

Who Is It For.

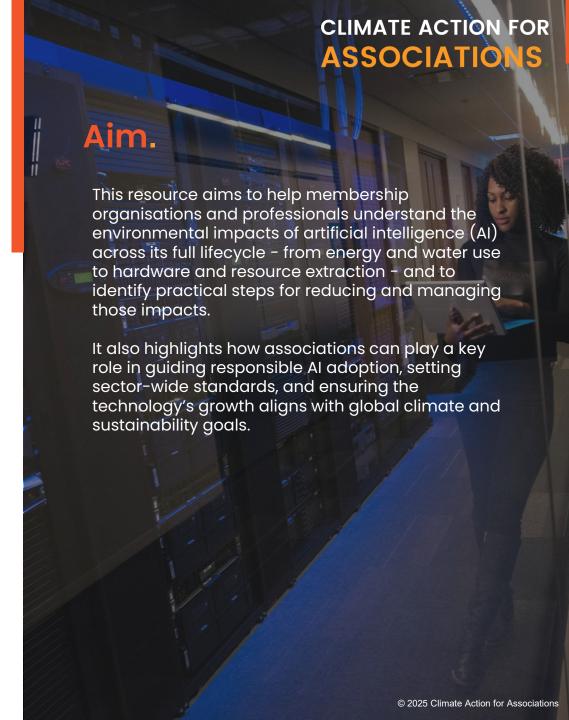
This resource is for staff within membership organisations and their members (including sustainability managers, SLTs, other) who are looking to:

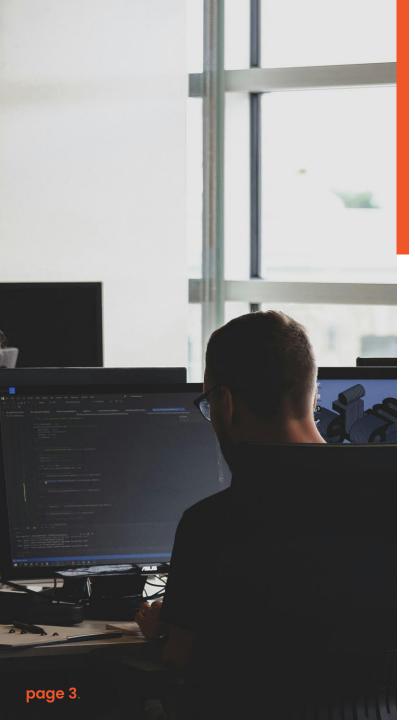
- Create a responsible AI use policy
- Use AI responsibly
- · Advocate for responsible use of Al
- Understand the impacts of AI and how to mitigate them
- Engage with stakeholders on responsible use of Al.

Can I share this with members?

This guide is a condensed summary of recent science on the environmental impacts of Al. It can:

- Be quoted as long as CAFA is credited as a reference
- Be shared with your members in its pdf format
- It can therefore not be plagiarised and it is not intended for public distribution.





CLIMATE ACTION FOR ASSOCIATIONS.

Introduction.

Over the past few years, artificial intelligence (AI) has followed an incredible path from research to commercial maturity which has sent shockwaves across the economy – not only due to its impacts on human creativity, but also because of the reactions it has prompted in people around the world.

Al has inspired awe over its ability to answer the most technical prompts with great results – and it sometimes equally amuses or frustrates us with flaws that draw a clear line between human and artificial intelligence. Although, such flaws seem to be increasingly less obvious or frequent as time goes on – a testament to the speed at which the sector is evolving.

