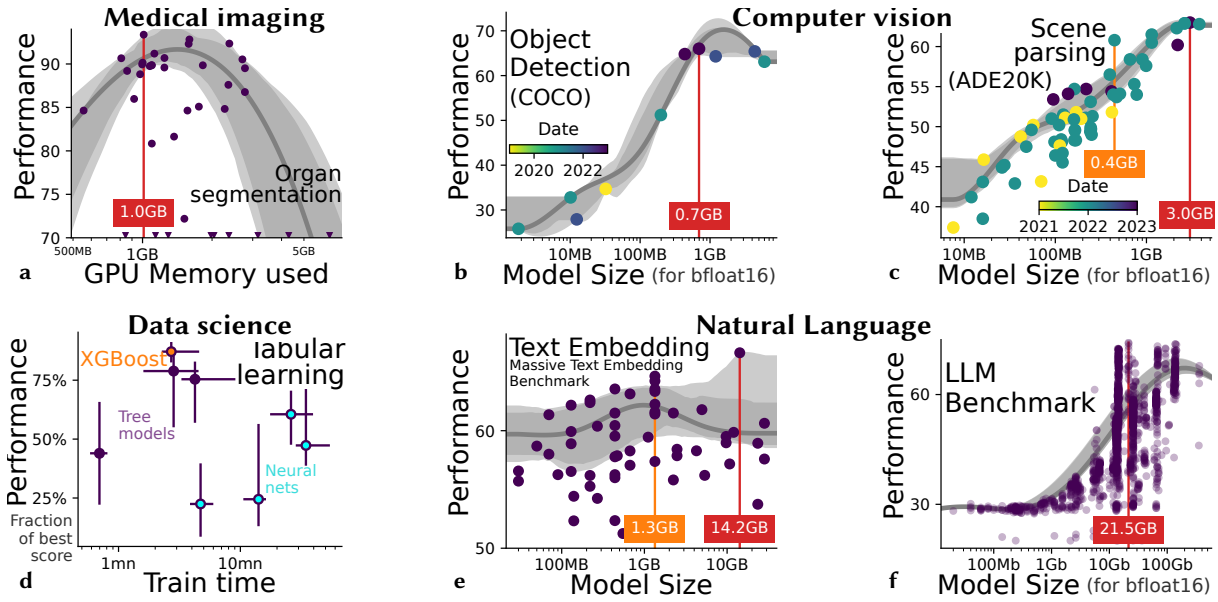

Figure 2: Performance as a function of scale saturates across various tasks. Plots of performance as a function of scale (time or memory footprint) on benchmark data from a) a medical image segmentation challenge [70], b) computer-vision object detection [62, COCO] and c) scene parsing [139, ADE20K], d) tabular learning [44], e) text embedding [82], and f) text understanding [9]. Details in Appendix B.

by these benchmarks. And beating the benchmark is currently how SOTA is defined (and how best paper awards are allotted, tenure given, funding secured, etc.). This persists even as this benchmark-centric evaluation often comes with ill-advised practices [37], both in terms of reproducibility of results [10, 72, 93] as well as the metrics used for measuring performance [95, 134]. In fact, one problem that the ML community currently faces is that many benchmarks were created to assess the performance of models in the context of the academic research field, not as a stand in for measures of contextualized performance, especially given the diversity of downstream contexts of applications in which AI models can be used, or to support broad claims made by marketers or companies given rising industrial interest and investment in AI.

Furthermore, with the advent of generative models including LLMs, the ML community has faced new evaluation challenges given the openendedness of model outputs. It is no longer possible to simply evaluate prediction on labels in the test set when there is no canonical 'correct answer'. This has resulted in a diversification of model evaluation practices, with the development of many different benchmarks intended to evaluative different aspects of generative model performance [22, 58] – this proliferation of non-standard approaches further confuses the landscape of assessment. To make matters worse, examining training datasets often reveals evidence of data contamination [32], including the presence of benchmark datasets inside training corpora [27]. This renders evaluations against these benchmarks questionable. All these shortcomings in benchmarks suggest that we should be very cautious regarding the validity of current evaluations of industrial models, given that the datasets used to train them are often not accessible to carry out the kind of public scrutiny necessary to identify such contamination or to understand the suitability of a given evaluation approach to the model at hand.

These issues have led to research on more rigorous model evaluation, including libraries [40, 124] and open leaderboards [9]. However, the crux of the matter remains that ML benchmarks and evaluations remain unrepresentative proxies of downstream model performance, and are not a substitute for more holistic assessments. Importantly, such benchmarks cannot speak to whether a given model is suited to a given purpose and context. In addition, the "cost functions" used for many of these comparisons are not absolute and result in, e.g., the fixation on specific metrics such as accuracy or precision, while ignoring others, such as robustness or efficiency, which remain crucial in real-world applications [103]. Finally, some ad-hoc benchmarks claim to measure ill-defined performance properties that are in fact impossible to verify (e.g. claiming that certain LLMs exhibit artificial general intelligence [19]). Considering a set of model capabilities much more narrow and better defined that general intelligence, model reasoning, thorough inspection of model capabilities has shown that benchmarks cannot be trusted as they are full of shortcuts exploited by models [79]. Such frail measures of high-level model capabilities impact the way the field as a whole tracks and perceives progress towards an ill-defined, illusory, and unattainable goal of general intelligence.

3 Consequence 1: An unsustainable trajectory

As we have shown in the sections above, focusing all of our research efforts on the pursuit of bigger models narrows the field, and scale is not an efficient solution to every problem. There are many interesting problems in AI that do not require scale. The singular

pursuit of scale also comes with worrying consequences, which we examine below.

Efficiency can have rebound effects. As we have seen from historical data, the compute required to create and deploy SOTA AI models grows faster than the cost of compute decreases. And yet, if we succeed at making these technologies generally useful, deployment at massive scale would be the logical next step, making these SOTA models accessible to increasing quantities of users. We may hope that efficiency improvements will come to the rescue, but the recent history of AI suggests the opposite. It is well-known phenomenon in economics that when the efficiency of a general-use technology increases, the falling costs lead to an increased demand, resulting in an overall increase in resource usage. This is referred to as "Jevons Paradox" [55]. Indeed, the AlexNet paper, which helped launch the current bigger-is-better era of AI research, was itself introducing a profound efficiency improvement, showing that a few GPUs could compete with tens of thousands of CPUs. One could assume that this would make AI training "democratically accessible", but in fact the opposite happened, with subsequent generations of AI models using even larger quantities of GPUs and training for a longer amount of time – with the training of the most recent generations of LLMs being measured in *millions* of GPU hours. This counter-intuitive dynamic has been documented numerous times in energy, agriculture, transportation, and computing [137], where observed trends have led to truisms such as Wirth's law, which asserts that software demands outpace hardware improvements [131].

3.1 The costs of scale-first AI don't add up

In fact, in many organizations, compute costs are already a major roadblock to developing and deploying ML models [89]. This is true even at large, tech-savvy companies such as Booking.com, who have found that gains in model performance cease to translate into gains in value to the company at a certain point, at which other factors such as efficiency or robustness become more important [13]. Increase in model scale aggravates this problem: even after the costs of training are accounted for, large models come with large deployment costs. Indeed, the price of a single inference (i.e. model query) has been growing in the last two decades despite hardware efficiency improvements (Figure 3).

