Policy Primer

Cohere For AI

Efficient AI

Increasing the Energy Efficiency of AI Models

Aidan Peppin, Ahmet Üstün, Sudip Roy, Ye Shen, Sara Hooker

January 2025

Summary

The growing global demand for digital products and services is increasing energy needs and their associated climate impacts. AI technologies are contributing to this challenge, due to the need for energy-intensive computing hardware, usually housed in data centers, to train and deploy AI models.

One approach to addressing this is improving AI model efficiency: reducing the amount of computing power that AI models require for training and real world use, while maintaining or increasing their performance.

A range of techniques are emerging to achieve this, across model design, pre-training, data efficiency, fine-tuning, model compression, and hardware optimization. While there are some trade-offs to be navigated, such as ensuring efforts to reduce model size do not reduce model accuracy, research focused on increasing model efficiency demonstrates that bigger is not always better – or even necessary – when it comes to AI models.

Work to improve model efficiency may also mark a potential shift away from ever-larger models, with smaller, more efficient models increasingly matching or outperforming larger ones. This trend is driven by financial cost considerations, hardware availability and optimization, energy demand concerns, and efforts to democratize AI access across regions with limited resources.

This policy primer outlines some of the challenges around measuring AI model efficiency systematically, and the techniques being developed to improve model efficiency. It focuses on work that can be done at the model developer layer, as opposed to the hardware or energy supply layers. It concludes with a range of considerations for those working to develop, deploy, or govern AI models:

1. Match model size to task requirements, avoiding oversized solutions for simple problems.
2. Develop and adopt standardized methods for measuring and reporting model efficiency.
3. Support the trend toward smaller models by prioritising research and development efforts that advance model efficiency.

CONTENTS

1. Introduction

Increasing global energy demands and their associated climate impacts are one of the most pressing challenges for current societies. One contributing factor to this challenge is the increasing energy requirements of digital products and services, largely due to their reliance on energy-intensive data centers. The increasing adoption of AI models and systems at scale is contributing to these growing energy demands.[1]

From 2017-2021 the electricity used by Meta, Amazon, Microsoft, and Google was estimated to have more than doubled,[2] and global data center electricity consumption has grown by 20-40% annually to now total between 1 and 1.3% of global energy demand and 1% of greenhouse gas emissions.[3] The International Energy Agency predicts that data center energy consumption could grow to more than 1,000TWh in 2026, more than double the 440TWh in 2022.[4]

While the evidence around how much AI is driving increased energy and carbon intensity is still nascent, one estimate suggests AI causes 10-20% of data center electricity consumption as of May 2024.[5] This is a product of the computing hardware required to develop and run AI models, known as "compute", such as energy-intensive GPUs housed in data centers.[6] AI training compute increased 300,000x from 2012-2018[7], and it is likely that the growing volume of compute used in training AI models and enabling their at-scale deployment could have big implications for energy use and associated climate impacts, such as carbon emissions and water usage.[8, 9] This growing energy demand of AI compute places strain on local and national

[1] Treviso, M. *et al.* (2023) 'Efficient Methods for Natural Language Processing: A Survey', *Transactions of the Association for Computational Linguistics*, 11, pp. 826–860. https://doi.org/10.1162/tacl_a_00577.
[2] Ralph Hintemann and Simon Hinterholzer. 2022. Cloud computing drives the growth of the data center industry and its energy consumption. Data centers 2022.
[3] https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
[4] International Energy Agency (2024) Electricity 2024, IEA. https://www.iea.org/reports/electricity-2024/executive-summary.
[5] Electric Power Research Institute (2024) *Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption*. https://www.epri.com/research/products/3002028905.
[6] Hooker, S. (2020) 'The Hardware Lottery'. arXiv. https://doi.org/10.48550/arXiv.2009.06489.
[7] Schwartz, R. et al. (2019) 'Green AI'. arXiv. https://doi.org/10.48550/arXiv.1907.10597.
[8] Treviso, M. et al. (2023) 'Efficient Methods for Natural Language Processing: A Survey'. arXiv. https://doi.org/10.48550/arXiv.2209.00099.
[9] Li, P. et al. (2023) 'Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models'. arXiv. http://arxiv.org/abs/2304.03271.

energy infrastructure, creating tensions between meeting both rising data center demand and continuing to supply homes, businesses, and critical services.[10, 11]

Given this background, attention has turned towards how the increased use of AI technologies is contributing to increased energy demand, and the associated impacts on both the global climate and local and national electricity infrastructure. A key aspect of this is understanding ways to reduce the compute demands of AI models and systems while maintaining (or improving) their performance; i.e. increasing the efficiency of AI models and systems. This could be a key component in reducing the energy demands and impacts of AI as it is increasingly adopted around the world.