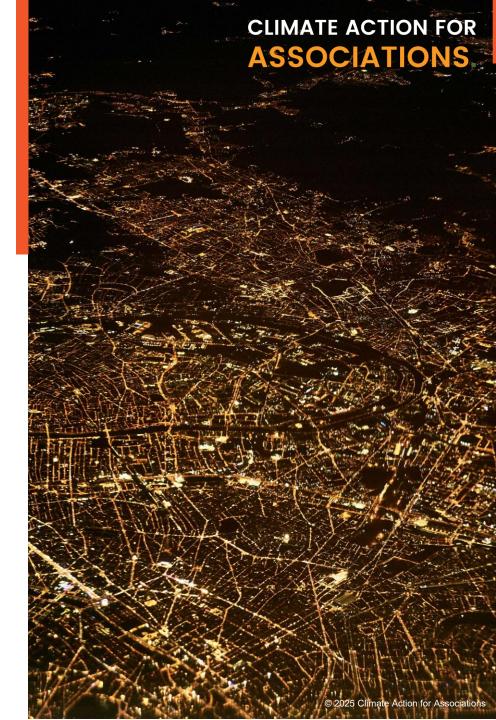
Despite this evolution – or precisely because of it – AI has also become a rising source of concern, and rightfully so. Once a frontier technology, AI now notoriously knows no bounds. Notwithstanding the world's need for sufficiency and moderation, on one hand, it is being used more and more widely: from cybersecurity to creating recipes, from solving complex research problems to correcting emails, from writing articles and papers to crafting content and memes. Even for professionals, AI is everywhere. Notetaking tools, content creation, chat bots, data records, and even algorithms – it is becoming a new layer to every single digital tool we use.

On the other hand, AI is always improving and is therefore fuelling the development of new technologies and products, from new AI models to humanoid robots. From where we stand in time, AI's path forward looks like an exponential one – just like its homonym in physics, a singularity in AI could very much absorb an infinity of energy and resources. And even before such an event occurs, the expansion of AI could generate astronomical costs due to its growth curve.

A possibility that seems to be on the horizon for Sam Altman himself – in 2025 the CEO of OpenAI told the US Congress that the number one most important and fruitful investment the government could make for AI development, was to invest in more energy.

But what is the environmental cost of this expansion? In a world where climate change and environmental risks still aren't taken as seriously as they should be, at a time when our attention should be focused on preventing avoidable harm to happen to people and businesses around the world in the form of said environmental risks, and in a context where energy (both literally and figuratively) must be carefully allocated to priority areas, the rapid growth of a *new* sector that could affect all others is a source of concern.

As such, many businesses and membership organisations are looking for information and data on AI – that's where this resource comes in.







The environmental impacts of Artificial Intelligence (AI) are substantial and span the entire AI lifecycle, encompassing energy, water, and mineral resources, and resulting greenhouse gas (GHG) emissions (Luccioni, Trevelin, and Mitchell 2025).

While early concerns focused on initial development, the deployment and everyday use (called "inference") of these systems now present the primary environmental challenge (Desislavov, Martínez-Plumed, and Hernández-Orallo 2023; Moore et al. 2025).

Inference costs typically exceed training costs because of the multiplicative factor of using the system many times, and inference can account for up to 90% of a model's total lifecycle energy use (Desislavov, Martínez-Plumed, and Hernández-Orallo 2023; Jegham et al. 2025). For highly used models like GPT-4, the carbon emissions from serving inferences for just 121 days can equate to the emissions from its initial training (Fu et al. 2024).



1. Energy Use.

Energy demand needs to fall – because energy is predominantly fossil fuel based, and low carbon alternatives are not deploying fast enough (Berners-Lee 2025). Yet, energy demand is the foundational environmental cost of AI, driven by the vast computational resources required by modern AI models, often housed in power-hungry data centres (Huang et al. 2025).

Al queries, particularly those utilising large language models (LLMs) like ChatGPT, are far more energy-intensive than traditional computing tasks (EPRI 2024; Luccioni, Trevelin, and Mitchell 2025).

A single ChatGPT query is estimated to consume about 2.9 watt-hours (Wh) of electricity, nearly ten times the electricity used for a traditional Google search, which uses about 0.3 Wh each (Brandtech Group 2025; EPRI 2024; Luccioni, Trevelin, and Mitchell 2025).

The complexity of the task is the primary determining factor of direct energy and carbon impact (Brandtech Group 2025). Generative tasks (such as creating text, images, or captions) consume significantly more energy than discriminative tasks (such as classification or simple question answering) (Luccioni, Jernite, and Strubell 2024).