Prohibitive costs of model deployment at scale threatens the business model of AI. Due to these costs, major investors are questioning the financial health of the current AI massive growth [3, 21]. For instance, the compute required to run ChatGPT as a publicly accessible interface cost OpenAI $ 700 000 per day in 2023 [116], and OpenAI's 2024 estimated balance is a $ 5 B loss [54]. According to Alphabet's chairman, a search using Bard, which is powered by Google's recent PaLM 2 model [7], is 10 times more expensive than a pre-Bard Google search [25]. In a similar vein, Bornstein et al. [17] reports that in generative AI gross margins are "more often as low as 50-60%, driven largely by the cost of model inference". The problem of costs is increasingly pressing and many recent announcements around industry R&D have focused on the cost or energy efficiency of new developments: OpenAI announced their recent GPT-4o mini as "advancing cost-efficient intelligence" [86]. DeepSeek-v3 [63] is a very large (671B parameters) and capable

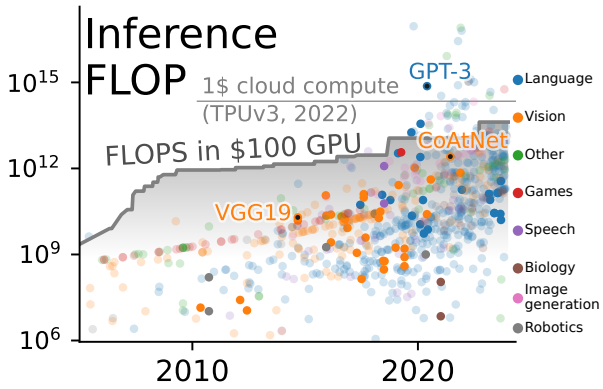

Figure 3: The cost of a single inference is growing faster than compute is improving– The x axis displays compute cost in FLOP (number of Floating Point Operations) for an inference of a model for notable models, while the grey line displays FLOPS (Floating Point Operations per Seconds) for $ 100 of GPU (sources of data detailed in appendix A.1).

language model which positions itself putting forward much decreased training and inference costs. The modernBERT paper [127] tackles language processing while specifically avoiding generative modeling to decrease costs. Microsoft CEO Satya Nadella even argued in late 2024 that progress in generative AI should be measured in "token per dollar per watt", as opposed to metrics focusing only on accuracy- or cost [42]. As models are improving, the costs of one query are going down for a given model size. However, will this be enough? "GenAI" product development is calling not only for increased model scale –*eg* to alleviate hallucinations–, but also increasing quantities of inferences by these models, adding them in user-facing applications on the web, but also in smartphones, smart connected devices, cars, etc.

Beyond generative AI, self-driving cars provide cautionary example. Here again the economics are challenging – 'autonomous' driving has turned out to rely heavily on the support of human workers and expensive sensors [51]. As of 2024, the field of companies endeavoring to commercialize fully autonomous vehicles has shrunk considerably. Cruise suspended its efforts, leaving only Waymo remaining [109].

3.2 Worrying environmental impacts

The emphasis on scale in AI also comes with consequences to the planet, since training and deploying AI models requires raw materials for manufacturing computing hardware, energy to power infrastructure, and water to cool ever-growing datacenters. The energy consumption and carbon footprint of AI models has grown with models such as LLMs emitting up to 550 tonnes of CO_2 during their training process [31, 66, 68, 113]. But the most serious sustainability concerns relate to inference, given the speed at which AI models are being deployed in user-facing applications. It is hard to gather meaningful data regarding both the energy cost and carbon emissions of AI inference because this varies widely depending on the deployment choices made (e.g. type of GPU used, batch size, precision, etc.). However, high-level estimates from Meta attribute

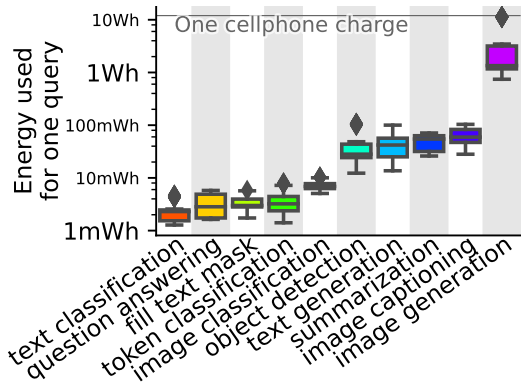

Figure 4: A single inference uses more energy for models with broad purposes. Data from Luccioni et al. [69].

approximately one-third of their AI-related carbon footprint to model inference [133], whereas Google attributes 60% of its AI-related energy use to inference [92]. Comparing estimates of the energy required for inference for LLMs [69] to ChatGPT's 10 million users per day [88] reveals that within a few weeks of commercial use, the energy use of inference overtakes that of training. Given the efforts to make AI more ubiquitous, these numbers could go up by many orders of magnitude. In fact, we are already seeing chatbots applied in consumer electronics such as smart speakers and TVs, as well as appliances such as refrigerators and ovens. The increased use of AI in user-facing products is putting tech companies' climate targets at risk, with both Microsoft and Google announcing in 2024 that they would miss the sustainability targets they set in previous years due to the energy demands of AI [78, 99].

Comparing different model architectures also shows that "general-purpose" (multi-task or zero-shot) models have higher energy costs than specialized models that were trained or fine-tuned for a specific task [69, and Figure 4]. This means that the current trend of using generative AI approaches for tasks such as Web search, which were previously carried out using "legacy" (information retrieval and embedding-based) systems, stands to increase the environmental costs of our day-to-day computer use significantly. The impact of a models' inference may seem a trifle compared to say a plane ride, however here again the challenge comes from the huge number of times they can be used. Data centers are already putting noticeable pressure on electricity networks of modern countries such as the US [84], leading tech companies to fund nuclear power plants for their exclusive usage [71]. Taking a step back, according to recent figures, global data centre electricity consumption represents 1-1.3% of global electricity demand and contributes 1% of energy-related greenhouse gas emissions [23, 52]. It is hard to estimate what portion of this number is attributable to AI. However, a recent report from the International Energy Agency estimates that electricity consumption from data centres and AI is set to double in 2 years, surpassing that of Japan (1 000 TWh) in 2026. [53].

4 Consequence 2: More data, more problems

As ML datasets grow in size (Figure 5), they bring a slew of issues ranging from documentation debts to a lack of auditability to biases. We discuss these below.

Data size in tension with quality. In recent years, pretraining has become the dominant approach in both computer vision [100, 115] and natural language processing [29, 64]. This approach requires access to large datasets, which have grown in size, from millions of images for datasets such as ImageNet [28] to gigabytes of textual documents for C4 [96] and billions of image-text pairs in LAION-5B [107]. The premise of pretraining is that more data will improve model performance and ensure maximal coverage in model predictions, with the assumption that the more data used for pretraining, the more representative the model's coverage will be of the world at large. However, numerous studies have shown that neither image nor text datasets are broadly representative, reflecting instead a select set of communities, regions and populations [32, 67, 97, 102]. In fact, recent research on the LAION datasets showed that as these datasets grow in size, the issues that plague them also multiply, with larger datasets contain disproportionally more problematic content than smaller datasets [16]. Thiel [117] reported that the LAION datasets included child sexual abuse material, prompting the datasets to be taken down from several hosting platforms. But with dozens of image generation models trained on LAION variants, some already deployed in user-facing settings, it is hard to assess and mitigate the negative effects of this grim reality [105].

