Vanderbilt University Law School

Legal Studies Research Paper Series
Accepted Paper Number 25-11

THE ENERGY AND ENVIRONMENTAL FOOTPRINT OF AI

Professor Michael P. Vandenbergh
Vanderbilt Law School
Ethan I. Thorpe
Climate Governance Fellow
Vanderbilt Law School
Professor Jonathan Gilligan
Vanderbilt University

Article

The Energy and Environmental Footprint of AI

Michael P. Vandenbergh,[*] Ethan I. Thorpe[**] & Jonathan M. Gilligan[***]

15 MICH. J. ENV'T. & ADMIN. L __ (forthcoming)

Abstract

Artificial intelligence (AI) has the potential to create major economic and social benefits but also to rapidly escalate electricity demand and its associated environmental impacts. Information has been a cornerstone of environmental law for half a century, and this Article argues that providing information to individual, corporate, and other users about the electricity demand and environmental impacts of AI can reduce those impacts without delaying development of the technology. Little is known about how AI large language models (LLMs) compare on these issues, though, and to address this shortcoming the Article provides the first comparison of the outputs of four AI environmental footprint calculators. The Article finds that running the same AI query through all four calculators produces substantial differences in outputs, with one calculator producing an estimate more than 50 times higher than another for the same type of query. These differences suggest that substantial improvements are needed in the disclosure of information, whether through international, national, state, or private standards, to provide reliable estimates of energy use and environmental impacts to users. In turn, more accurate, easily available information can create incentives for reducing the costs, energy demand, and environmental impacts of AI even in a deregulatory era.

---

[*] David Daniels Allen Distinguished Chair of Law, Vanderbilt University Law School, Director, Climate Change Research Network, and Co-Director, Energy, Environment and Land Use Program. For comments, we thank Emily Moburg, J.B. Ruhl, David Stein, and Mark Williams, and the participants at the Artificial Intelligence (AI) and Energy Agenda Roundtable co-sponsored by the Vanderbilt Policy Accelerator—Energy, Environment and Land Use Program. Excellent research assistance was provided by Sam Holmes, Hannah Kupfer, Jane Elizabeth Miller, and Gabrial Xiong. Financial support was provided by the Vanderbilt University Law School Dean's Fund and the Sally Shallenberger Brown EELU Program Fund. The opinions expressed in the Article are solely those of the authors.

[**] Fellow, Private Climate Governance Lab, Energy, Environment and Land Use Program, Vanderbilt University Law School,

[***] Professor of Earth & Environmental Science and Civil & Environmental Engineering, and Associate Director, Climate Change Research Network, Vanderbilt University Law School.

I. Introduction

Information is the foundation of environmental governance,[1] yet surprisingly little information is publicly available about the most important and energy-intensive development in decades: artificial intelligence (AI). More than half of American adults have used Large Language Models (LLMs), with 34% of LLM-users doing so at least once per day.[2] These tools are not just for personal use. In the second half of 2024, 78% of businesses reported using generative AI, a broader category that encompasses LLMs and other generative technologies, for at least one function.[3] AI is even helping to construct the next generation of programming as over 63% of professional programmers reported using generative AI tools for their code in mid-2024.[4]

As AI becomes mainstream, concerns have emerged about its environmental footprint.[5] Data centers, large facilities stocked with state-of-the-art computer chips, are necessary to perform the computationally intense calculations underlying AI.[6] They must operate constantly because

---

[1] *See, e.g.*, Daniel C. Esty, *Environmental Protection in the Information Age*, 79 N.Y.U. L. REV. 115, 115 (2004) (noting the relationship between emerging technologies and environmental information disclosures); Cass R. Sunstein, *Informational Regulation and Informational Standing: Akins and Beyond*, 147 U. PA. L. REV. 613, 614 (1999) (exploring information disclosure in environmental statutes).

[2] Lee Rainie, *Close Encounters of the AI Kind: The Increasingly Human-like Way People Are Engaging with Language Models*, IMAGINING THE DIGITAL FUTURE CENTER (Mar. 2025), https://imaginingthedigitalfuture.org/reports-and-publications/close-encounters-of-the-ai-kind/close-encounters-of-the-ai-kind-main-report/.

[3] Alex Singla et al., *The State of AI: Global Survey*, MCKINSEY & CO. (Mar. 12, 2025), https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[4] *AI | 2024 Stack Overflow Developer Survey*, STACK OVERFLOW (2024), https://survey.stackoverflow.co/2024/ai#sentiment-and-usage-ai-sel-prof.

[5] *See* Amy L. Stein, *Artificial Intelligence and Climate Change*, 37 YALE J. ON REG. 890, 891 (2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3665760 (examining the growing role of artificial intelligence in the economy, security, and law); James O'Donnell & Casey Crownhart, *We did the math on AI's energy footprint. Here's the story you haven't heard*, MIT TECH. REV. (May 20, 2025), https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/; Pranshu Verma & Shelly Tan, *A bottle of water per email: the hidden environmental costs of using AI chatbots*, WASH. POST: POWER GRAB (Sept. 18, 2024), https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/; Renée Cho, *AI's Growing Carbon Footprint*, ST. OF THE PLANET (Jun. 9, 2023), https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/.

[6] These can be truly massive; large data enters can be millions of square feet and house tens-of-thousands of chips. *See* Sean Farmer, *The Stone in the Cloud: Planning the Resource Demands of Data Centre Industry through Land Use Law*, 56 U.B.C. L. REV. 435 (Nov. 2023), https://commons.allard.ubc.ca/ubclawreview/vol56/iss2/3/ ("a data centre is a building that contains equipment running software designed to "process data requests-to receive, store and deliver-to 'serve' data, such as games, music, emails and apps, to clients over a network."); Alexandra Jonker & Alice

power outages can cause data loss and other problems.[7] US data centers consumed 176TWh in 2023,[8] equivalent to the average annual consumption of 16.3 million American households.[9] It is hard to disaggregate AI from other workloads that data centers are responsible for (a report from the International Energy Agency estimates that AI is responsible for ~15% of total data center energy), but AI is *the* key driver for projected increases in overall data center energy demand.[10] This figure may triple to 580TWh by 2028,[11] more than all the electricity use by Germany, France, or Brazil in 2019.[12] The North American Electric Reliability Corporation, responsible for ensuring reliable electricity supply across the continent, has warned that growing electricity demand from data centers and increased cooling needs due to climate change could cause rolling blackouts across most of the US.[13]

Generating this electricity emits millions of tons of carbon, including 61 million metric tons (MMT) in the US in 2023.[14] This is equivalent to adding an additional 13 million cars to the road for a year. [15] If current emissions intensities persist, this figure could reach 197 MMT by 2028, equal to about half of total residential emissions in 2022.

Gomstyn, *What Is an AI Data Center?*, IBM (Feb. 21, 2025), https://www.ibm.com/think/topics/ai-data-center.

[7] Andreja Velimirovic, *Data Center Power Outage*, PHOENIXNAP (Sept. 12, 2024), https://phoenixnap.com/blog/data-center-power-outage.

[8] ARMAN SHEHABI ET AL., 2024 UNITED STATES DATA CENTER ENERGY USAGE REPORT, at 5 (2024), https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf.

[9] U.S. ENERGY INFO. ADMIN., HOW MUCH ELECTRICITY DOES AN AMERICAN HOME USE?, https://www.eia.gov/tools/faqs/faq.php?id=97&t=3 (last updated Jan. 8, 2024).

[10] INT'L ENERGY AGENCY, *Energy Demand from AI*, https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai (last visited Aug. 11, 2025) at 56.