For instance, energy demand can be 50 to 25,000 times higher for an LLM than a traditional machine learning model, and 50 to 100 times higher for image generation than for text generation (The Shift Project 2025).

Generally speaking, tasks involving images, like image generation, tend to be the most energy-intensive tasks overall, with image-based tasks requiring more energy than text-based ones (Brandtech Group 2025; Luccioni, Jernite, and Strubell 2024).

Deployment infrastructure is a critical determinant of AI sustainability and energy use (Jegham et al. 2025). For instance, GPT-40 mini consumes approximately 20% more energy than GPT-40 due to reliance on older AI00 GPU nodes, illustrating that deployment infrastructure can overshadow model size in determining real-world energy use (Jegham et al. 2025).

- Energy demand needs to fall
- Energy demand is rising in part because of data centres including AI use
- Not all AI tasks and models and processes are created equal:
 - o Use of AI accounts for 90% of its impact over its lifecycle
 - simple ML (machine learning) can require 50-25,000x less energy than LLMs (large language models)
 - Text can require 50-100x less energy than image generation.
 - Video generation, while not as studied as image and text generation, can be expected to use more energy by at least one or several orders of magnitude.
- Al providers are not transparent about their energy use.

2. Carbon and Greenhouse Gas (GHG) Emissions.

Emissions are divided into two main categories: operational carbon refers to the emissions generated by the electricity consumed during the hardware's use phase, and embodied carbon stems from the emissions associated with the manufacturing and disposal of the physical hardware (Huang et al. 2025; Luccioni, Trevelin, and Mitchell 2025).

Even the initial training of large AI models results in substantial carbon footprints. Training a model like GPT-3 was estimated to produce around 550 metric tons of CO2e, which is roughly equivalent to the lifetime emissions (including fuel) of five average cars (Jegham et al. 2025; Luccioni, Trevelin, and Mitchell 2025; Strubell, Ganesh, and McCallum 2019).

Operational carbon emissions are highly variable because they depend heavily on the carbon intensity of the local electricity grid supplying the data centre, which is the main factor influencing the final quantity of emissions (Huang et al. 2025; Luccioni, Trevelin, and Mitchell 2025). Converting energy consumption to carbon emissions using location–specific carbon intensity factors is crucial for accurate assessment (Huang et al. 2025; Nguyen et al. 2024). Identical AI tasks performed in different geographic regions can show huge differences in generated carbon, sometimes approaching a 15-fold variance (Huang et al. 2025).





Current data centre design and deployment could lead to 630-920 MtCO₂e emissions by 2030 (which is up to twice the entire emissions of France). This would be in large part due to the development of AI, which could increase data centre power demand by 60% or more by 2030, which are themselves often powered by fossil fuels.

Tech firms in the US are increasingly building gas power plants to meet the energy needs of data centres, with more than 80 gas power plants currently in development for that purpose (The Shift Project 2025). et al. 2025).

- Greenhouse gas emissions from AI are largely due to energy use, because AI is energy intensive, and most energy grids are still predominantly reliant on fossil fuels.
- Greenhouse gas emissions from AI are largely due to demand, because AI is driving the deployment of new data centres and therefore new fossil fuel projects.

3. Water Consumption.

Water is a critical resource consumed by the AI ecosystem, primarily for cooling data centres (Luccioni, Trevelin, and Mitchell 2025; Moore et al. 2025). Data centres house hundreds of thousands of servers performing intensive computation, requiring constant cooling to prevent overheating (Luccioni, Trevelin, and Mitchell 2025). One key method involves pumping clean water through radiators in the data centre, which absorbs the heat; a significant portion of this water evaporates in the process (Luccioni, Trevelin, and Mitchell 2025).