While a growing wave of research has proposed a more 'data-centric' approach to data collection and curation [80, 138], attempts to actually document the contents of ML datasets have been hampered by their sheer size, which requires compute and storage beyond the grasp of most members of the ML research community [65], before the challenges of classifying and understanding the contents of a given dataset are even addressed. As the field works with ever larger datasets, we are also incurring more and more 'documentation debt' wherein training datasets are too large to document both during creation and post-hoc [11]. In a nutshell, this means that we do not truly know what goes in to the models that we rely on to answer our questions and generate our images, apart from some high-level statistics. And this, in turn, hampers efforts to audit, evaluate, and understand these models.

Invasive data gathering. From a privacy perspective, the perceived need for ever-growing datasets implicitly incentivizes more pervasive surveillance, as companies gather our clicks, likes and searches to feed models that are then applied to sell us consumer products or impact the order the search results that we see. In fact, the majority of the revenue of the Big Tech companies leading the AI revolution comes from targeted advertising - representing 80% of Google's [6] and 90% of Meta's [75, 76] annual revenue in 2022.

Also, while much of ML data gathering efforts have been operating under the assumption that copyright laws do not apply for data gathered from the Internet and used to train ML models (or, at the least, that training ML models constitutes 'fair use'), last year has seen a series of copyright lawsuits filed against many companies and organizations. This includes lawsuits from authors, artists and newspapers [45, 106, 110]. While is is too early to tell what the

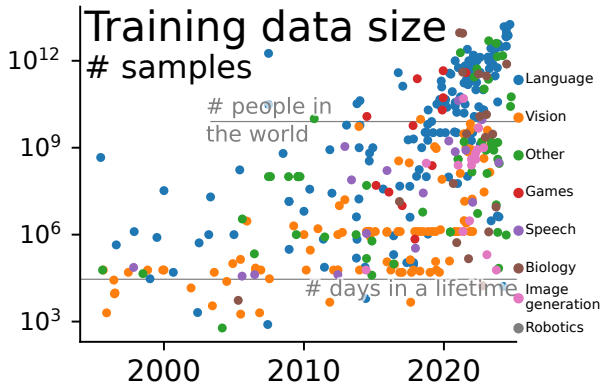

Figure 5: A sharp increase in amount of data used for training – Training data size (number of samples) as a function of time for notable models. Data from Epoch [34], specific details in Appendix A.

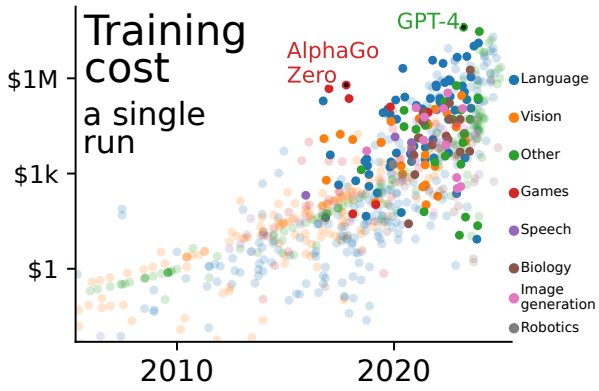

Figure 6: The increase in training costs leaves many aside – Training cost (US$, estimated) as a function of time for notable models. Data from Epoch [34], specific details in Appendix A.

verdicts of these lawsuits will be, they will undoubtedly impact on the way in which ML datasets are created and leveraged. The European Union's General Data Protection Regulation (GDPR) law sets some limitations in terms of data gathering and privacy in the EU, even if unevenly enforced. However, the United States and many other jurisdiction have yet to pass federal privacy laws, offering their citizens little to no privacy protection against predatory data gathering practices and surveillance, which the rush to scale AI incentivises.

5 Consequence 3: Narrow field, few players

5.1 Scale shouldn't be a requirement for science

The bigger-is-better paradigm shapes the AI research field, yet we are in a moment where expensive and scarce infrastructure is viewed as necessary to conduct cutting-edge AI research and where companies are increasingly focused on providing resources to large-scale actors, at the expense of academics and hobbyists [81]. Recent research argues that "a compute divide has coincided with a reduced representation of academic-only research teams in compute intensive research topics, especially foundation models." [14, p.1] We also see evidence of this in the way in which the AI field is both growing–as measured by number of PhDs in AI, and shrinking–as measured by graduates remaining in academia. Stanford HAI [111] reports that "the proportion of new computer science PhD graduates from U.S. universities who specialized in AI jumped to 19.1% in 2021, from 14.9% in 2020 and 10.2% in 2010.". This dynamic sees academia increasingly marginalized in the AI space and reliant on corporate resources to participate in large-scale research and development of the kind that is likely to be published and recognized. And it arguably disincentives work that could challenge the bigger-is-better paradigm and the actors benefiting from it.

5.2 Scale comes with a concentration of power

The expense and scarcity of the ingredients needed to build and operate increasingly large models benefits the actors in AI that have access to compute and distribution markets. This works to concentrate power and influence over AI, which, in turn, provides incentives for those with sufficient resources to perpetuate the bigger-is-better AI paradigm in service of maintaining their market advantage.

The scale game benefits large players. Market concentration is apparent in compute resources, and is most evident when examining Nvidia GPUs. The cost and scarcity of these resources make it increasingly difficult for an academic lab, or even most startups, to purchase and rack sufficient hardware on premises. An H100, currently Nvidia's most powerful chip, costs up to $40,000. Recent shortages in chips have intensified the gap between the "GPU rich" and "GPU poor" [8] – notably, this constrained supply has benefited the large cloud providers who have privileged access to such hardware [17].

For training and deploying large models, most AI practitioners operating outside of large cloud companies must rent access to compute from these companies. Since the cost of training ML systems has grown exponentially, this can entail costs for large scale models in the tens to hundreds of millions of dollars, which is beyond the budget of the vast majority of academic researchers (Figure 6). However, tech companies are more able to invest this money to fuel their AI research and competitiveness – Meta alone is estimated to have spent $18 billion on GPUs in 2024 [48], while OpenAI announced early 2025 StarGate, a joint venture with SoftBank, Oracle, and MGX, planning more than $ 100 billion AI-infrastructure investments [87], bigger than the GDP of most countries in the world. While AI startups continue to raise significant capital, the majority of this capital ("up to 80-90% in early rounds") is paid to cloud providers [17]. Notably, the flow of capital here is often circular: three large cloud companies "contributed a full two-thirds of the $27bn raised by fledgling AI companies in 2023" [46].

These circular investment deals can blur the boundary between "investment" and "acquisition". For instance, the deal between Microsoft and OpenAI [currently under investigation by the FTC, 36], involved providing OpenAI access to Azure compute, in exchange for exclusive license to integrate OpenAI's GPT models,

and a promise of $1 trillion in profit delivered to Microsoft prior to any revenue being directed to OpenAI's social mission.

Shaping practices and applications. AI dominance is one factor that has helped vault companies into shapers of global economic forecasts and markets. The S&P 500, used by US investors as an index of market performance, is significantly shaped by the fates of large US tech companies [101]. This indicates that the interests of large AI companies, and particularly their emphasis on large-scale AI, have implications well beyond the tech industry, since a downturn in the AI market would impact global financial markets beyond. This creates market incentives, divorced from scientific imperatives, for perpetuating the bigger-is-better paradigm on which these companies are betting.