This policy primer explores recent research associated with understanding the energy demands of AI model training and deployment, and efforts towards more efficient AI models as one approach to mitigating their energy and associated climate impacts. This primer focuses on research on model efficiency to draw insights for those working in policy and governance for how to support efforts to reduce the energy and carbon impacts of AI models.

Note: the climate and environmental impacts of AI models go beyond energy consumption for training and deployment, including aspects such as embodied carbon of physical hardware and data centers. This primer focuses primarily on the impacts of the models themselves to focus the scope.

2. Understanding AI Energy Requirements

Training frontier AI models involves performing billions of computing operations.[12] These operations are resource intensive and require large-scale, specialized hardware, usually housed in data centers. Measuring the total amount of energy used by this hardware to train an AI model, and the associated carbon emissions, is challenging to do accurately, and there are few studies which explore this, especially for recent models.[13, 14] For example, one estimate suggests training GPT-3 (a model released in 2020 with 175 billion parameters) consumed 1287 MWh of electricity, and resulted in carbon emissions of 502 metric tons of carbon.[15] However, it is difficult to systematically analyse or compare model training energy use because of the lack of shared hardware and differences in approaches to training. Beyond the impacts of training AI models, the deployment of AI models – for example in widely-used consumer facing services or in large-scale organisational processes – also adds to the energy and carbon costs of data

[10] Coskun, A. (2024) *AI supercharges data center energy use – straining the grid and slowing sustainability efforts*, *The Conversation*. http://theconversation.com/ai-supercharges-data-center-energy-use-straining-the-grid-and-slowing-sustainability-efforts-232697.
[11] Jackson, A. (2024) *Power-Hungry Data Centres Put Pressure on Ireland's Grid*. https://datacentremagazine.com/critical-environments/power-hungry-data-centres-put-pressure-on-irelands-grid.
[12] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv. http://arxiv.org/abs/2407.05694.
[13] Strubell, E., Ganesh, A. and McCallum, A. (2019) 'Energy and Policy Considerations for Deep Learning in NLP'. arXiv. https://doi.org/10.48550/arXiv.1906.02243.
[14] Luccioni, A.S., Viguier, S. and Ligozat, A.-L. (2022) 'Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model'. arXiv. https://doi.org/10.48550/arXiv.2211.02001.
[15] Patterson, D. *et al.* (2021) 'Carbon Emissions and Large Neural Network Training'. arXiv. https://doi.org/10.48550/arXiv.2104.10350.

centers.[16, 17] These deployment impacts are more important to address because "models are generally trained only once, while they are used for inference thousands if not millions of times".[18] "Inference" refers to when a trained AI model makes predictions, classifications, or decisions to produce outputs based on new data; in other words, inference is the name given to the operations AI models perform when they are used in practice. The energy costs of individual model inferences are relatively small, estimated to range from 0.002 kWh for every 1000 text classification inferences, to 0.047kWh per 1000 text generation inferences or 2.907 kWh per 1000 image generation inferences.[19, 20] While these numbers seem small, the cost of deployment would surpass training within "a few weeks" for models that are used in high volume.[21]

Difficulties Measuring Energy Use in a Standardized Way

The word 'estimated' used so far to refer to AI models' energy use and carbon impacts in this section is key: measuring a model's energy use is non-trivial. This is due to a range of challenges to accurately measuring an AI model's energy use, such as different hardware types having different energy requirements, and different data centers relying on different energy sources.[22, 23] Because of this complexity, there is a lack of standardized methodologies for quantifying and comparing energy consumption and carbon emissions of AI models. A handful of tools and techniques have emerged, such as Code Carbon[24] or LLM-Carbon.[25] However, all take different approaches, so their results can't be systematically compared.[26] Additionally, many existing approaches to measure model efficiency only capture certain aspects and do not correlate, meaning developers lack a comprehensive way to measure and report model efficiency.[27] One effort to address this challenge is the AI Energy Star project, which aims to "help developers and users of AI models to take energy consumption into account by testing a sufficiently diverse array of AI models for a set of popular use cases to establish an expected range of energy consumption, and then rate models depending on where they lie on this range."[28]

Despite these challenges related to accurately measuring the energy use and therefore carbon impacts of AI models, it is clear AI model training and use is energy intensive. While efforts to systematically understand the energy requirements of AI models are still nascent, they already offer useful insights by highlighting the factors that influence AI models' energy consumption.