Historical estimates suggest that training models like GPT-3 required more than 700 kilolitres (kL) of water for cooling alone (Jegham et al. 2025). When scaled up (e.g., GPT-40 serving hundreds of millions of daily requests), the annual water consumption can be equivalent to the annual drinking needs of almost 1.2 million people (Jegham et al. 2025). Water is also used in the hardware manufacturing process, specifically for rinsing the different layers of semiconductor wafers that form the CPU and GPU chips (Luccioni, Trevelin, and Mitchell 2025).

Key Takeaway

• Water is used for cooling datacentres and manufacturing hardware. Water use therefore increases with demand.







Building sophisticated components like Graphics Processing Units (GPUs) requires raw materials and minerals, including silicon, aluminum, copper, tantalum, lithium, gallium, germanium, palladium, cobalt, and tungsten (Luccioni, Trevelin, and Mitchell 2025). Mining these metals results in environmental costs, since hundreds of tonnes of ore typically need to be processed to get a single ton of relatively common metals such as copper or aluminum (Luccioni, Trevelin, and Mitchell 2025).

Key Takeaway

 Raw materials and minerals are used for manufacturing hardware. Resource depletion and associated environmental costs (pollution, greenhouse gas emissions, biodiversity loss) increase with demand.



5. Overall Trends and Sustainability Paradox.

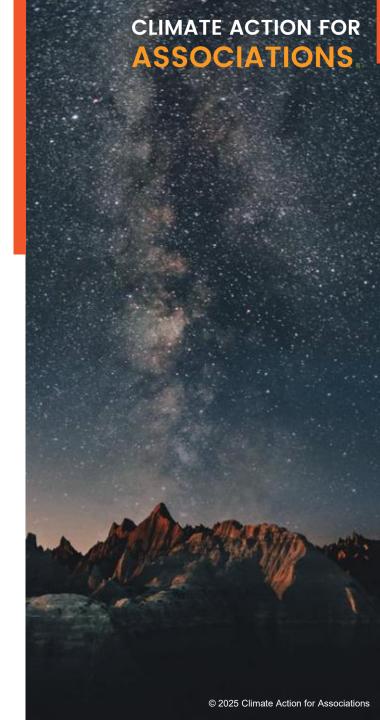
A key challenge is the rapidly increasing ubiquity and usage of Al. Although hardware and algorithms are constantly becoming more energy-efficient, the total environmental impact often continues to rise because the efficiency gains are overwhelmed by the explosion in total Al usage, a phenomenon known as the rebound effect or Jevons paradox (Brandtech Group 2025; Jegham et al. 2025; Polimeni and Polimeni 2006; Stojkovic et al. 2024). As Al becomes cheaper and faster, overall demand expands, intensifying environmental strain (Jegham et al. 2025). In essence, as Al becomes easier to access and use, people and businesses use it more – for all its applications, from surveillance and security to content creation and data analysis, which in turn creates new demand (The Shift Project 2025).

The unsustainability of AI is therefore mainly driven by the same dynamic that hides behind the unsustainability of other sectors and industrial human activities in their current form: the more demand there is for a product or service, the more supply increases. And the more supply increases, the more demand increases.

This is because of a two-dimensional effect.

1. When demand increases, there is:

- a. An incentive to meet that demand with competitive prices and value
- o. An incentive to optimise production to make efficiency gains for lower costs
- Financial means to scale up supply by way of creating new companies or scaling existing activities – such financial means arise from the extra demand boosting the confidence of investors and banks.





2. When supply increases in multiple sectors at once (which is almost always the case), products and services become more widely available to support the development of other technologies and products, which helps scale up supply and create new niche applications and products, creating new markets and by extension, new demand.

This, in turn, creates new and cheaper offerings in larger quantities, that can find new applications in new markets that hadn't been reached before. In turn, these markets also increase their demand, calling for more scaling, more cost-efficiencies, and more product specialisation. Al is following that exact path – and it's exceptionally good at following it. The internet, the web, smartphones, laptops have provided a foundational infrastructure for Al to move from basic machine learning used for example for algorithmic purposes, to highly capable LLMs available on-demand on a smartphone. This was largely amplified by OpenAl's strategy of opening its first LLM to the public, and to the "wow" effect that drew in what probably is excessive investment from both public and private sources.