The expense of creating large-scale AI models also makes the need to commercialize these models more pressing – even if inference can be costly [122]. Here, established cloud and platform companies also have an advantage in the form of access to large and existing markets. Cloud infrastructure offerings are a readily available means to commercialize AI models, allowing startups and enterprises to rent access to, e.g., generative AI APIs as part of cloud contracts. The ecosystem of 'GPT wrapper' companies licensing access to the model via Microsoft's Azure cloud is an obvious example. McKinsey [74], looking at the 2022 landscape of AI, noted that "commercialization is likely tied to hosting. Demand for proprietary APIs (e.g. from OpenAI) is growing rapidly." Notable here is that the business model for commercializing generative AI in particular is still being worked out [17].

Consequences for innovation. When a few actors control an economy, it stifles innovation, as illustrated by recent history of computing. The 1990s were marked by the explosion of personal computing, and the race to commercialize networked computation; this innovation reshaped the economy and the labor market [20]. From there, the actors involved became concentrated into a small number of firms, becoming enablers and gatekeepers of personal-computing software. Resulting market capture created a rent, to the expense of the rest of the economy, including IT consumers [4]. This, in turn, led to a fall of innovation and growth because of lack of incentives for leaders and lack of market access for competitors [4]; desktop software has been moving slowly.

Today we face a similar scenario. Even the most well-resourced AI startups–like OpenAI, Anthropic, or Mistral– currently need compute resources so large that they can only be leased from a handful of dominant companies. And the primary pathway to a more widespread commercialization of AI models also lies through these companies. For example, Microsoft licensed their GPT API to 'wrapper' startup Jasper before they launched ChatGPT – Jasper provided writing help, offering services very similar to ChatGPT [90]. This, and OpenAI's launch of their GPT app store, inflamed concerns that Microsoft and OpenAI would leverage their informational advantage and privileged ability to shape the AI market and to compete with smaller players dependent on their platform. This echos practices that Amazon Marketplace has been criticised for, in which they used data collected about buyers and sellers to inform product development that directly competed with marketplace vendors [73].

Scale comes with consequences to society. The growth and profit imperative of corporations dictates finding sources of revenue that can cover the costs of AI systems of ever-growing size. These often puts them at odds with responsible and ethical AI use, which puts the emphasis on more consensual and incremental progress [57]. We see this dynamic at play in OpenAI's shifting stance and policies as they have become more closely tied to Microsoft, a choice they made admittedly to access scarce computational resources. This is illustrated by the 2024 change to their acceptable use policy, which significantly softened the proscription against using their products for "military and warfare"; this change unlocked new revenue streams [112], connected OpenAI's large scale AI to Microsoft's existing US military partnerships [135], and later pushed AI to direct battlefield application via a partnership with the defense-tech company Anduril [85]. The public did not have a say in this determination, nor, we assume, did the global stakeholders whose interests may not be safeguarded by the militaries that OpenAI chooses to provide services to. And yet, the concentrated private industry power over AI creates a small, and financially incentivized segment of AI decision makers. We should consider how such concentrated power with agency over centralized AI could shape society under more authoritarian conditions.

There are many other examples of financial incentives shaping the use of automated decision systems in ways that are not socially beneficial. Ross and Herman [104] showed United Health used a diagnostic algorithm to deny patients care, an application now at the center of a class action lawsuit filed by patients and their families [119]. The bigger-is-better AI paradigm is also exacerbating geopolitical concentration and tensions. Beyond the power of large corporations, government imperatives are also in play, particularly given that the large AI industry players are primarily located in the US. These tensions are apparent in the efforts on the part of the US government to limit the supply of computing resources such as GPUs to China in the global race for AI dominance [5]. While a discussion of whether such limits are warranted is outside the scope of this paper, this dynamic highlights the role of AI as a source of geopolitical power, one that given the current concentration of AI resources threatens to undermine technological sovereignty globally.

A broader landscape of risks. Deploying AIs is society may come with a variety of risks [12, 33, 128]. Some risks are already visible in existing systems such as the problems of fairness and replicating biases which lead to under-diagnosis in historically under-served populations [108]. Other risks are more speculative given today's systems but are to be considered in future scenario where AIs become much more powerful [24]. A concentration of power fostered by very large-scale AI systems that can be operated and control by only very few actors makes these multiple risks more problematic, as it challenges checks and balance mechanisms.

6 Ways Forward: small can also be beautiful

Machine learning research is central to defining technical possibilities and building the narrative in AI. There is no magic bullet that unlocks research on AI systems both small-scale and powerful, but shared goals and preferences do shape where research efforts go. We believe that the research community can and must act to pursue

scientific questions beyond ever-larger models and datasets, and needs to foster scientific discussion engaging the trade-offs that come with scale. As such, we call the research community to adopt the following norms:

Assigning value to research on smaller systems. By defining research agendas (which papers and talks are accepted, which questions are asked), the research community can shape how AI progresses and is appreciated. We can deepen our understanding of performance by diversifying our benchmarks, investigating the limitations of these benchmarks, and building bridges to other communities to pursue tasks and problems that are relevant in different contexts. All work on large AI system should also be compared to simpler baselines, to open the conversation on trade-offs between scale and other factors. And we should value research on open research questions such as uncertainty quantification or causality, even when it is conducted on smaller models, as it can bring valuable insights for the field as a whole.

Talking openly about size and cost. A scientific study in machine learning should report compute cost, energy usage and memory footprint for training and inference. Measuring these requires additional work, but it is small compared to the work of designing, training, and evaluating models, and can be done with available open-source tools like TensorBoard and Code Carbon. We should account for efficiency as well as performance when comparing models for instance, reporting metrics such as samples/kWh or throughput.

Holding reasonable expectations. Experiments at scale are costly, and we cannot expect everyone in the ML community to be "GPU rich". When assessing the merits of a scientific study, we should refrain from requiring additional experiments more costly than those already performed. And, as ML researchers, we should keep in mind that *1)* not every problem is solved by scale *2)* solving a problem that is only present at small scales is valuable as it decreases costs.

Quantitatively framing the changes that we call for can help anchoring them in the empirical practices of the AI community. For this, it is useful to consider, as in Figure 7, the trade-off between task performance (e.g. performance on an ML benchmark) and the computing resources used. In this space, the state of the art appears as a pareto-optimal frontier. Successes that typically make the headlines are "resource intensive": they move the field in the upper right direction, where using more computing resources is associated to better task performance. We must also celebrate progress in resource efficiency, moving the Pareto frontier to the right: less resource for the same performance. We believe that this framing and the discussions it entails are useful to build the discussion, and we advocate their more systematic use by AI practitioners.

7 Conclusion

As researchers and practitioners in the AI field, we find that the current AI paradigm rests on an unhealthy taste for scale. The common sense that 'bigger makes better' shapes research, directs capital, grounds popular narratives, and comes with dire consequences. These include economic inequalities that further divide the world's winners and losers; climate harms in which the proper way to use scarce and valuable resources is determined by AI's resource imperatives, not democratic decision; metastasis of surveillance as a hunger for more, and more data further erodes privacy; polluted information ecosystems via AI's reflection of warped data, its encoding of uncool internet dynamics and preferences as 'intelligence', and its application for tuning and calibrating social media feeds for engagement; and an artificially narrow AI research field as the resources required to do cutting edged inquiry–the kind that is awarded with grants and prestige–are increasingly pooled in the hands of the large companies, which also leads to the structural exclusion of small actors from research and development. Narrowing of research and inquiry, and the concomitant ability for the industry to shape which questions deserve resources and which do not, also shapes what we, the public, know about AI and those behind it. It also stunts development of smaller, often more useful, AI systems and data collection processes which often prove more useful in real life contexts than the large-scale systems. Given this, we believe that de-emphasizing scale as the north star of AI is imperative. Instead, the AI field should refocus on smaller, more elegant and nimble models. Those that can be run on widely-available hardware, at moderate costs, using more carefully created data; those that can be built around particular environments and context within which their real utility and benefit can be measured. Such a refocus will serve to address the many pathological symptoms associated with the race to scale, from widening inequality to climate harms, at a common root.