---

[16] Luccioni, A.S., Jernite, Y. and Strubell, E. (2024) 'Power Hungry Processing: Watts Driving the Cost of AI Deployment?' https://doi.org/10.1145/3630106.3658542.
[17] Wu, C.-J. *et al.* (2024) 'Beyond Efficiency: Scaling AI Sustainably', *IEEE Micro*, pp. 1–8. https://doi.org/10.1109/MM.2024.3409275.
[18] Argerich, M.F. and Patiño-Martínez, M. (2024) 'Measuring and Improving the Energy Efficiency of Large Language Models Inference', IEEE Access, 12, pp. 80194–80207. https://doi.org/10.1109/ACCESS.2024.3409745.
[19] Luccioni, A.S., Jernite, Y. and Strubell, E. (2024) 'Power Hungry Processing: Watts Driving the Cost of AI Deployment?' https://doi.org/10.1145/3630106.3658542.
[20] Argerich, M.F. and Patiño-Martínez, M. (2024)
[21] Luccioni, A.S., Jernite, Y. and Strubell, E. (2024), p.9.
[22] Treviso, M. et al. (2023)
[23] Argerich, M.F. and Patiño-Martínez, M. (2024)
[24] See: https://codecarbon.io/#about
[25] Faiz, A. *et al.* (2024) 'LLMCarbon: Modeling the end-to-end Carbon Footprint of Large Language Models'. arXiv. https://doi.org/10.48550/arXiv.2309.14393.
[26] Luccioni, A.S., Jernite, Y. and Strubell, E. (2024)
[27] Dehghani, M. *et al.* (2022) 'The Efficiency Misnomer'. arXiv. https://doi.org/10.48550/arXiv.2110.12894.
[28] Luccioni, S. *et al.* (2024) 'Light bulbs have energy ratings — so why can't AI chatbots?', *Nature*, 632(8026), pp. 736–738. https://doi.org/10.1038/d41586-024-02680-3.

The next section explores how these factors can be targeted in ways that reduce models' energy requirements while maintaining – or improving – model performance. In other words, increasing models' efficiency.

3. Improving Model Efficiency

For LLMs in particular, there are a range of different techniques that model developers can apply to yield efficiency gains made across the entire lifecycle, from training to deployment.[29, 30] These include (but are not limited to) the following:

- **Model design**: Model builders and developers can make choices about different model architectures, and ways to optimize them, to reduce compute requirements. For example, architectures such as transformers[31] or state-space models (SSMs), offer different efficiency advantages.[32] To optimize an architecture, techniques such as Mixture of Experts (MoE) can be applied, which uses a collection of specialized sub-models, or 'experts', to enhance overall performance without increasing the computational burden.[33] This works by passing inputs through only certain parts of a model, rather than the whole neural network. [34] Other optimization methods include multi-query attention and grouped-query attention, which improve how inputs are processed across transformer models' attention heads.[35]
- **Training and data efficiency**: at the pre-training stage, techniques to improve how data is used can reduce the compute required to train a model. This includes data pruning which removes low-quality data points from training datasets that don't contribute highly to the model learning process,[36, 37] as well as data deduplication which eliminates repeated content, leading to better performance and faster convergence.[38] Additionally, strategic choices made during pre-training can minimize issues that can arise later when attempting to compress the model, for example, training a model to tolerate large levels of compression.[39] One of the biggest impacts for pretraining efficiency is automating the exploration of data mixes and parameter architectures to find a suitable combination before scaling it up - via a process known as fast hyperparameter exploration.[40] Another technique that can be employed to improve