- Demand and supply pull each other up in feedback loops: supply increases to meet demand, demand increases to adapt to new level of supply.
- Demand rebounds higher than efficiency gains as per the Jevons effect (or rebound effect).



What Can Users Do to Reduce Their Footprint?

While the systemic nature of Al's environmental footprint is daunting, there remain concrete steps that individuals, organisations, and industry leaders can take to mitigate it:

- 1. Big Picture Management of the organisation's footprint
- 2. Adjust Usage and Queries: ensure AI is used responsibly and efficiently
- 3. Optimise Model Selection: ensure the right model is used for the right task
- 4. Consider Infrastructure Deployment: where possible, mitigate the impact of data centres by limiting use of cloud and preferring providers operating in countries with low carbon grids
- 5. Advocate for Transparency, Accountability, and Decarbonisation: ask Al providers to disclose their emissions, and advocate for decarbonisation of energy.

1. Big Picture Management.

To ensure best AI usage practices are put in place in a given organisation, the organisation needs a clear view over its footprint and how AI contributes or might contribute to it.

- Baseline and Monitor the Organisation's Footprint: By measuring emissions and environmental footprint, users can track the evolution of said footprint (increases or reductions) which will in turn support the organisation's strategy for managing this footprint for AI as well as for other technologies and areas of the value chain.
- **Define a Carbon Budget**: An environmental baseline measurement should be used to plan science-based targets & reductions, and a corresponding carbon budget for the organisation. This will in turn help define a sub-budget for Al use specifically, allowing the footprint to be managed.
- **Define an Organisation Wide Policy**: Based on the footprint (see above) and on options for mitigation (see below), define a policy for use of Al. This will be useful not only for environmental mitigation but likely for data protection too.

- Measure and monitor the organisation's footprint.
- Define an environmental budget for Al.





CLIMATE ACTION FOR ASSOCIATIONS.

2. Adjust Usage and Queries.

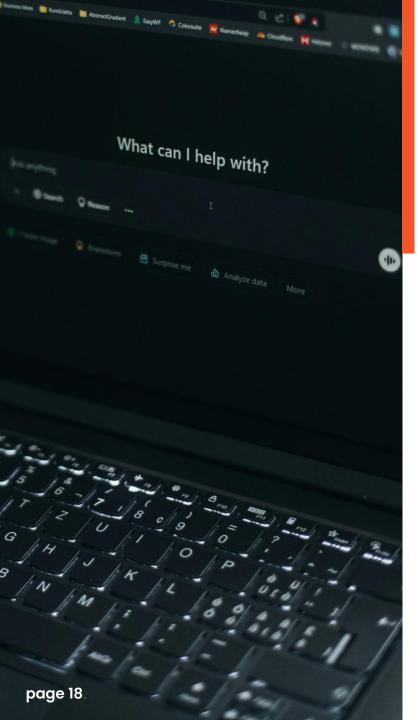
Since inference (deployment and use) accounts for the largest share of an AI model's lifetime costs, altering how the system and alternatives are used can save energy.

- Identify Alternatives: Before using AI, determine if a traditional search engine (e.g., Google) would suffice. Avoid using high-energy-consuming AI models for simple tasks. A task specific AI model e.g. for translation, is far more efficient that a general model like GPT.
- Avoid Use of Image and Video Generation: Image generation consumes up to 100 times more energy than text generation. This is dependent on image size, resolution and complexity. It's also fair to say that AI images lose impact and purpose. Not only that, graphic designers and animators do a better, more authentic job!
- Optimise Prompts to Reduce Iterations: Improving prompt efficiency with generative AI saves energy by reducing the number of iterations required to achieve a satisfactory result, thereby limiting energy-hungry generations (Brandtech Group 2025).
- Be Concise and Limit Output Size: Energy consumption and runtime are strongly correlated with the size of the output (Brandtech Group 2025; Luccioni, Jernite, and Strubell 2024). Users can ask for shorter outputs in prompts, using phrases like "be concise" or "limit your answer to approximately 100 words," to reduce impact and cost (Brandtech Group 2025).