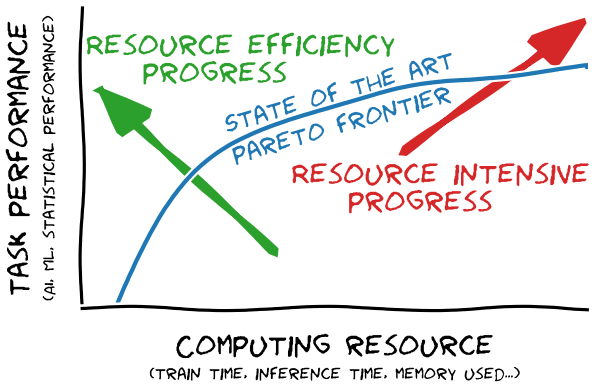

Figure 7: Pareto optimality and different contributions The state-of-the-art should be considered as a Pareto frontier in the trade-off space between task performance and computing resource. Visible contributions are often "resource intensive progress", increasing task performance by increase computing resources. We must also celebrate resource efficiency progress that decrease computing resources for the same task performance.

Positionality Statement

The authors of this paper come from a variety of backgrounds and perspectives, including sociotechnical studies, cognitive science, and theoretical and applied machine learning. We work at multiple different academic, non-profit and for-profit institutions

that span Europe and North America, and all engage in both analyzing and building computational technologies. Our opinions are informed both by our academic backgrounds and our practical work in software, data science, and machine learning/AI, coupled with extensive readings in other fields including economics, sustainability and sociology.

References

[1] Mohamed Abdalla and Moustafa Abdalla. 2021. The Grey Hoodie Project: Big Tobacco, Big Tech, and the Threat on Academic Integrity. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. ACM. https://doi.org/10.1145/3461702.3462563

[2] Mohamed Abdalla, Jan Philip Wahle, Terry Lima Ruas, Aurélie Névéol, Fanny Ducel, Saif Mohammad, and Karen Fort. 2023. The Elephant in the Room: Analyzing the Presence of Big Tech in Natural Language Processing Research. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.734

[3] Daron Acemoglu, Brian Janous, Jim Covello, Kash Rangan, and Eric Sheridan. 2024. Gen AI: too much spend, too little benefit? Goldman Sachs, Global Macro Research. https://www.goldmansachs.com/images/migrated/insights/pages/gs-research/gen-ai--too-much-spend%2C-too-little-benefit-/TOM_AI%202.0_ForRedaction.pdf

[4] Philippe Aghion, Antonin Bergeaud, Timo Boppart, Peter J Klenow, and Huiyu Li. 2023. A Theory of Falling Growth and Rising Rents. *The Review of Economic Studies* 90, 6 (02 2023), 2675–2702. https://doi.org/10.1093/restud/rdad016 arXiv:https://academic.oup.com/restud/article-pdf/90/6/2675/52976654/rdad016_supplementary_data.pdf

[5] Alexandra Alper, Karen Freifeld, and Stephen Nellis. 2023. Biden cuts China off from more Nvidia chips, expands curbs to other countries. https://www.reuters.com/technology/biden-cut-china-off-more-nvidia-chips-expand-curbs-more-countries-2023-10-17/.

[6] Inc Alphabet. 2023. FORM 10-K. https://abc.xyz/assets/investor/static/pdf/20230203_alphabet_10K.pdf?cache=5ae4398.

[7] Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. PaLM 2 Technical Report. arXiv:2305.10403 [cs.CL]

[8] Alistair Barr. 2023. The tech world is being dividing into 'GPU rich' and 'GPU poor.' Here are the companies in each group. *Business Insider* (2023). https://www.businessinsider.com/gpu-rich-vs-gpu-poor-tech-companies-in-each-group-2023-8

[9] Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open LLM Leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

[10] Anya Belz, Shubham Agarwal, Anastasia Shimorina, and Ehud Reiter. 2021. A systematic review of reproducibility research in natural language processing. *arXiv preprint arXiv:2103.07929* (2021).

[11] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

[12] Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, et al. 2024. International Scientific Report on the Safety of Advanced AI (Interim Report). *arXiv preprint arXiv:2412.05282* (2024).

[13] Lucas Bernardi, Themistoklis Mavridis, and Pablo Estevez. 2019. 150 successful machine learning models: 6 lessons learned at booking. com. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 1743–1751.

[14] Tamay Besiroglu, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. 2024. The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny? arXiv:2401.02452 [cs.CY]

[15] Joseph R Biden. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. (2023).

[16] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets. *arXiv preprint arXiv:2311.03449* (2023).

[17] Matt Bornstein, Guido Appenzeller, and Martin Casado. 2023. Who Owns the Generative AI Platform? https://a16z.com/who-owns-the-generative-ai-platform/.

[18] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, et al. 2021. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems* 3 (2021), 747–769.

[19] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[20] Ricardo J Caballero. 2010. Creative destruction. In *Economic growth*. Springer, 24–29.

[21] David Cahn. 2024. AI's $600B Question. Sequoia VC. https://www.sequoiacap.com/article/ais-600b-question/

[22] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109* (2023).

[23] Copenhagen Centre on Energy Efficiency. 2020. Greenhouse gas emissions in the ICT sector: Trends and methodologies. https://c2e2.unepdtu.org/wp-content/uploads/sites/3/2020/03/greenhouse-gas-emissions-in-the-ict-sector.pdf.

[24] Andrew Critch and Stuart Russell. 2023. Tasra: A taxonomy and analysis of societal-scale risks from ai. *arXiv preprint arXiv:2306.06924* (2023).

[25] Jeffrey Dastin and Stephen Nellis. 2023. For tech giants, AI like Bing and Bard poses billion-dollar search problem. https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/.

[26] Tom Davidson, Jean-Stanislas Denain, Pablo Villalobos, and Guillem Bas. 2023. AI capabilities can be significantly improved without expensive retraining. *arXiv preprint arXiv:2312.07413* (2023).

[27] Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating Data Contamination in Modern Benchmarks for Large Language Models. *arXiv preprint arXiv:2311.09783* (2023).

[28] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 248–255.

[29] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[30] Xinyi Ding and Eric C Larson. 2019. Why Deep Knowledge Tracing Has Less Depth than Anticipated. *International Educational Data Mining Society* (2019).

[31] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the carbon intensity of ai in cloud instances. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 1877–1894.

[32] Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758* (2021).

[33] Andrés Domínguez Hernández, Shyam Krishna, Antonella Maia Perini, Michael Katell, SJ Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, et al. 2024. Mapping the individual, social and biospheric impacts of Foundation Models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 776–796.

[34] Epoch. 2023. Parameter, Compute and Data Trends in Machine Learning. https://epochai.org/data/pcd.