[11] SHEHABI ET AL., *supra* note 8, at 6. It's important to note that energy consumption predictions tend to exceed real energy load growth.

[12] INT'L ENERGY AGENCY, ELECTRICITY INFORMATION: OVERVIEW – ELECTRICITY CONSUMPTION, https://www.iea.org/reports/electricity-information-overview/electricity-consumption (last visited Aug. 11, 2025).

[13] NORTH AM. ELEC. RELIABILITY CORP., 2025 SUMMER RELIABILITY ASSESSMENT (May 2025), https://www.nerc.com/pa/RAPA/ra/Reliability%20Assessments%20DL/NERC_SRA_2025.pdf (finding that most of the US is at high or elevated risk of rolling blackouts).

[14] SHEHABI ET AL., *supra* note 8, 57. Some estimates found that data centers emit 48% more carbon per unit of electricity than the grid, though, which would imply an additional ~30MMT in 2023 and ~100MMT in 2028 barring significant decarbonization. *See* Gianluca Guidi et al., *Environmental Burden of United States Data Centers in the Artificial Intelligence Era*, ARXIV (Nov. 14, 2024), https://arxiv.org/html/2411.09786v1.

[15] U.S. ENV'T PROT. AGENCY, *Greenhouse Gas Emissions from a Typical Passenger Vehicle*, https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle (last updated Jun. 12, 2025).

AI also drives increases in water use. Roughly as much energy goes into cooling the average data center as powering its servers. [16] Data centers generate a great deal of heat, and consequently they require massive volumes of fresh water for cooling.[17] For instance, a small data center can use nearly seven million gallons of water in a single year on direct cooling[18] and an additional 10 million gallons indirectly from cooling the plants that generate the electricity it consumes.[19] A single small data center, therefore, uses enough fresh water each year to cover a football field to a depth of ten feet.[20]

Despite the increasing concerns about the stress on the electric grid, the environmental impacts of AI, and the contribution of AI to raising the price of electricity for residential and business users[21] and causing pressure on household water supplies,[22] federal action is unlikely. A provision banning states from regulating AI for ten years was almost adopted by Congress in 2025 and suggests that comprehensive federal legislation is

---

[16] Karthik Ramachandran, *As generative AI asks for more power, data centers seek more reliable, cleaner energy solutions, Deloitte Insights: Tech., Media & Telecommunications* (Nov. 19, 2024), https://www2.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2025/genai-power-consumption-creates-need-for-more-sustainable-data-centers.html. These figures can vary between data centers. Electricity for cooling is used in air cooling (i.e. air conditioning) and to lower the temperature of water that is used in water cooling. The other 20% is split between internal power conditioning, lighting, communications, and other backend functions. *See also* Henry Gunther & Julietta Rose, *Governing AI: The Importance of Environmentally Sustainable and Equitable Innovation*, 50 ELR 10888, 10889 (Nov. 2020), https://www.elr.info/articles/elr-articles/governing-ai-importance-environmentally-sustainable-and-equitable-innovation ("Google's DeepMind AI, using machine learning to train neural networks to predict the most efficient energy consumption in their data centers, was able to reduce the energy used to cool their centers by 40%.").

[17] *See* Whitni Simpson, *A Need to Regulate the Environmental Impacts of Artificial Intelligence (AI): Preserving Clean Water for Humans, Not Robots*, 38 TUL. ENV'T L.J. 133, 137-39 (Winter 2025), https://www.jstor.org/stable/27378753?seq=1 (discussing the massive amount of freshwater water consumed by AI systems for cooling, producing AI chips, and offsite energy production); Mya Garcia, *AI Uses How Much Water? Navigating Regulation of AI Data Centers' Water Footprint Post-Watershed Loper Bright Decision*, 57 TEX. TECH. L. REV. 517 (2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5064473 ("Google reported that the global water consumption for all of its data centers in 2021 was 'approximately 4.3 billion gallons of water'… The report also states that water cooling is typically the most efficient means of cooling, and has helped lessen data centers' carbon emissions while also mitigating climate change.").

[18] David Mytton, *Data Centre Water Consumption*, 4 NPJ CLEAN WATER 11 (Feb. 15, 2021), https://www.nature.com/articles/s41545-021-00101-w.

[19] SHEHABI ET AL., *supra* note 8, 57. 4.35L/kWh for a 1MW facility operating 8,670 hours/year = 38,106,000L.

[20] Football fields are 5,350 square meters, 17M liters = 17,000 cubic meters, meaning a depth of over 3 meters (>10 feet).

[21] Marc Levy, As electric bills rise, evidence mounts that data centers share blame. States feel pressure to act, ASSOCIATED PRESS (Aug. 8, 2025), https://apnews.com/article/electricity-prices-data-centers-artificial-intelligence-fbf213a915fb574a4f3e5baaa7041c3a; Ivan Penn & Karen Weise, *Big Tech's AI Data Centers Are Driving Up Electricity Bills for Everyone*, N.Y. TIMES (Aug. 14, 2025), https://www.nytimes.com/2025/08/14/business/energy-environment/ai-data-centers-electricity-costs.html.

[22] Eli Tan, *Their water taps ran dry when Meta built next door*, N.Y. TIMES (Jul. 14, 2025), https://www.nytimes.com/2025/07/14/technology/meta-data-center-water.html.

unlikely in the near term.[23] In addition, the Trump White House has adopted executive orders discouraging regulation of AI[24] and has adopted an AI "Action Plan" that expressly directs the Department of Commerce to remove the term climate change from a new standard emerging from the National Institute of Standards and Technology, a public-private standard-setting organization. [25] Some states have adopted regulations regarding AI, but none has targeted disclosure of AI-driven energy use or environmental impacts.[26] The European Union has regulatory initiatives underway, but it is unclear whether these initiatives will result in energy and environmental disclosures to corporate and household users in the U.S.[27]

Not surprisingly, given the absence of regulation and the risks of transparency, AI firms disclose only limited energy and environmental information, and most AI users are unaware of the environmental implications of their AI use.[28] Advocates argue that the economic benefits of expanding data centers and AI outweigh the environmental impact, so required disclosure is not necessary or would create greater harms by slowing down the development of beneficial AI applications.[29] In other

---

[23] *See* S.Amdt.2360 to H.R.1, 119th Cong. § 40012 (2025); S.Amdt.2814 to S.Amdt.2360 to H.R.1, 119th Cong. (as passed by Senate, Jul. 1, 2025); Matt Brown & Matt O'Brien, *Senate pulls AI regulatory ban from GOP bill after complaints from states*, PBS NEWS (Jul. 1, 2025), https://www.pbs.org/newshour/politics/senate-pulls-ai-regulatory-ban-from-gop-bill-after-complaints-from-states. For reviews of the barriers to public governance over the last several decades, see Michael P. Vandenbergh, *The Emergence of Private Environmental Governance*, 44 ENVTL. L. REP. 10125, 10132 fig.1 (2014); David Uhlmann, *The Quest for a Sustainable Future*, 1 MICH. J. ENVTL. & ADMIN. L. 1, 9 (2012); Richard J. Lazarus, *Congressional Descent: The Demise of Deliberative Democracy in Environmental Law*, 94 GEO. L.J. 619, 619 (2006).

[24] *See* Removing Barriers to American Leadership in Artificial Intelligence, Exec. Order No. 14,179, 3 C.F.R. § 8741 (Jan. 31, 2025); Accelerating Federal Permitting of Data Center Infrastructure, Exec. Order No. 14,318, 3 C.F.R. § 35385 (Jul. 28. 2025); Promoting the Export of the American AI Technology Stack, Exec. Order No. 14,320, 3 C.F.R. § 35393 (Jul. 28, 2025).