- Avoid Overuse of Generative Tasks: Recognise that tasks that generate new content (text generation, summarisation, image generation) are significantly more energy and carbon intensive than simpler discriminative tasks (Luccioni, Jernite, and Strubell 2024).
- Start New Chats for Unrelated Questions: Longer chat sessions consume more energy because the model must process the entire history (all previous tokens) for subsequent generations (Brandtech Group 2025).
- Regularly Delete Old Chats: Just like emails and inactive tabs, old chats use energy due to needing to be stored. Data storage in paper is passive in the sense that once the data is stored, it remains stored without additional energy for the data to keep being stored. Data storage in data centres is active, in the sense that energy is continuously required to keep the data available for retrieval.

- Prioritise other tools where possible.
- Craft prompts and tasks carefully for maximum efficiency.
- Avoid image and video generation.





3. Optimise Model Selection.

Choosing the right model for the task is often the most significant action users can take, as models vary wildly in energy demand (Luccioni, Jernite, and Strubell 2024).

- Prefer Task-Specific Models over General-Purpose LLMs: For tasks that are specific (such as text classification or simple question answering), using smaller models fine-tuned for a task is orders of magnitude more energy-efficient than relying on large, multi-purpose generative AI models (Luccioni, Jernite, and Strubell 2024).
- Opt for Smaller Models When Possible: Utilising models with fewer parameters generally reduces energy consumption (Brandtech Group 2025). However, users must recognise that deployment infrastructure can sometimes outweigh model size: GPT-40 mini, despite being substantially smaller, consumed more energy than GPT-40 because it relied on less efficient A100 GPU nodes instead of H100s or H200s (Jegham et al. 2025).
- Opt for models with good environmental input to performance ratio. These are ranked as follows by Jegham et al. (2025) using a DEA analysis (Data Envelopment Analysis (DEA) measures the eco-efficiency of models based on performance vs environmental cost).

Top ranked AI Models:

- 1. Claude-3.7 Sonnet 0.886 (highest eco-efficiency).
- 2. o4-mini (high) 0.867.
- 3. o3-mini 0.840.
- 4. GPT-4.1 mini 0.802.
- 5. GPT-40 0.762.

(These five are the explicitly highlighted top performers in Jegham et al. DEA discussion.)

Mid-range eco-efficiency:

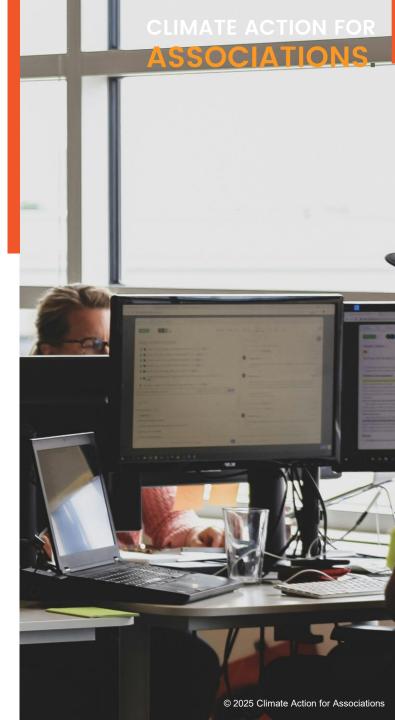
• LLaMA family (various sizes) - LLaMA models are resource-efficient but modest in performance, which limits eco-efficiency.

Least-efficient (bottom of DEA)

- GPT-4.5 described by Jegham et al as among the least efficient (no single numeric DEA value reported).
- DeepSeek-V3 0.060 (one of the two lowest scores).
- DeepSeek-R1 0.058 (lowest score reported).

(Jegham et al. also call out o3 and DeepSeek-R1 as extremely energy-intensive in per-prompt Wh for long prompts; their DEA positions are toward the low end as well.)

- Prefer specialised models.
- Prefer smaller models.
- Prefer models with a high performance-impact ratio.