[29] Treviso, M. et al. (2023)
[30] Wu, C.-J. *et al.* (2024) 'Beyond Efficiency: Scaling AI Sustainably', *IEEE Micro*, pp. 1–8. https://doi.org/10.1109/MM.2024.3409275.
[31] Vaswani, A. *et al.* (2023) 'Attention Is All You Need'. arXiv. http://arxiv.org/abs/1706.03762.
[32] Gu, A. and Dao, T. (2024) 'Mamba: Linear-Time Sequence Modeling with Selective State Spaces'. arXiv. https://doi.org/10.48550/arXiv.2312.00752.
[33] Gritsch, N. *et al.* (2024) 'Nexus: Specialization meets Adaptability for Efficiently Training Mixture of Experts'. arXiv. https://doi.org/10.48550/arXiv.2408.15901.
[34] Google, Gemma Team (2024) 'Gemma: Open Models Based on Gemini Research and Technology'. arXiv. https://doi.org/10.48550/arXiv.2403.08295.
[35] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032
[36] Marion, M. *et al.* (2023) 'When Less is More: Investigating Data Pruning for Pretraining LLMs at Scale'. arXiv. http://arxiv.org/abs/2309.04564.
[37] Chimoto, E.A. *et al.* (2024) 'Critical Learning Periods: Leveraging Early Training Dynamics for Efficient Data Pruning'. arXiv. https://doi.org/10.48550/arXiv.2405.19462.
[38] Hernandez, D. *et al.* (2022) 'Scaling Laws and Interpretability of Learning from Repeated Data'. arXiv. https://doi.org/10.48550/arXiv.2205.10487.
[39] Ahmadian, A. *et al.* (2023) 'Intriguing Properties of Quantization at Scale'. arXiv. https://doi.org/10.48550/arXiv.2305.19268.
[40] Paul, S., Kurin, V. and Whiteson, S. (2019) 'Fast Efficient Hyperparameter Tuning for Policy Gradient Methods'. https://www.cs.ox.ac.uk/people/shimon.whiteson/pubs/paulnips19.pdf

efficiency is 'model distillation', which involves training a smaller "student" model to mimic the behavior of a larger "teacher" model, enabling the smaller model to increase performance while maintaining a smaller size.[41]

- **Fine-tuning**: in fine-tuning, techniques can be applied to limit weight updates to a smaller subset of parameters, reducing memory and computational demands. This is known as 'parameter efficient finetuning' (PEFT). These techniques include applying "adapters" - a small number of new parameters tuned for specific tasks - or prompt tuning which guides parameter updates towards only the parameters most relevant to the prompt.[42]
- **Inference and compression**: To enhance inference efficiency, models can be compressed to reduce size dramatically while maintaining performance. Techniques include 'pruning', which removes redundant model weights within the model and 'quantization' which represents parameters using a smaller number of bits, allowing for faster computation and requiring less memory.[43, 44] These techniques can have a significant impact on memory requirements, resulting in requiring less hardware since the model can be squeezed into fewer GPUs, which is one of the most impactful ways to improve energy efficiency.
- **Hardware:** Dedicated AI hardware like GPUs or TPUs can accelerate training and inference. General-purpose hardware like CPUs can offer greater flexibility but often are slow at processing machine learning workloads. When choosing hardware, factors like memory capacity, processing power, and support for different data precision formats should be carefully considered. This is a double edged-sword: while it suggests opportunities to optimize certain hardware for certain models and functions, it can mean that attempts to run models or processes on other hardware, for example because optimal hardware is not available or too costly, leads to reduced efficiency.[45, 46]

It is important to note that techniques to improve model efficiency are not trade-off free. Efforts to make large language models (LLMs) smaller and faster can actually harm the performance of LLMs in certain languages and for specific tasks. This is because techniques like pruning and quantization, can target parts of the model that are focused towards infrequent, but not necessarily insignificant data patterns.[47] This is particularly acute for the "long-tail" of data – the less frequent or prevalent patterns within a dataset – which often includes low-resource languages or under-represented demographic groups.[48, 49] Researchers and model developers must therefore carefully balance how to increase efficiency with performance, focusing on what *kinds* of performance must be retained if accuracy is compromised as models are pruned or compressed.