[25] WHITE HOUSE, WINNING THE RACE: AMERICA'S AI ACTION PLAN 4 (Jul. 2025), https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf.

[26] *See, e.g.,* Austin Jenkins, *Tech policies are becoming public safety issues: State lawmakers say they must approach artificial intelligence and data privacy from a more dire angle*, PLURBUS NEWS (Aug. 13, 2025), https://pluribusnews.com/news-and-events/tech-policies-are-becoming-public-safety-issues/; BCLP LAW, *US State-by-State AI Legislation Snapshot*, https://www.bclplaw.com/en-US/events-insights-news/us-state-by-state-artificial-intelligence-legislation-snapshot.html (last visited Aug. 13, 2025).

[27] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on harmonised rules on artificial intelligence and amending Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828, 2024 O.J. (L 1689) 1 (EU), https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

[28] See Jane E. Miller & Michael P. Vandenbergh, Survey: Knowledge About AI Energy Demand and Environmental Impacts, University Law School Working Paper (August 2025).

[29] CBIZ INSIGHTS, *Do the Benefits of Generative AI Outweigh its Environmental Impact?* (Jul. 25, 2024), https://www.cbiz.com/insights/article/do-the-benefits-of-generative-ai-outweigh-its-environmental-impact; *see also* Ana Poscic, *The Intersection between Artificial Intelligence and Sustainability: Challenges and Opportunities*, 8 ECLIC 748, 757 (2024), https://hrcak.srce.hr/ojs/index.php/eclic/article/view/32300 ("Al can improve societal outcome like no poverty, quality education, clean water and sanitation, clean energy, sustainable cities, identify areas of poverty by satellite images").

words, unleashing AI will be such a benefit for combating environmental threats in the long term that its direct energy demand and environmental impacts should be overlooked.[30] It is accurate that AI tools have helped companies improve climate reporting and disclosure,[31] optimize energy generation to increase the use of renewables,[32] and reduce their environmental footprint,[33] and additional benefits are likely, but it is not at all clear that greater disclosure would hamper or delay these outcomes.

The lack of disclosure from major AI providers about energy and water consumption makes it very difficult for experts, policymakers, and the public to understand the effects of AI-driven data center electricity use.[34] Researchers can run experiments using open-source models,[35] but most popular models are private, and their owners do not report on their environmental footprint.[36] Without data from the source, AI users are left in the dark.

---

[30] James Temple, *Sorry, AI won't "fix" climate change,* MIT TECH. REV. (Sept. 28, 2024), https://www.technologyreview.com/2024/09/28/1104588/sorry-ai-wont-fix-climate-change; Chase DiBenedetto, *Google's former CEO: AI advances more important than climate conservation,* MASHABLE (Oct. 7, 2024), https://mashable.com/article/former-google-ceo-invest-ai-despite-climate-concerns; *see generally* Vasudha Sharma & Abheyshek Jamwal, *Artificial Intelligence's Role in Environmental Conservation: A Study on Harnessing Artificial Intelligence for Planetary Preservation*, 7 INT'L J.L. MGMT. & HUMAN. 297 (2024), https://ijlmh.com/paper/artificial-intelligences-role-in-environmental-conservation-a-study-on-harnessing-artificial-intelligence-for-planetary-preservation/.

[31] CDP WORLDWIDE, *CO2 AI Product Ecosystem*, https://www.cdp.net/en/insights/co2ai-product-ecosystem (last visited Aug. 11, 2025); PERSEFONI AI, *The future of climate management is here*, https://www.persefoni.com/persefoniai (last visited Aug. 11, 2025); CLARITY AI, *Make Climate Risk Understandable, Measurable, and Actionable,* https://clarity.ai/climate/ (last visited Aug. 11, 2025). For a discussion of private governance initiatives directed at environmental problems, *see* Michael P. Vandenbergh, *Private Environmental Governance*, 99 CORNELL L. REV. 129, 134-37 (2013).

[32] DEP'T OF ENERGY, *Artificial Intelligence for Energy*, https://www.energy.gov/topics/artificial-intelligence-energy (last visited Aug. 11, 2025).

[33] AMAZON, *How AI is helping Amazon buildings conserve water and improve energy efficiency around the world* (Mar. 11, 2025), https://www.aboutamazon.com/news/sustainability/amazon-ai-buildings-water-energy-efficiency; *see* Claudia Elena Marinica, *Artificial Intelligence - A Possible Key for Better Results on Tackling Climate Change*, 12 UNION JURISTS ROMANIA L. REV. 83, 90 (January-June 2022), https://internationallawreview.universuljuridic.ro/index.php/journal/article/view/7 (discussing the Capgemini Research Institute report "Climate AI: How artificial intelligence can power your climate action strategy" which finds that AI has the potential to help organizations decrease their energy intensity to 11-45% of targets of Paris Agreement).

[34] O'Donnell & Crownhart, *supra* note 5; *see* Stein, *supra* note 5, at 920 ("[t]he hope is that increasing transparency and accountability would make researchers put more effort into keeping these costs low and brining awareness to potential impacts of algorithms.").

[35] Sasha Luccioni, Bruna Trevelin, & Margaret Mitchell, *The Environmental Impacts of AI – Policy Primer*, HUGGING FACE BLOG (2024), https://www.sashaluccioni.com/AI%20+%20Environment%20Primer%20(Hugging%20Face).pdf.

[36] Companies like Google and Microsoft report their total environmental footprints in annual reports but do not specify what portion is attributable to AI. It can be especially challenging because most data centers perform handle more than just AI.

Organizations have developed AI environmental footprint calculators to fill some gaps,[37] but they rely on estimates and do not have a consistent methodology. They represent an important step in informational governance of AI, but the results can be inconsistent. For instance, the Ecologits calculator estimates that asking Meta's 8 billion parameter Llama 3.1 model to write a 50-token Tweet uses 0.189Wh of 1.51 0.026 energy.[38] (A token is a numerical representation of four or five characters that allows the model to process data more efficiently. For example, "resting" might be split into [rest] and [ing], corresponding to [123] and [456], while the word "waking" might be split into [wak] and [ing], corresponding to [789] and [456].[39]) On the other hand, ChatUI Energy estimates that the same prompt on the same model uses just 0.003Wh.[40] A person who uses one of these footprint calculators will thus come to very different conclusions about the energy costs of AI use than someone who chooses the other. Neither calculator provides the information necessary to determine which estimate is more accurate without disclosures from the operators themselves.

These difficulties are exacerbated by the latest generation of LLMs. Open AI's GPT-5 incorporates multiple different models with different energy use and environmental impact and automatically chooses which model to use for processing a given query, a process described as "automatic routing." As a result, users have neither control over nor knowledge about which model their query uses and what the energy and environmental impact will be.[41] This practice could be beneficial if routing decisions prioritize resource efficiency, but the algorithm's opacity clouds users' awareness of their environmental impact.

---

[37] *See e.g.*, HUGGING FACE BLOG, *EcoLogits Calculator*, https://huggingface.co/spaces/genai-impact/ecologits-calculator (last visited Aug. 11, 2025); Julien Delavande, *Chat UI Energy: Tracking Energy Use in ChatUI*, HUGGING FACE SPACES https://jdelavande-chat-ui-energy.hf.space/ (last visited Aug. 12, 2025); DELOITTE, *AI Carbon Footprint Calculator*, https://www.deloitte.com/uk/en/services/consulting/content/ai-carbon-footprint-calculator.html (last visited Aug. 13, 2025); AIX HUMAN, *AI Environmental Footprint*, https://www.aixhuman.co/ (last visited Aug. 13, 2025).