4. Consider Infrastructure Deployment.

The physical location and method of deployment are critical factors in minimising operational carbon and water use.

- Choose Low-Carbon Data Centre Locations: The carbon emissions from identical AI tasks can show a variance approaching 15-fold depending on the carbon intensity of the local electricity grid (Huang et al. 2025). When given the option (e.g., using configurable cloud solutions), users should select data centres in regions utilising low-carbon energy sources (Brandtech Group 2025).
- Prefer Self-Hosted Solutions When Possible: Deploying open-source or private models to a configurable cloud platform enables users to select low-carbon locations (Brandtech Group 2025).

- If possible, choose data centres located in low-carbon grids.
- · Host AI and LLMs on premises and private servers.



5. Advocate for Transparency, Accountability, and Decarbonisation

To facilitate better environmental decision-making, users and practitioners should demand greater reporting transparency across the AI ecosystem. Additionally, advocating for decarbonisation will can lead to more impactful collective change and increased trust in the impact of individual changes.

- Advocate For A Low Carbon Economy: Transitioning energy grids from fossil fuels to renewable and low carbon sources, as well as reducing overall energy use through what is sometimes called efficiency, is at the core of decarbonising many sectors, not just Al. By advocating for such change, users can have an impact over several sectors, while creating a shared culture of change and a shared expectation of fossil fuel phase out. This can in turn affect demand for energy intensive products and services more generally.
- Demand Transparent Reporting: Encourage AI developers and providers to offer transparent reporting of per-inference energy, water, and carbon metrics (Luccioni, Trevelin, and Mitchell 2025; Jegham et al. 2025). The absence of this information makes it difficult for users to make informed, efficient choices (Luccioni, Jernite, and Strubell 2024).





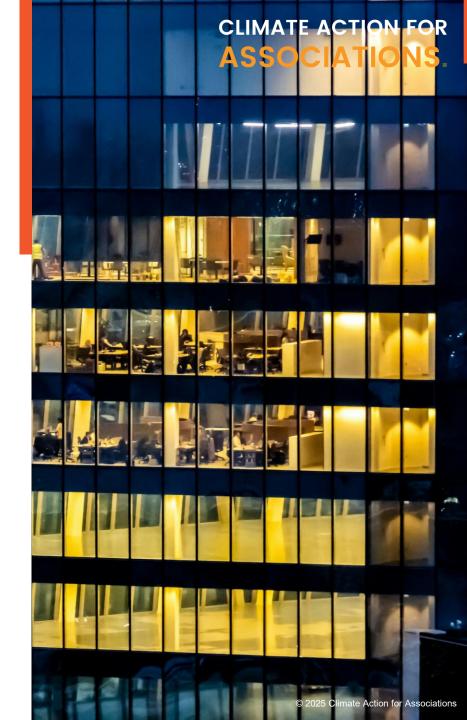
- **Promote Eco-Efficiency Metrics**: Advocate for the use of evaluation standards that contextualise model capability against its environmental cost (Jegham et al. 2025).
- Report Compute Details in Research: Researchers should report the training time, computational resources, and model sensitivity required for their models to enable direct comparison and assessment of environmental cost (Strubell, Ganesh, and McCallum 2019).

- Advocate for phasing out fossil fuels, reducing energy use, and for decarbonisation.
- Advocate for transparency of energy efficiency and environmental footprint of models.

The Role of Membership Organisations.

Membership organisations have a tremendous role to play in mitigating the impact of Al. For instance, membership organisations can:

- Advocate for responsible AI policy, such as:
 - Setting explicit carbon/energy budgets for the data centre/Al sector and integrating them into national decarbonisation strategies
 - Monitoring and data centre / Al growth management
 - Abandoning deployments that cannot stay within science aligned trajectories.
 - Regulation and infrastructure planning to ensure data centre growth doesn't conflict with other transition priorities (electricity grid, land/water resources, etc.)
- In coordination with experts and other membership organisations, either advocate for or carry out estimations of full lifecycle impacts (manufacture, use, infrastructure) of digital/AI services
- Support members on using AI responsibly and sustainably
- Support members on using AI for sustainability where efficiency gains are significant and potential rebound effects are anticipated and managed
- Train professionals and SLTs in responsible AI use
- Embed responsible AI use criteria in professional competency frameworks