[35] Maria H Eriksson, Mathilde Ripart, Rory J Piper, Friederike Moeller, Krishna B Das, Christin Eltze, Gerald Cooray, John Booth, Kirstie J Whitaker, Aswin Chari, et al. 2023. Predicting seizure outcome after epilepsy surgery: do we need more complex models, larger samples, or better data? *Epilepsia* 64, 8 (2023),

2014–2026.
[36] Federal Trade Commission. 2024. FTC Launches Inquiry into Generative AI Investments and Partnerships. https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-launches-inquiry-generative-ai-investments-partnerships.
[37] Peter Flach. 2019. Performance evaluation in machine learning: the good, the bad, the ugly, and the way forward. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9808–9814.
[38] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. 2024. Mobile ALOHA: Learning Bimanual Mobile Manipulation with Low-Cost Whole-Body Teleoperation. *arXiv preprint arXiv:2401.02117* (2024).
[39] Reviewer fySy. 2024. Review of Calibrated on Average, but not Within Each Slice: Few-shot Calibration for All Slices of a Distribution. https://openreview.net/forum?id=T11rD8k578. "I wonder if the problem of this paper would vanish as the data and model become larger. Could you please demonstrate it?".
[40] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation. https://doi.org/10.5281/zenodo.10256836
[41] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.
[42] Guillaume Grallet. 2024. Microsoft CEO Satya Nadella speaks about artificial intelligence, research and the "dance floor" of a thriving competition. Le Point. https://www.lepoint.fr/high-tech-internet/exclusive-microsoft-ceo-satya-nadella-speaks-about-artificial-intelligence-research-and-the-dance-floor-of-a-thriving-competition-05-11-2024-2574496_47.php
[43] Léo Grinsztajn, Edouard Oyallon, Myung Jun Kim, and Gaël Varoquaux. 2023. Vectorizing string entries for data processing on tables: when are larger language models better? *arXiv preprint arXiv:2312.09634* (2023).
[44] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems* 35 (2022), 507–520.
[45] Michael Grynbaum and Ryan Mac. 2023. The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.
[46] George Hammond. 2024. Big Tech outspends venture capital firms in AI investment frenzy. https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532.
[47] Karen Hao and Deepa Seetharaman. 2023. Cleaning up ChatGPT takes heavy toll on human workers. *The Wall Street Journal* (2023).
[48] Kali Hays. 2024. Zuck's GPU flex will cost Meta as much as $18 billion by the end of 2024. https://ca.style.yahoo.com/zucks-gpu-flex-cost-meta-171750011.html.
[49] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[50] Jessi Hempel. 2018. Fei-Fei Li's Quest to Make AI Better for Humanity. https://www.wired.com/story/fei-fei-li-artificial-intelligence-humanity/.
[51] Tim Higgins. 2022. Slow Self-Driving Car Progress Tests Investors' Patience. https://www.wsj.com/articles/investors-are-losing-patience-with-slow-pace-of-driverless-cars-11669576382.
[52] Ralph Hintemann and Simon Hinterholzer. 2022. Cloud computing drives the growth of the data center industry and its energy consumption. *Data centers 2022. ResearchGate* (2022).
[53] IEA. 2024. Electricity 2024. https://www.iea.org/reports/electricity-2024.
[54] Mike Isaac and Erin Griffith. 2024. OpenAI Is Growing Fast and Burning Through Piles of Money. https://www.nytimes.com/2024/09/27/technology/openai-chatgpt-investors-funding.html.
[55] W S Jevons. 1865. *The Coal Question.* MacMillan, London.
[56] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. 2022. Causal Machine Learning: A Survey and Open Problems. arXiv:2206.15475 [cs.LG] https://arxiv.org/abs/2206.15475
[57] Lauren Klein and Catherine D'Ignazio. 2024. Data feminism for AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 100–112.
[58] Bernard Koch, Emily Denton, Alex Hanna, and Jacob G Foster. 2021. Reduced, reused and recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716* (2021).
[59] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* 25 (2012).
[60] Dhiraj Kumar. 2022. Kaggle Survey 2022 Data Analysis. https://www.kaggle.com/code/dhirajkumar612/kaggle-survey-2022-data-analysis.
[61] Marcus Law. 2024. Meta Ramps up AI Efforts, Building Massive Compute Capacity. https://technologymagazine.com/articles/meta-ramping-up-ai-efforts-expanding-gpu-capacity.
[62] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.
[63] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).
[64] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[65] Alexandra Luccioni and Joseph Viviano. 2021. What's in the box? an analysis of undesirable content in the Common Crawl corpus. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 182–189.
[66] Alexandra Sasha Luccioni and Alex Hernandez-Garcia. 2023. Counting carbon: A survey of factors influencing the emissions of machine learning. *arXiv preprint arXiv:2302.08476* (2023).
[67] Alexandra Sasha Luccioni and David Rolnick. 2023. Bugs in the data: How ImageNet misrepresents biodiversity. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 14382–14390.
[68] Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2022. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *arXiv preprint arXiv:2211.02001* (2022).
[69] Sasha Luccioni, Yacine Jernite, and Emma Strubell. 2024. Power hungry processing: Watts driving the cost of AI deployment?. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 85–99.
[70] Jun Ma and Bo Wang. 2023. MICCAI FLARE23: Fast, Low-resource, and Accurate oRgan and Pan-cancer sEgmentation in Abdomen CT. https://codalab.lisn.upsaclay.fr/competitions/12239.
[71] C Mandler. 2024. Three Mile Island nuclear plant will reopen to power Microsoft data centers. https://www.bloomberg.com/graphics/2024-ai-power-home-appliances.
[72] Benjamin Marie, Atsushi Fujita, and Raphael Rubino. 2021. Scientific credibility of machine translation research: A meta-evaluation of 769 papers. *arXiv preprint arXiv:2106.15195* (2021).
[73] Dana Mattioli. 2020. Amazon Scooped Up Data From Its Own Sellers to Launch Competing Products. https://www.wsj.com/articles/amazon-scooped-up-data-from-its-own-sellers-to-launch-competing-products-11587650015.
[74] McKinsey. 2022. The state of AI in 2022—and a half decade in review. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2022-and-a-half-decade-in-review.
[75] Inc Meta. 2023. FORM 10-K. https://www.sec.gov/Archives/edgar/data/1326801/000132680123000013/meta-20221231.htm.
[76] Inc Meta. 2023. Meta Reports Fourth Quarter and Full Year 2023 Results; Initiates Quarterly Dividend. https://www.prnewswire.com/news-releases/meta-reports-fourth-quarter-and-full-year-2023-results-initiates-quarterly-dividend-302051285.html.
[77] Cade Metz. 2023. The Secret Ingredient of ChatGPT Is Human Advice. https://www.nytimes.com/2023/09/25/technology/chatgpt-rlhf-human-tutors.html.
[78] Rachel Metz. 2024. Google's Emissions Shot Up 48% Over Five Years Due to AI. *Bloomberg* (2024).
[79] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229* (2024).
[80] Margaret Mitchell, Alexandra Sasha Luccioni, Nathan Lambert, Marissa Gerchick, Angelina McMillan-Major, Ezinwanne Ozoani, Nazneen Rajani, Tristan Thrush, Yacine Jernite, and Douwe Kiela. 2022. Measuring data. *arXiv preprint arXiv:2212.05129* (2022).
[81] Sebastian Moss. 2018. Nvidia updates geforce eula to prohibit data center use. https://www.datacenterdynamics.com/en/news/nvidia-updates-geforce-eula-to-prohibit-data-center-use/
[82] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316* (2022).
[83] W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences* 116, 44 (2019), 22071–22080.
[84] Leonardo Nicoletti, Naureen Malik, and Andre Tartar. 2024. AI needs so much power it is making your's worse. https://www.bloomberg.com/graphics/2024-ai-power-home-appliances.
[85] James O'Donnell. 2024. OpenAI's new defense contract completes its military pivot. MIT Technology review. https://www.technologyreview.com/2024/12/04/1107897/openais-new-defense-contract-completes-its-military-pivot
[86] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

[87] OpenAI. 2025. Announcing The Stargate Project. https://openai.com/index/announcing-the-stargate-project/

[88] Will Oremus. 2023. AI chatbots lose money every time you use them. That is a problem. https://www.washingtonpost.com/technology/2023/06/05/chatgpt-hidden-cost-gpu-compute/. *Washington Post* (2023).