[38] HUGGING FACE BLOG, *EcoLogits Calculator*, https://huggingface.co/spaces/genai-impact/ecologits-calculator (last visited Aug. 11, 2025).

[39] OPENAI, *What are tokens and how to count them?*, https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them (last updated Jul. 30, 2025).

[40] Delavande, *supra* note 37.

[41] Benji Edwards, *The GPT-5 rollout has been a big mess,* ARS TECHNICA (Aug. 11, 2025), https://arstechnica.com/information-technology/2025/08/the-gpt-5-rollout-has-been-a-big-mess/; *see also* OPEN AI, *Using GPT-5*, https://platform.openai.com/docs/guides/latest-model (last visited Aug. 15, 2025).

This Article begins by examining the ongoing debate about the energy demand of AI inference. It then examines what we know about AI inference's energy consumption, how the landscape is changing, and how AI energy demand might evolve. It also compares four popular footprint calculators and demonstrates that different assumptions in these footprint calculators result in substantial differences in estimated consumption. The Article concludes with a call for viable public or private governance initiatives that increase transparency and drive progress on energy efficiency and environmental impacts without discouraging or delaying the benefits of AI.

II. Does Inference Matter?

AI deployment has two main phases: training and inference. Training is the learning phase, where the AI model is fed massive amounts of information. Using complex algorithms, the model processes this data to identify patterns and learn connections, much like teaching someone to recognize different animals by showing them countless pictures. Inference is the application phase. Here, the trained model uses the patterns it learned from the processed information to analyze new, unseen data and make predictions, generate outputs, or perform other tasks. This is what happens when you ask an AI tool a question or ask it to perform a task; it looks through its memory for similar data and predicts an output based on those patterns.[42]

Early in the public deployment of AI, the resource consumption of training took center stage due to its high up-front costs.[43] Some estimated that ChatGPT 3.0, the model that gained instant notoriety in late-2022, used as much electricity in training as 1,000 households do in a year.[44] Aided by research into efficient training design and more efficient, AI-focused chips,

---

[42] New developments of reasoning models include additional steps where the model breaks problems down into smaller steps and "checks its work." *See* Vinayakh, *How Reasoning Models are transforming Logical AI thinking*, MICROSOFT DEVELOPER COMMUNITY BLOG (Feb. 4, 2025), https://techcommunity.microsoft.com/blog/azuredevcommunityblog/how-reasoning-models-are-transforming-logical-ai-thinking/4373194.

[43] *See* Alesia Zhuk, *Artificial Intelligence Impact on the Environment: Hidden Ecological Costs and Ethical-Legal Issues*, 1 J. DIGIT. TECHS. & L. 932, 934 (2023), https://www.lawjournal.digital/jour/article/view/303?locale=en_US (*citing* Emma Strubell, Ananya Ganesh & Andrew McCallum, *Energy and Policy Considerations for Deep Learning in NLP*, ARXIV (June 2019), https://doi.org/10.48550/arXiv.1906.02243 (finding in 2019 that training a single AI model can emit as much carbon dioxide as the lifetime emissions of five cars)).

[44] Sarah McQuate, *Q&A: UW researcher discusses just how much energy ChatGPT uses*, UW NEWS (Jul. 27, 2023), https://www.washington.edu/news/2023/07/27/how-much-energy-does-chatgpt-use/.

AI developers have been relatively successful at reducing training's energy and environmental burden.[45] The environmental costs of training are hard to abate, but it can be reasonably assumed that the energy intensity of training will continue to decrease.[46]

On the other hand, inference's environmental footprint may grow for three main reasons. First, more people will use AI intentionally (e.g., talking to ChatGPT) and unintentionally (e.g., AI tools embedded in email, video conferencing, and other services).[47] Second, the proliferation of reasoning models will increase the complexity and energy demand of LLMs.[48] Third, the deployment of agents and other AI-empowered systems that continually operate will increase the volume and regularity of inference.[49]

Access to LLM inference is currently priced below the cost of delivering the service, however, and providers are choosing to lose billions of dollars in order to build a customer base.[50] At some point AI providers will need to generate profits, and a steep increase in the cost of inference to consumers may substantially reduce demand for inference. Some industry analysts worry that consumers' demand curves are incompatible with the

---

[45] For instance, Chinese startup DeepSeek made headlines when they announced they had trained their V3 model using just 2.8 million GPU hours, far less than their competitors used to train similarly sized models. GPU hours are a good analogue for energy consumption since each GPU draws additional power each second it is being used in training or inference. *See* DEEPSEEK-AI, *DeepSeek-V3 Technical Report* (Dec. 27, 2024), https://arxiv.org/html/2412.19437v1. Meanwhile, Nvidia's new chips are up tot 25x more efficient when handling AI workloads than their predecessors. *See* Catalyst w/ Shayle Kann, *Can chip efficiency slow AI's energy demand?*, LATITUDE MEDIA (Jul. 18, 2024), https://www.latitudemedia.com/news/catalyst-can-chip-efficiency-slow-ais-energy-demand/.

[46] There is a clear trend of decreasing energy intensity of computing operations (measured as the energy consumed to perform a fixed number of computations, such as Wh/FLOP), but the increasing complexity of models and diminishing returns in training successive generations could very well lead to growing energy and environmental costs of training, despite growing efficiency in the computing hardware . *See, e.g.,* Jared Fernandsez, Luca Wehrstedt, Leonid Shamis, Mostafa Elhoushi, Kalyan Saladi, Yonatan Bisk, Emma Strubell, & Jacob Kahn, *Efficient hardware scaling and diminishing returns in large-scale training of language models*, TRANS. MACH. LEARNING RES. (2025), https://openreview.net/forum?id=p7jQEf3wlh.

[47] "Batching" – sending many queries at once – is a common strategy for data centers to improve efficiency. It is possible that increasing the volume of requests will lead to relatively more efficient batching, but the gross increase will almost certainly increase overall energy demand.

[48] Maximilian Dauner & Gudrun Socher, *Energy Costs of Communicating with AI*, FRONT. COMMUN. (Jun. 19, 2025), https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2025.1572947/full.

[49] *See* Patrick K. Lin, *The Cost of Training a Machine: Lighting the Way for a Climate-Aware Policy Framework That Addresses Artificial Intelligence's Carbon Footprint Problem*, 34 FORDHAM ENV'T L. REV. 1, 18 (Spring 2023), https://ir.lawnet.fordham.edu/elr/vol34/iss2/1/ ("In contrast, even though training typically involves repetition, inference is performed every single day, nonstop, for as long as the AI system or device is in use.").

[50] Ashley Capoot, *OpenAI's Altman is still looking to spend after GPT-5 launch and is 'willing to run the loss'*, CNBC NEWS (Aug. 8, 2025), https://www.cnbc.com/2025/08/08/chatgpt-gpt-5-openai-altman-loss.html.

revenue necessary to keep these companies solvent.[51] Thus, considerable uncertainty exists about demand for AI inference after the price to consumers reflects the actual cost of providing the services.

New developments in technology and AI software can be expected to increase the efficiency of AI models and reduce energy demand, but these developments will be competing with rapid expansion of AI. The net effect of model training and inference on these countervailing trends is unclear, but the result is pressure on the electric grid and the environment in the near term and risks to economic and environmental goals in the long run. Three main camps have emerged in the debate about the environmental footprint of AI inference: inference is environmentally irrelevant, the value of inference outweighs it environmental costs, and inference poses substantial environmental risks.