- Set sector-wide standards and rules for responsible AI use to:
 - ensure a level playing field
 - avoid competitive runaways (when organisations race to implement new processes and technologies to stay ahead of the competition, to the detriment of the collective good)
 - ensure the functional need of AI systems is justified, and that AI isn't deployed by default or because "it's possible"
- For IT, AI, research and manufacturing sectors: help members choose architectures, hardware, usage models that minimise energy/carbon

(Many of these recommendations were adapted for membership organisations by CAFA and sourced from The Shift Project 2025).

As membership organisations, we have a unique leverage point in shaping the responsible use of AI. By embedding environmental accountability in professional standards, supporting evidence-based policy, and helping members adopt efficient practices, associations can ensure AI innovation supports - rather than undermines - the climate transition.

- Advocate for responsible AI policy.
- Embed responsible AI use in support delivered to members, in the form of resources, training, competency frameworks, etc.
- Develop or promote standards and rules for responsible AI use across your sector/ profession to foster a level playing field.

References.

- Berners-Lee, Mike. 2025. A Climate of Truth: The Climate Crisis and How to Fix It.
- Brandtech Group. 2025. Generative AI Environmental Impact Study.
- Desislavov, Radosvet, Fernando Martínez-Plumed, and José Hernández-Orallo. 2023. "Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning." Sustainable Computing: Informatics and Systems 38: 100857.
- Electric Power Research Institute (EPRI). 2024. Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption.
- Fu, Zhenxiao, Fan Chen, Shan Zhou, Haitong Li, and Lei Jiang. 2024. "LLMCO2: Advancing Accurate Carbon Footprint Prediction for LLM Inferences." Accessed January 22, 2025.
- Huang, Hongzhen, Kunming Zhang, Hanlong Liao, Kui Wu, and Guoming Tang. 2025. "WattsOnAl: Measuring, Analyzing, and Visualizing Energy and Carbon Footprint of Al Workloads."
- Jegham, Nidhal, Marwan Abdelatti, Lassad Elmoubarki, and Abdeltawab Hendawi. 2025. "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference."
- Luccioni, Alexandra Sasha, Yacine Jernite, and Emma Strubell. 2024. "Power Hungry Processing: Watts Driving the Cost of Al Deployment?" In ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24), 85–99.
- Luccioni, Sasha, Bruna Trevelin, and Margaret Mitchell. 2025.
 The Environmental Impacts of AI Policy Primer.

- Moore, Hayden, Sirui Qi, Ninad Hogade, Dejan Milojicic, Cullen Bash, and Sudeep Pasricha. 2025. "Sustainable Carbon-Aware and Water-Efficient LLM Scheduling in Geo-Distributed Cloud Datacenters." In Great Lakes Symposium on VLSI 2025 (GLSVLSI '25).
- Nguyen, Sophia, Beihao Zhou, Yi Ding, and Sihang Liu. 2024. "Towards Sustainable Large Language Model Serving." ACM SIGENERGY Energy Informatics Review 4 (5): 134–140.
- Polimeni, John M., and Raluca lorgulescu Polimeni. 2006. "Jevons' paradox and the myth of technological liberation." Ecological Complexity 3 (4): 344–353.
- Stojkovic, Jovan, Chaojie Zhang, Íñigo Goiri, Josep Torrellas, and Esha Choukse. 2024. "DynamoLLM: Designing LLM Inference Clusters for Performance and Energy Efficiency."
- Strubell, Emma, Ananya Ganesh, and Andrew McCallum. 2019. "Energy and Policy Considerations for Deep Learning in NLP." In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.
- The Shift Project. 2025. IA, données, calcul: quelles infrastructures dans un monde décarboné?.
 https://theshiftproject.org/app/uploads/2025/09/RF-PIA-1.pdf.