[89] Andrei Paleyes, Raoul-Gabriel Urma, and Neil D Lawrence. 2022. Challenges in deploying machine learning: a survey of case studies. *Comput. Surveys* 55, 6 (2022), 1–29.

[90] Arielle Pardes. 2022. The best little unicorn in Texas: Jasper was winning the AI race then ChatGPT blew up the whole game. https://www.theinformation.com/articles/the-best-little-unicorn-in-texas-jasper-was-winning-the-ai-race-then-chatgpt-blew-up-the-whole-game.

[91] EU Parliament. 2023. EU AI Act: first regulation on artificial intelligence. *Accessed June* 25 (2023), 2023.

[92] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2022. The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink. https://doi.org/10.48550/ARXIV.2204.05149

[93] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *The Journal of Machine Learning Research* 22, 1 (2021), 7459–7478.

[94] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*. 2641–2649.

[95] Matt Post. 2018. A Call for Clarity in Reporting BLEU Scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics, Belgium, Brussels, 186–191. https://www.aclweb.org/anthology/W18-6319

[96] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[97] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

[98] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine* 1, 1 (2018), 18.

[99] Akshat Rathi and Dina Bass. 2024. Microsoft's AI Push Imperils Climate Goal as Carbon Emissions Jump 30%. *Bloomberg* (2024).

[100] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640 [cs.CV]

[101] Joe Rennison and Eli Murray. 2023. How Big Tech Camouflaged Wall Street's Crisis. *International New York Times* (2023), NA–NA.

[102] Anna Rogers. 2021. Changing the world by changing the data. *arXiv preprint arXiv:2105.13947* (2021).

[103] Anna Rogers and Sasha Luccioni. 2024. Position: Key Claims in LLM Research Have a Long Tail of Footnotes. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=M2cwkGleRL

[104] Casey Ross and Bob Herman. 2023. UnitedHealth pushed employees to follow an algorithm to cut off Medicare patients' rehab care. https://www.statnews.com/2023/11/14/unitedhealth-algorithm-medicare-advantage-investigation.

[105] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[106] Adam Schrader. 2023. A Class Action Lawsuit Brought by Artists Against A.I. Companies Adds New Plaintiffs. https://news.artnet.com/art-world/lawyers-for-artists-suing-ai-companies-file-amended-complaint-after-judge-dismisses-some-claims-2403523.

[107] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems* 35 (2022), 25278–25294.

[108] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. 2021. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine* 27, 12 (2021), 2176–2182.

[109] David Shepardson and Ben Klayman. 2023. GM's Cruise suspends supervised and manual car trips, expands probes. https://www.reuters.com/business/autos-transportation/gms-cruise-suspends-supervised-manual-car-trips-expands-probes-2023-11-15/.

[110] Zachary Small. 2023. Sarah Silverman Sues OpenAI and Meta Over Copyright Infringement. https://www.nytimes.com/2023/07/10/arts/sarah-silverman-lawsuit-openai-meta.html.

[111] Stanford HAI. 2023. The AI Index Report: Measuring Trends in Artificial intelligence. https://aiindex.stanford.edu/report/

[112] Brad Stone and Mark MBergen. 2024. OpenAI Working With U.S. Military on Cybersecurity Tools. https://time.com/6556827/openai-us-military-cybersecurity/.

[113] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243* (2019).

[114] Richard Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)* 13, 1 (2019).

[115] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[116] The Economic Times. 2023. OpenAI faces financial challenges amid user decline: Experts predict bankruptcy concerns. https://economictimes.indiatimes.com/news/new-updates/openai-faces-financial-challenges-amid-user-decline-experts-predict-bankruptcy-concerns/articleshow/102711336.cms.

[117] David Thiel. 2023. *Identifying and Eliminating CSAM in Generative ML Training Data and Models*. Technical Report. Technical report, Stanford University, Palo Alto, CA, 2023.

[118] Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F. Manso. 2022. The Computational Limits of Deep Learning. arXiv:2007.05558 [cs.LG]

[119] US district court, district of Minnesota. 2023. Class action complaint. https://litigationtracker.law.georgetown.edu/wp-content/uploads/2023/11/Estate-of-Gene-B.-Lokken-et-al_20231114_COMPLAINT.pdf.

[120] Stef Van Buuren and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of statistical software* 45 (2011), 1–67.

[121] Ben Van Calster, David J McLernon, Maarten Van Smeden, Laure Wynants, and Ewout W Steyerberg. 2019. Calibration: the Achilles heel of predictive analytics. *BMC medicine* 17 (2019), 1–7.

[122] Jonathan Vanian and Kif Leswing. 2023. ChatGPT and generative AI are booming, but the costs can be extraordinary. *CNBC News* (2023). https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html

[123] Gaël Varoquaux and Veronika Cheplygina. 2022. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine* 5, 1 (2022), 48.

[124] Leandro Von Werra, Lewis Tunstall, Abhishek Thakur, Alexandra Sasha Luccioni, Tristan Thrush, Aleksandra Piktus, Felix Marty, Nazneen Rajani, Victor Mustar, Helen Ngo, et al. 2022. Evaluate & Evaluation on the Hub: Better Best Practices for Data and Model Measurement. *arXiv preprint arXiv:2210.01970* (2022).

[125] Kiri Wagstaff. 2012. Machine learning that matters. *ICML* (2012).

[126] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533* (2022).

[127] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. *arXiv preprint arXiv:2412.13663* (2024).

[128] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 214–229.

[129] Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (2021), 50–55.

[130] David Gray Widder, Sarah West, and Meredith Whittaker. 2023. Open (for Business): big tech, concentrated power, and the political economy of open AI. *Concentrated Power, and the Political Economy of Open AI (August 17, 2023)* (2023).

[131] Niklaus Wirth. 1995. A plea for lean software. *Computer* 28, 2 (1995), 64–68.

[132] Ben Wodecki. 2023. Want Better Image Models? More Compute is All You Need. https://aibusiness.com/computer-vision/want-better-image-models-scale-up-training-not-just-switch-architectures.

[133] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga Behram, James Huang, Charles Bai, et al. 2021. Sustainable AI: Environmental Implications, Challenges and Opportunities. *arXiv preprint arXiv:2111.00364* (2021).

[134] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016.