Inference is Environmentally Irrelevant

Individuals who hold the opinion that inference does not raise important environmental concerns point to estimates like the widely cited 3Wh per ChatGPT query or OpenAI CEO Sam Altman's more recent assertion that an average ChatGPT interaction uses 0.34Wh.[52] Meanwhile, actions that typically do not garner much attention for their environmental footprint like using a microwave or TV use as much or more energy (20Wh and 3.3Wh per minute of use, respectively). Critics of the view that inference matters for energy use rely on this type of comparison to argue that AI inference is a drop in the bucket and attention should be focused elsewhere.[53] As some models migrate from centralized data centers to individuals' devices, this comparison may grow more salient because AI energy demand will be embodied in charging one's computer or phone – something most people do every day already.

---

[51] *See , e.g.,* Edward Zitron, AI is a money trap. WHERE'S YOUR ED AT BLOG (Aug. 6, 2025), https://www.wheresyoured.at/ai-is-a-money-trap/.

[52] Sasha Luccioni, Boris Gamazaychikov, Theo Alves de Costa, & Emma Strubell, *Misinformation by Omission: The Need for More Environmental Transparency in AI*, ARXIS (Jun. 18, 2025), https://arxiv.org/pdf/2506.15572; Sam Altman, *The Gentle Singularity*, SAM ALTMAN BLOG (Jun. 10, 2025), https://blog.samaltman.com/the-gentle-singularity.

[53] Andy Masley, *Using ChatGPT is not bad for the environment*, THE WEIRD TURN PRO: AI & THE ENVIRONMENT (Jan. 13, 2025), https://andymasley.substack.com/p/individual-ai-use-is-not-bad-for; Altman, *Id*; Marcel Salathé, *Does ChatGPT use 10x more energy than a standard Google search?*, ENGINEERING PROMPTS (Jan. 16, 2025), https://engineeringprompts.substack.com/p/does-chatgpt-use-10x-more-energy.

Critics of this perspective, on the other hand, argue that energy estimates are unsubstantiated, and that the scale of AI adoption will make even small contributions accumulate into a major environmental challenge.[54] Both the 3Wh and 0.34Wh estimates come from questionable sources and, without disclosures from AI operators, it is impossible to know exactly how much energy each inference uses. Furthermore, it is hard to explain the projections that AI will cause a major spike in U.S. energy demand in the coming years without also accepting that implication that inference will consume substantially more energy in the future since inference is responsible for about 90% of the lifecycle energy of an AI model.[55] Models that can run locally, meanwhile, do not necessarily consume less energy; in fact, it is possible that moving away from data centers that are ruthlessly optimized for efficiency will increase energy consumption as it spreads across millions of individual users.[56]

The Value of Inference Outweighs its Environmental Costs

Another perspective assumes that disclosure and regulation should not occur because the utility of the technology outweighs even a significant environmental footprint.[57] Inference is responsible for up to 90% of deployed AI tools' energy footprints, but little is known about its marginal costs.[58] AI leaders like Google's Eric Schmidt and OpenAI's Sam Altman

---

[54] Eshta Bhardwaj, Rohan Alexander, & Christoph Becker, *Limits to AI Growth: The Ecological and Social Consequences of Scaling*, ARXIV (Jan. 29, 2025), https://arxiv.org/html/2501.17980v1; Christian Bogmans, Patricia Gomez-Gonzalez, Ganchimeg Ganpurev, Giovanni Melina, Andrea Pescatori, & Sneha D Thube, *Power Hungry: How AI Will Drive Energy Demand*, INT'L MONETARY FUND: WORKING PAPERS (Aug. 21, 2025), https://www.imf.org/en/Publications/WP/Issues/2025/04/21/Power-Hungry-How-AI-Will-Drive-Energy-Demand-566304. Kevin M. Stack & Michael P. Vandenbergh, *The One Percent Problem*, COLUMBIA L. REV. 111, 1388 (2011), https://scholarship.law.vanderbilt.edu/faculty-publications/226/ ("Our concern is… when the one percent argument is made in circumstances where small contributors account for so much of a regulatory problem that the social goal cannot be met without regulating many one percent sources."). Michael P. Vandenbergh, *The Individual as Polluter*, Env. L. Rep. (Nov. 2005), https://ssrn.com/abstract=847804.

[55] SHEHABI ET AL., *supra* note 8; INT'L ENERGY AGENCY, *Energy Demand from AI*, https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai (last visited Aug. 11, 2025); Radosvet Desislavov, Fernando Martínez-Plumed, & José Hernández-Orallo, *Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws of deep learning*, 38 SUSTAINABLE COMPUTING: INFORMATICS AND SYSTEMS 100857 (Apr. 2023), https://www.sciencedirect.com/science/article/pii/S2210537923000124#b2.

[56] Eugene Gorelik, *Cloud Computing Models* (Jan. 2013) (Master's thesis, Master of Mgmt. & Master of Eng'g, Massachusetts Institute of Technology) https://dspace.mit.edu/handle/1721.1/79811.

[57] This may overlap with the first group since those who believe inference's footprint is minimal are also likely to believe that the benefits of AI are greater than those costs; *see* Newal Chaudhary, *AI and the Fight against Climate Change: Opportunities and Challenges for Environmental Law*, 6 INT'L J.L. MGMT. & HUMAN. 390, 391-94 (2023), https://ijlmh.com/paper/ai-and-the-fight-against-climate-change-opportunities-and-challenges-for-environmental-law/ (describing the different ways in which AI may be able to help solve climate change and environmental degradation).

[58] Desislavov, *supra* note 55.

have suggested that unleashing AI will enable us to solve major societal problems including climate change and environmental degradation.[59] This may be true, but it places all of society's chips on a single high-stakes gamble that AI will invent ways to rapidly and inexpensively remove greenhouse gases from the atmosphere and alleviate other environmental harm, such as groundwater depletion. It also contradicts the many estimates that AI will drive a long-term surge in energy demand, much of which will be met with new natural gas-fueled plants.[60]

Those with less optimistic projections may point to AI's ability to improve efficiency or accelerate research into greener technologies as a net gain for the environment. For example, if AI can meaningfully decrease the amount of time programmers have their machines on, are working in air-conditioned offices, are brewing coffee to meet a crunch, etc., it will result in a net decrease in resource consumption even if the AI tools are resource intensive. Moreover, if AI makes previously impossible research achievable, it may grant new insights that help us combat environmental problems. Again, if this is true, why are projections suggesting that major increases will occur in net electricity demand, causing utilities to make major investments in new gas-fired electricity generation and increasing stress on corporate climate commitments?[61]

Both positions maintain that the output of AI is more valuable than the resources that go into powering it. They are likely to promote expanding AI since they see the potential gains as greater than the costs. It remains important to consider that the way in which AI services are powered can change the analysis (e.g., coal-powered data centers will have different social consequences than solar with storage or nuclear power).[62] Within this view, reducing or shifting the timing of energy demand can increase the margin between AI's positive and adverse effects. The argument that AI will produce net social and environmental benefits is not in tension with

---

[59] DiBenedetto, *supra* note 30; Temple, *supra* note 30; Lin, *supra* note 49, at 5 ("A 2018 survey conducted by Intel and the research firm Concentrix found that 74 percent of business-decision makers working in environmental sustainability agree that Al will help solve long-standing environmental challenges.").
[60] SHEHABI ET AL., *supra* note 8, 5; INT'L ENERGY AGENCY, *supra* note 55.
[61] It is true that projections tend to overestimate future energy demand. See discussion infra at notes 82-90.
[62] Eamon Farhat, *AI data center growth means more coal and gas plants, IEA says*, BLOOMBERG (April 9, 2025), https://www.bloomberg.com/news/articles/2025-04-10/ai-data-center-growth-means-more-coal-and-gas-plants-iea-says; Mason Adams, *Georgia Power asks to keep coal plants burning to meet AI demand*, SOUTHEAST ENERGY NEWS (Feb. 3, 2025), https://www.canarymedia.com/newsletters/georgia-power-asks-to-keep-coal-plants-burning-to-meet-ai-demand.

enhanced disclosure or energy-saving initiatives, so long as disclosure does not delay or substantially increase the costs of AI.