---

[41] Aryabumi, V. *et al.* (2024) 'Aya 23: Open Weight Releases to Further Multilingual Progress'. arXiv. https://doi.org/10.48550/arXiv.2405.15032.
[42] Zadouri, T. et al. (2023) 'Pushing Mixture of Experts to the Limit: Extremely Parameter Efficient MoE for Instruction Tuning'. arXiv. https://doi.org/10.48550/arXiv.2309.05444.
[43] Ahmadian, A. et al. (2023) 'Intriguing Properties of Quantization at Scale'. arXiv. https://doi.org/10.48550/arXiv.2305.19268.
[44] Ogueji, K. et al. (2022) 'Intriguing Properties of Compression on Multilingual Models'. https://doi.org/10.48550/arXiv.2211.02738.
[45] Mince, F. et al. (2023) 'The Grand Illusion: The Myth of Software Portability and Implications for ML Progress'. arXiv. http://arxiv.org/abs/2309.07181.
[46] Hooker, S. (2020) 'The Hardware Lottery'. arXiv. https://doi.org/10.48550/arXiv.2009.06489.
[47] Hooker, S. *et al.* (2020) 'Characterising Bias in Compressed Models'. arXiv. https://doi.org/10.48550/arXiv.2010.03058.
[48] Hooker, S. (2021) 'Moving beyond "algorithmic bias is a data problem"', *Patterns*, 2(4), p. 100241. https://doi.org/10.1016/j.patter.2021.100241.
[49] Ahia, O., Kreutzer, J. and Hooker, S. (2021) 'The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation'. arXiv. https://doi.org/10.48550/arXiv.2110.03036.

Nevertheless, applying techniques such as these can reduce the compute required to run an AI model, and therefore reduce the energy demands and any associated carbon or other climate impacts. But model efficiency matters for more than just energy demands and carbon. Increasingly large-scale and compute-intensive AI models also present barriers to access, as many researchers and organizations lack the resources, funding, or skills to access and make use of large-scale models, because they cannot afford to build their own compute infrastructure at-scale to run models themselves, or because access to commercial model APIs is financially prohibitive.[50] This means that many potential beneficial applications of AI models may be disproportionately distributed, with well-resourced organizations in developed nations making use of AI, while communities in less-developed regions remain underserved.

Addressing the AI Access Gap

At Cohere For AI, in addition to advancing fundamental research on model efficiency, we develop resources and provide support to widen access to AI models, such as through our Aya models and datasets and our Grants Program.

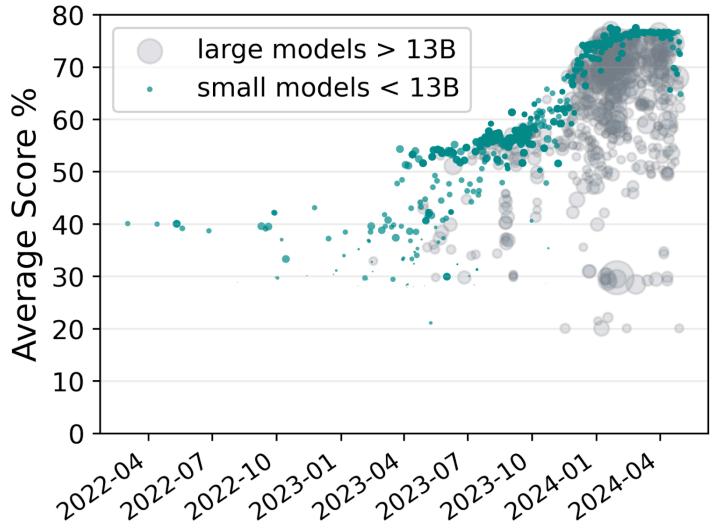
Aya is a multi-year, global initiative to advance the state-of-the-art in multilingual AI and bridge gaps between people and cultures across the world. Involving over 3,000 independent researchers across 119 countries, Aya is an open science project to create new models and datasets that expand the number of languages covered by AI. To date, we have released a family of 3 models, which we have released openly and optimized for efficiency, so that they can be used without the need for high-performance compute.

Our Grants Program addresses the fact that access to the resources needed to conduct machine learning research, such as compute and state-of-art large language models (LLMs), is not always evenly distributed. Not all research institutions have their own compute clusters or expertise to run models, and many existing research grants don't allow researchers to spend their funds on accessing proprietary models. This is especially acute for researchers outside of well-funded institutions and labs (largely in North America), but it is also the case for many researchers outside of computer science departments who want to apply LLMs to their research fields – from health to education. In an effort to help narrow this gap, Cohere For AI launched our Research Grant program in July 2023. These grants provide academic researchers and developers with subsidized access to the Cohere API to support their research into advancing safe, responsible LLM capabilities and applications. An added benefit of such approaches is that instead of different organizations serving their own models individually, large organizations like Cohere offer optimized inference for models, meaning much more efficient utilization of compute and power.