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. arXiv:1609.08144 [cs.CL]
[135] Chloe Xiang. 2023. The Defense Department Now Has GPT-4 Thanks to Microsoft. https://vice.com/en/article/y3wwwb/the-defense-department-now-has-gpt-4-thanks-to-microsoft.
[136] Hugo Yèche, Rita Kuznetsova, Marc Zimmermann, Matthias Hüser, Xinrui Lyu, Martin Faltys, and Gunnar Ratsch. 2021. HiRID-ICU-Benchmark—A Comprehensive Machine Learning Benchmark on High-resolution ICU Data. In *NeurIPS*.
[137] Richard York and Julius Alexander McGee. 2016. Understanding the Jevons paradox. *Environmental Sociology* 2, 1 (2016), 77–87.
[138] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, and Xia Hu. 2023. Data-centric ai: Perspectives and challenges. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 945–948.
[139] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 633–641.
[140] Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-Tuning Language Models from Human Preferences. arXiv:1909.08593 [cs.CL]

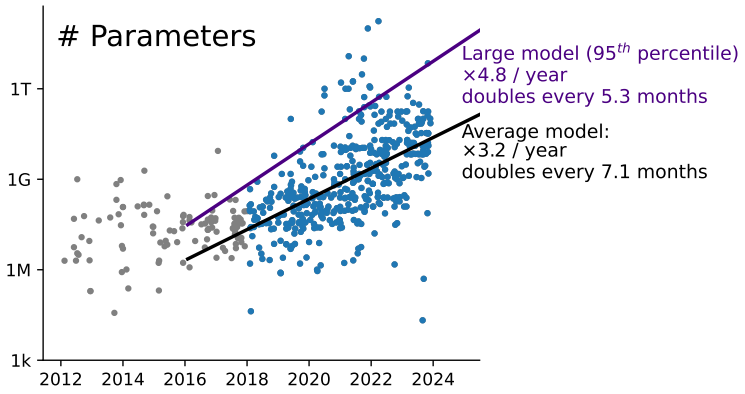

Figure 8: Doubling time of models, estimated between years 2018 and 2024, for the average model, and the 95^{th} percentile. Data from Epoch [34]; the estimation is detailed in subsection A.3.

A Methodology for the plots on historical trends

Here we give details on the plots on historical trends (Figure 1, Figure 3, Figure 6, Figure 5). All the Python code to generate the plots is available as supplementary material and will be deposited on github.

A.1 Sources of data

Evolution of notable AI systems. The data on the notable AI systems is adapted from Epoch [34] (https://epochai.org/data/epochdb/table). Here, a system is considered notable if it is highly cited. The data has been collected manually from publications.

Cost of memory. We used data on the cost of memory from OurWordInData: https://ourworldindata.org/grapher/historical-cost-of-computer-memory-and-storage

Cost of GPU FLOPs. The cost of GPU FLOPs (FLOPs denote the number of floating point operations per second) can be found here: https://epochai.org/blog/trends-in-machine-learning-hardware

An estimate of the cost of compute with TPUv3 can be found here: https://blog.heim.xyz/palm-training-cost: in 2022 "*We can rent a TPUv3 pod with 32 cores for $32 per hour. So, that's one TPUcore-hour per dollar.*"

Largest supercomputer. Data on the computing power of the largest super computer on Earth can be found here: https://en.wikipedia.org/wiki/File:Supercomputers-history.svg

FLOPs of other computer. Data on the computing power of other computers, including the PlayStation 4 can be found here: https://en.wikipedia.org/wiki/FLOPS

A.2 Imputation of missing values

Not all features of notable AI systems are reported, for instance inference cost, compute cost in dollars, may often be missing. For all plots but the one about dataset size[3] we impute the missing quantities from the reported one. Specifically, we use an imputation with chained equations [120, MICE]. We relate number of parameters, training FLOP, inference FLOP, training cost, to each other as well as to dataset size, number of epochs, application domain, and time the model was published.

The imputed data points are displayed on the various figures with transparency.

A.3 Estimation of doubling time

We estimate the doubling time of the number of parameters for the span 2018..2024. For this we use a linear model in log space: an ordinary least square regression to compute the evolution of the average model size as a function of time, and a quantile regression (pinball loss) to compute the evolution of the large models (95^{th} percentile of the size distribution). We find a doubling time of 7.1 months for the average model, which corresponds to a yearly increase of ×3.2 per year, and a doubling time of 5.3 months for the large models, which corresponds to an increase of ×4.8 per year (Figure 8).

The corresponding increase is very rapid: in three years, the size large models is multiplied by 110, and that of average models by 33. If growth continues at this pace, in 10 years large models will be 6.5 million times bigger, and average models 100 000 times bigger.

B Methodology for the plots on benchmarks

Here we give details on the plots on the benchmarks (Figure 2).

[3]We did not impute the dataset size as, unlike the other quantity, it does not need to have links to other quantities.

B.1 Sources of data

The difficulty with doing an analysis of how performance relates to scale is that they few benchmarks report a measure of scale. We used the following sources:

Tabular data. The data on tabular learning comes from the benchmark of Grinsztajn et al. [44], with detailed results downloaded from https://figshare.com/ndownloader/files/40081681. The timings and performance results were measured by Grinsztajn et al. [44] by running the algorithms for their benchmark.

Medical imaging. The data on medical imaging comes from an organ segmentation challenge: the MICCAI FLARE23 challenge (Fast, Low-resource, and Accurate oRgan and Pan-cancer sEgmentation in Abdomen CT): https://codalab.lisn.upsaclay.fr/competitions/12239#learn_the_details-testing_results The memory result was OCRed from the image of the table on the page. As the table report area under the GPU memory curve (as measured by the challenge organizers), we divided by time to have average GPU memory.

Object detection, COCO data. We consider the coco computer-vision object detection benchmark [62] and retrieve benchmark results from https://paperswithcode.com/sota/object-detection-on-coco. To have a model size in RAM (to compare with other benchmarks) we assume that parameters are bfloat16, which is on the low end of the memory footprint.

Scene parsing, ADE20K. We consider the ADE20K computer-vision scene parsing [139] and retrieve benchmark results from https://paperswithcode.com/sota/semantic-segmentation-on-ade20k. Here also we extrapolate memory footprint from number of parameters assume bfloat16 representation.

Massive Text Embedding Benchmark. We consider the massive text embedding benchmark [82], downloading results from https://huggingface.co/spaces/mteb/leaderboard.

LLM benchmark. We consider the "open LLM leaderboard" [9], downloading results from https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.

B.2 Fit of the .75 quantile

To study the dependence between model size and performance, we fit the .75 quantile: the goal is to look at the "best" model for a given parameter size, however given the scarcity of data we use the .75 quantile as conditional α-quantiles are hard to estimate for high α values.

We use a order-3 polynomial expansion to capture non-linear trends, and a quantile regression. We estimate prediction intervals by bootstrap.

C Labor and the human cost

While we, as a computer-science community, are well attuned to computational costs, other costs come in when developing large AI models. In particular, recent breakthroughs in ML rely on significant human labor to create data at various points throughout the development and deployment life cycle. ImageNet, the dataset that helped catalyze the current AI era, was only possible due to significant click worker labor, namely Amazon Mechanical Turk, which replaced costly and slow student labor [28, 50]. Subsequently, the Mechanical Turk platform, and the ability it gave companies and researchers to remotely direct large numbers of low paid workers, became critical to construct emblematic computer vision datasets like MS-COCO [62] and Flickr-30K [94]. More recently, reinforcement learning from human feedback (RLHF) has become a central to conversational large language models, allowing them to better respond to human preferences [140] and to provide outputs that stay within the boundaries of politically and socially acceptable expression. In fact, RLHF was hailed as the 'secret ingredient' of ChatGPT [77]. RLHF continues the legacy of Mechanical Turk, paying workers, often in developing countries, a few dollars an hour for labor that is repetitive, exhausting, and often emotionally taxing [47]. Even though each worker is paid little, given the scale of RLHF endeavors, it remains inaccessible to average research labs and startups, who cannot afford to pay the thousands of people required to carry out this tuning [130].