In short, some AI advocates argue that the alternatives to AI are relatively more environmentally harmful than AI, that disclosing and reducing emissions will slow the development of AI, and that slowing its development is harmful on net. They also contend that rapid hardware and software upgrades will minimize future environmental costs.[63]

Inference Poses Substantial Environmental Risks[64]

Finally, some argue that AI inference is a serious threat to environmental goals. Increased energy demand from the data centers serving AI have led to major increases in projected energy demand, and meeting this demand may conflict with carbon emissions goals.[65] Some assert that the aggregate effect of inference is large even if any individual use is small.[66] Since AI needs data centers and data centers need firm energy, fossil fuels make up a larger share of data centers' electricity supply than the rest of the grid ("firm energy" refers to sources that generate a consistent amount of energy over an indefinite period of time).[67] AI infrastructure in the US alone uses as much energy and emits as much carbon as many countries.[68]

These challenges are likely to grow in the coming years. Projections that data center energy demand could triple by 2028 suggest that emissions

---

[63] Proponents may argue back that the Jevons Paradox suggests otherwise. *See* Greg Rosalsky, *Why the AI world is suddenly obsessed with a 160-year-old economics paradox*, PLANET MONEY NEWSLETTER (Feb. 4, 2025), https://www.npr.org/sections/planet-money/2025/02/04/g-s1-46018/ai-deepseek-economics-jevons-paradox.

[64] Many people who hold this position would also point to water and critical mineral consumption. These are essential parts of the discussion, but this piece focuses only on the energy demand of AI.

[65] Vinayakh, *supra* note 42; GOOGLE LLC, *Google Environmental Report 2025*, at 105 (June 2025), https://www.gstatic.com/gumdrop/sustainability/google-2025-environmental-report.pdf (showing that Google's emissions have increased nearly 50% since 2021, challenging their net-zero target); MICROSOFT CORP., *2025 Environmental Sustainability Report: Accelerating progress to 2030*, at 12 (May 29, 2025), https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/2025-Microsoft-Environmental-Sustainability-Report.pdf (finding that Microsoft's emissions have increased 23.4% relative to base year).

[66] Nidhal Jegham, Marwan Abdelatti, Lassad Elmoubarki, & Abdeltawab Hendawi, *How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference*, ARXIV (May 15, 2025), https://arxiv.org/html/2505.09598v2.

[67] Guidi et al., *supra* note 14.

[68] Guidi et al., *supra* note 14 (suggesting US data centers "produced 105 million tons CO_2e" in 2024, more than Peru, New Zealand, or Greece); Monica Crippa et al., *GHG Emissions of All World Countries – 2024 Report*, EUR 31802 EN, JRC 138862 (Publications Off. of the Eur. Union 2024), https://edgar.jrc.ec.europa.eu/report_2024.

from electricity supply would almost certainly increase.[69] The projected load growth is also being used to justify adding new fossil fuel infrastructure and bringing older facilities back online.[70] More use also means that chips – which have significant embodied emissions – need to be replaced more frequently.[71]

One way to bypass this debate is to induce transparent disclosure from AI providers without adopting regulations that are likely to delay the development of the technology and its environmental benefits. It is possible that AI will follow its predecessors by falling well short of expected energy demand; [72] the problem is that investing in fossil fuel infrastructure locks countries into decades of reliance on those fuels even if AI energy demand never materializes. Missing from the debate are options that facilitate disclosure in ways that will not delay the development of the technology but will reduce the risk of locking into capital investments that will become outdated and inconsistent with environmental goals in future years.

III. How Large is AI Energy Demand?

In the absence of consistent reporting from AI developers and data center operators, researchers have conducted independent studies estimating the energy demand of AI inference. This section highlights estimates of AI energy consumption in the literature and on public-facing footprint calculators. We find that estimates tend to follow certain patterns (e.g. model size is correlated to energy consumption) but also exhibit significant variation in their estimates of energy consumption.

The lone lifecycle assessment of a major AI model's inference comes from French startup Mistral AI whose Large 2 model emits 1.14g CO_2 equivalent for a 400-token response (roughly one page of generated text).[73]

---

[69] SHEHABI ET AL., *supra* note 8; *see* Stein, *supra* note 5, at 917 (AI data centers are projected to consume between 8% (best case) and 21% (expected) of global electricity in 2025).

[70] Matteo Wong, *The False AI Energy Crisis*, THE ATLANTIC: TECHNOLOGY (Feb. 11, 2025), https://www.theatlantic.com/technology/archive/2025/02/ai-energy-crisis-fossil-fuels/681653/.

[71] It is possible that replacing chips more regularly can result in less energy consumption, but the net effect is entirely dependent on the relative efficiency of the original and replacement chips.

[72] Juan Pablo Carvallo, Peter H. Larsen, Alan H. Sanstad, & Charles A. Goldman, *Long term load forecasting accuracy in electric utility integrated resource planning*, 119 ENERGY POLICY 410 (Aug. 2018), https://doi.org/10.1016/j.enpol.2018.04.060; Michael Wara, Danny Cullenward, & Rachel Teitelbaum, *Peak Electricity and the Clean Power Plan*, 28 ELECTRICITY J. (May 2015), https://doi.org/10.1016/j.tej.2015.04.006.

[73] MISTRAL AI, *Our contribution to a global environmental standard for AI* (Jul. 22, 2025), https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai.

As we noted above, one token usually corresponds to four or five characters and represents a single unit of data that the model takes in or produces.[74] Although Mistral AI does not disclose energy consumption (which is the metric used by many studies to control for variation in energy intensity of electricity supply), we can infer based on the average emissions intensity of the US grid (0.367g CO_2e per kWh) that the Mistral AI Large 2 model uses about 3.11Wh for 400-tokens.[75] This may serve as a valuable benchmark, but it is difficult to extrapolate it to other models due to differences in model architecture, size, data center hardware, and other factors. This estimate does fall within the range suggested by the literature, though.

The Literature

Estimates of the environmental footprint at all stages of AI development and deployment suffer from substantial uncertainty.[76] In lieu of first-hand disclosures of inference's environmental footprint, researchers are forced to make a series of difficult choices with major implications for their findings.[77] In addition, no standardized way to report findings has emerged; some researchers report that they index findings "by query" while others use specific token counts or GPU hours. The range of variables that meaningfully affect estimates makes it nearly impossible to generate direct comparisons. Given the underlying challenges, rather than

---

[74] OPENAI, *supra* note 39.