[50] Ahia, O., Kreutzer, J. and Hooker, S. (2021) 'The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation'. arXiv. https://doi.org/10.48550/arXiv.2110.03036.

4. Trends Towards Smaller Models

The very recent history of AI development has focused on larger and more compute-intensive models. This is because, to-date, performance increases have generally come from increasing the size and scale of a model: in effect "throwing compute" at the problem.[51] However, this trend is shifting. While in the near- to medium-term it is likely that models will continue to get larger as researchers scale compute to continue making performance gains, it is likely that increasingly smaller, more task-specific, and therefore more efficient models will start to emerge as methods for increasing efficiency continue to advance. This is evidenced by recent trends of smaller models matching or outperforming larger ones, as seen in submissions to the OpenLLM leaderboard since April 2022.[52] This is also exemplified by Cohere For AI's own multilingual model, Aya Expanse, which is an open-weight 32bn parameter model that outperforms several significantly larger models on a range of languages, including Llama (400bn parameters) and Mistral Large 2 (123bn).[53, 54]



Smaller models submitted to the OpenLLM leaderboard from April 2022 to April 2024 show matching or increased performance against many larger models.

There are several factors that may be motivating this trend towards more efficient models:

1. Efforts to minimize financial costs associated with developing and deploying AI models, in terms of hardware needed to run them and the energy needed to power that hardware.
2. Efforts to minimize the climate impact of AI models, in terms of the carbon emissions associated with energy used for training and inference.

[51] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv. http://arxiv.org/abs/2407.05694. p6.
[52] Hooker, S. (2024) 'On the Limitations of Compute Thresholds as a Governance Strategy'.
[53] Cohere For AI (2024) Aya Expanse: Connecting our world, https://cohere.com/blog/aya-expanse-connecting-our-world
[54] See: Scale AI's SEAL leaderboard, https://scale.com/leaderboard

3. Efforts to widen access to AI models for research and development across settings that lack the resources (hardware, funding, talent) as well as enable deployment of AI models to consumer or edge devices which have lower compute power.

While (1) is driven by market forces, such as commercial users seeking the most cost effective option, that may not lead to (2) and (3) automatically. For example, model efficiency is not the only contributing factor to AI's energy and carbon impact – choices made by data center providers about hardware types have an impact too, and are often outside of the control of those developing and using models. Nevertheless, the benefits of improving model efficiency remain, as efforts to reduce the computational requirements of models will scale as models are deployed and widely.

5. Conclusion & Considerations

The growing energy demands of AI models and their associated climate impacts are a pressing challenge. While methods to accurately and systematically measure AI models' energy and carbon intensity are still nascent, efforts to improve their efficiency are growing in their effectiveness and maturity. A range of techniques, including those outlined in this primer, can reduce the compute required to run an AI model, and therefore reduce its energy demands.

As AI models are increasingly deployed in practice and at scale across a range of real-world applications, the approaches to developing improving model efficiency outlined in this paper offer a range of considerations for those developing, using, and governing AI technologies:

1. **Match the size of the model to the scale of the task**. Not all AI use cases require the largest, most performant models: those adopting AI must be careful not to use sledgehammers to crack nuts. It is crucial to ensure that the size of the model deployed is proportionate to the scale of the task to which it is applied, which means calculating models' energy footprints and requirements.
2. **Standardize methods for measuring and reporting model efficiency.** To achieve (1), model developers need to reduce models' compute requirements, and AI adopters need to choose the more efficient models. Key to this is developing standardized ways to measure and report models' compute or energy requirements. This will allow easy comparison between models based on their efficiency. Initiatives such as AI energy scores offer an approach towards this.[55]
3. **Advance efforts towards more efficient models.** Model developers, AI adopters, policymakers should continue to support efforts and make investments in model efficiency across all layers of the stack, from hardware innovation to model architecture choices and optimizations. This is key to not only managing AI energy and carbon footprint, but also to enable broader access to AI for underresourced communities.

[55]Luccioni, S. *et al.* (2024) 'Light bulbs have energy ratings — so why can't AI chatbots?', *Nature*, 632(8026), pp. 736–738. https://doi.org/10.1038/d41586-024-02680-3.