[75] U.S. ENERGY INFORMATION ADMIN., *How much carbon dioxide is produced per kilowatthour of U.S. electricity generation?*, https://www.eia.gov/tools/faqs/faq.php?id=74&t=11 (last updated Dec. 11, 2024) (showing the emissions intensity of US grid). There is reason to believe they used the US grid's average since they benchmark it against time streaming video in the US. If they were to use the French emissions intensity, Large 2 would consume 23.75Wh per 400 tokens because the French grid only emits 0.048g CO_2e per kWh. *See* EUROPEAN ENV'T AGENCY, *Greenhouse gas emission intensity of electricity generation in Europe* (Jun. 27, 2025), https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1.

[76] *See* Ian Schneider, Hui Xu, Stephan Benecke, David Patterson, Keguo Huang, Parthasarathy Ranganathan, & Cooper Elsworth, *Life-Cycle Emissions of AI Hardware: A Cradle-To-Grave Approach and Generational Trends*, ARXIV (Feb. 2025), https://arxiv.org/pdf/2502.01671 (models of how to assess the lifecycle impacts of AI); Lynn H. Kaack, Priya L. Donti, Emma Strubell, George Kamiya, Felix Creutzig, & David Rolnick, *Aligning artificial intelligence with climate change mitigation*, 12 NATURE CLIMATE CHANGE 518 (2022), https://www.nature.com/articles/s41558-022-01377-7#Sec5. *See also* O'Donnell & Crownhart, *supra* note 5 (explaining how estimating AI's environmental footprint is nearly impossible without more information). Most research on AI inference measures compute costs from running the programs but does not account for cooling and other data center operations that can make up as much as 50% of energy demand associated with a given inference. This paper will use the doubling approach to account for non-compute energy consumption and refer to these figures as "secondary costs."

[77] For example, whether a researcher assumes that their inference is being powered by a Nvidia A100, H100, or B200 GPU can make the resulting energy estimate vary by more than an order of magnitude. *See* NVIDIA CORP., *What Is MLPerf?*, https://www.nvidia.com/en-au/data-center/resources/mlperf-benchmarks/ (last visited Aug. 11, 2025). The problem is, unless they are running the model on their own hardware (which has its own tradeoffs in terms of accuracy) they cannot possibly know what kind of chip(s) processed their query. Variation then compound over each assumption.

identifying the most accurate assessment, this section assesses the research landscape showing that academics and the public are largely unable to determine the true costs of AI inference.[78]

The limitations of attempting to develop a per-query carbon footprint are exacerbated by the move by OpenAI automatic routing that depends upon private, proprietary criteria. This routing algorithm does not allow the user to choose the inference model, nor to find out which model will be used. Thus, the uncertainty in calculating emissions per query comprises both the variation within each inference model from one query to another and also the variation in which model a query will be routed to.[79]

The most commonly cited estimates of AI energy demand are 3Wh and 2.9Wh per query, often referenced together to provide corroboration.[80] The problem is that these numbers are completely contrived.[81] The first estimate dates to a 2009 blog post from Google[82] and a 2023 interview with Alphabet Chairman John Hennesy, who has no affiliation with OpenAI.[83] A 2023 SemiAnalysis report seeking to estimate the financial cost of integrating ChatGPT-like functionality into every Google search spawned the second figure when researcher Alex de Vries extrapolated the report's findings to estimate energy consumption.[84] De Vries himself later cast doubt on the accuracy of his calculation saying it "felt like 'grasping at straws'… because he had to rely on third-party estimates he could not replicate."[85]

---

[78] To control for at least one of these variables, this section pulls research conducted since November 2022 to avoid irrelevant findings based on outdated model constructions. While any discrete cutoff is somewhat arbitrary, the release of GPT-3 in November 2022 marked a significant change in the AI landscape.

[79] Delavande, *supra* note 37.

[80] Luccioni, *supra* note 52 (finding that 53% of news reports on "ChatGPT energy consumption" cite the 3Wh figure).

[81] *Id*., at 4-5 (lays out several key uncertainties in the estimate).

[82] Urs Hölzle, *Powering a Google search*, GOOGLE: OFFICIAL BLOG (Jan. 11, 2009), https://googleblog.blogspot.com/2009/01/powering-google-search.html.

[83] Jeffrey Dastin & Stephen Nellis, *Focus: For tech giants, AI like Bing and Bard poses billion-dollar search problem*, REUTERS (Feb. 22, 2023), https://www.reuters.com/technology/tech-giants-ai-like-bing-bard-poses-billion-dollar-search-problem-2023-02-22/.

[84] Dylan Patel & Afzal Ahmad, *The Inference Cost Of Search Disruption – Large Language Model Cost Analysis // $30B Of Google Profit Evaporating Overnight, Performance Improvement With H100 TPUv4 TPUv5*, SEMIANALYSIS (Feb. 9, 2023), https://semianalysis.com/2023/02/09/the-inference-cost-of-search-disruption/; Alex de Vries, *The gorwing energy footprint of artificial intelligence*, 7 JOULE 2191 (Oct. 18, 2023), https://www.sciencedirect.com/science/article/pii/S2542435123003653#bib5. The original SemiAnalysis study never mentions energy or electricity, nor does it claim to analyze the energy demand of ChatGPT.

[85] Sophia Chen , *How much energy will AI really consume? The good, the bad and the unknown*, NATURE (Mar. 5, 2025), https://www.nature.com/articles/d41586-025-00616-z.

Even if these estimates were accurate at the time, they have quickly become outdated due to hardware and software developments. The Nvidia A100 chip, used by SemiAnalysis in the 2023 report, has been succeeded by two new generations of chips (known as Hopper and Blackwell chips) that boast significantly more computational power per watt.[86] What's more, Google's TPUs and Groq's LPUs are optimized for the math and algorithms that power LLMs (both chips are specifically optimized to process the complex linear algebra equations that power LLMs whereas GPUs were originally designed for 3D graphics that require slightly different equations) and claim to be even more efficient than Nvidia GPUs.[87]

More efficient chips will not necessarily decrease energy consumption if software grows more complex, demanding more functions to fulfill a given request. For instance, a team led by Sasha Luccioni in 2024 observed a relationship between the size of a model and energy consumption due to the additional computational intensity of larger models.[88] For the most energy-intensive text-to-text task studied by the Luccioni team (a task called "summarization"), the smallest model consumed 0.114Wh per request whereas the largest consumed 0.416Wh per request.[89] Research by teams led by Siddharth Samsi in 2023 and Erik J. Husom in 2024 corroborated the positive relationship between model size and energy consumption.[90] Further evidence that model size may nullify

---

[86] NVIDIA CORP., *NVIDIA Blackwell Platform Arrives to Power a New Era of Computing* (Mar. 18, 2024), https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing.

[87] Molly McHugh-Johnson, *Aska a Techspert: What's the difference between a CPU, GPU, and TPU?*, THE KEYWORD (Oct. 30, 2024), https://blog.google/technology/ai/difference-cpu-gpu-tpu-trillium/ https://groq.com/blog/what-nvidia-didnt-say.

[88] Alexandra Sasha Luccioni, Yacine Jernite, & Emma Strubell, *Power Hungry Processing: Watts Driving the Cost of AI Deployment?*, ARXIV (Oct. 15, 2024), https://arxiv.org/pdf/2311.16863. To clarify, they draw a relationship between model size and emissions, but because they calculate emissions by multiplying energy consumption by a constant emissions factor, the same relationship would exist for energy. *See also* Siddhartha Samsi et al., *From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference*, ARXIV (Oct. 4, 2023), https://arxiv.org/pdf/2310.03003 (finds that larger LLaMa models consistently produce less words per second and consume more energy per second for multiple tasks, datasets, and types of chips); Erik Johannes Husom, Arda Goknil, Lwin Khin Shar, & Sagar Sen, *The Price of Prompting: Profiling Energy Use in Large Language Models Inference*, ARXIV (Jul. 4, 2024), https://www.arxiv.org/pdf/2407.16893.

[89] Sasha Luccioni, *CO_2 Inference: Complete Emissions Data*, GITHUB (2003), https://github.com/sashavor/co2_inference/tree/main/data/emissions/complete. Accounting for secondary costs. Smallest model is Flan-T5 base (222M parameters) and largest is Flan-T5 xxl (11B parameters). On average, summarization tasks across all models consumed 0.098Wh/request when accounting for secondary costs. Tests were run on the researchers' own A100 hardware.

[90] *See* Samsi et al., *supra* note 88 (finding that larger LLaMa models consistently returned fewer words per second and consumed more energy per second across different tasks, databases, and types of chips);

chip efficiency improvements comes from a 2025 estimate that OpenAI's "largest and best" model, GPT-4.5,[91] consumes 6.72Wh per request even though the authors assumed that GPT-4.5 loads are processed on H100 or H200 chips, which are more efficient than the A100s used in previous research.[92]

Other design choices can complicate the relationship between model size and energy consumption. As the authors in the GPT-4.5 study point out, its high energy consumption "suggests inefficiencies rooted in model architecture."[93] In the first study of OpenAI's open source GPT-OSS models co-created with Hugging Face, researchers found that the 20B parameter model consumed 2.02Wh per 100-token response and the 120B model consumed 8.31Wh per 100 tokens when using A100 chips, somewhat higher but within range of past studies on similarly sized models given the relatively older hardware.[94] The former is particularly interesting because it was designed to run on everyday consumer devices without a dedicated GPU.[95] It is probable that running them locally will increase inference energy consumption because individuals' devices are not optimized in the same way as data centers. Furthermore, if models migrate to operate locally, it could further confuse efforts to trace AI energy consumption since it will be spread across millions of users who are unlikely to track their computer's energy.

Several new developments suggest other factors may be yet more consequential. First, the introduction of reasoning models may substantially increase energy demand.[96] Unlike traditional models that predict the next token based on probability, reasoning models introduce a "thinking" phase in which they break the problem into several steps and analyze its responses along the way. The additional thinking stage included

---

Husom et al., *supra* note 88 (showing that larger CodeLLaMa, Gemma, and LLaMa3 models consumer more energy per response).

[91] OPENAI, *Introducing GPT-4.5.* (Feb. 27, 2025), https://openai.com/index/introducing-gpt-4-5/.

[92] Jegham, *supra* note 66. Importantly, the query length (100 token prompt, 300 token response) is much longer than Luccioni et al. 2024 (54 token prompt, 8.52-10.68 token response). Though it may play a role, it is very unlikely that the difference in prompt and response length accounts for most or all of the discrepancy.

[93] *Id*.

[94] Sasha Luccioni, *The GPT-OSS models are here… and they're energy-efficient!*, HUGGING FACE BLOG (Aug. 7, 2025), https://huggingface.co/blog/sasha/gpt-oss-energy.

[95] OPENAI, *Introducing gpt-oss* (Aug. 5, 2025), https://openai.com/index/introducing-gpt-oss/.

[96] *See* Vinayakh, *supra* note 42 (description of how reasoning models work and how they differ from traditional LLMs).

in these models means the models perform additional computations before arriving at a response. Three of the four most consumptive models studied by a team led by Nidhal Jegham in 2025 are reasoning models, with Deepseek R1 leading the pack at 23.82Wh per response.[97] Although reasoning is a powerful tool for some applications, it is unnecessary for many common functions, so defaulting to these more powerful models may increase AI's energy demand without meaningfully improving the output.

Second, and most important, is evolution in task structure. Text-to-text functions (like asking a question to a chatbot) tend to be orders of magnitude less consumptive than image or video generation. Image generation consumed nearly 60 times as much energy per response than the most intensive text-to-text operation studied by the Luccioni team (5.81Wh and 0.098Wh, respectively).[98] Whereas text-to-text tasks usually have the model generate a string of text based on the probability that one token will follow its predecessor, image and video generation usually involves thousands of inferences over dozens of "inference" or "diffusion steps." These "diffusion models" begin with random noise before gradually refining the image over many iterations (i.e. diffusion steps).[99] More diffusion steps typically lead to a superior output, but they also tend to be more energy intensive since the model is effectively performing additional individual inferences before returning your image.[100] More recently, researchers estimated that Stable Diffusion 3 Medium consumes 1.22Wh per image, significantly less than the Luccioni team's original calculation but still far more consumptive than text-to-text functions on similarly sized models.[101]

Video generation is likely even more intensive, but research about its specific demand is sparse. In 2025, a team led by Julien Delavande

---

[97] Jegham, *supra* note 66. The other reasoning models, OpenAI o3 and o1, consumed 7.03 and 4.45Wh/response, respectively. GPT-4.5 was the third most consumptive despite note being a reasoning model due to its large size and "inefficiencies rooted in model architecture."

[98] Luccioni, *supra* note 88. Figures doubled to account for secondary costs.

[99] *See* Dave Bergman & Cole Stryker, *What are diffusion models?*, IBM (Aug. 21, 2024), https://www.ibm.com/think/topics/diffusion-models (more about diffusion models).

[100] O'Donnell & Crownhart, *supra* note 5. Suggests that "doubling the number [of] diffusion steps to 50 just about doubles the energy required." Output quality probably experiences decreasing marginal returns relative to energy consumption, implying that additional diffusion steps are not always necessary. *See also* Zeyu Yang, Karel Adámek, & Wesley Armour, *Double-Exponential Increases in Inference Energy: The Cost of the Race for Accuracy*, ARXIV (Dec. 12, 2024), https://arxiv.org/html/2412.09731v1.

[101] O'Donnell & Crownhart, *supra* note 5. Cost is to generate a standard-quality 1024x1024 pixel image using 50 diffusion steps. The model has 2B parameters.

studied open-source text-to-video models and found that energy consumption for a single response ranged by almost an order of magnitude, from 0.28-218.94Wh.[102] The researchers explained that factors like model size, the number of diffusion steps, resolution, video length, and architectural design are responsible for the discrepancy.[103] Increasing these factors can improve the video that the model generates by making it longer, higher resolution, and more responsive to the original prompt, but it comes with a high cost. Given the dazzling ability of closed-source models like OpenAI's Sora and Google's Veo 3, it is almost certain that generating videos using these models is more energy intensive than even the highest estimate included in the 2025 estimate by the Delavande research team.

We were unable to find any studies directly comparing the inference costs of traditional LLMs to the inference costs of multi-modal large language models (MLLMs)(an MLLM can perform multiple task types including text, image, and video). It is reasonable to assume that some relationship exists between the number of task structures an MLLM can perform and its energy consumption for a given task because MLLMs typically require more parameters and computational intensity. Future research should focus on comparing the energy demand for multi-modal and unimodal models when performing the same task.

Finally, agentic deployment could drastically increase total energy demand for AI. An agent, as opposed to a traditional or reasoning model, combines dynamic reasoning with the use of external tools to independently perform a specific function.[104] For example, a day trader might create an AI agent to automatically execute transactions based on live market data or a company might automate customer service by giving a

---

[102] Julien Delavande, Sasha Luccioni, & Régis Pierrard, *How Much Power does a SOTA Open Video Model Use?*, HUGGING FACE BLOG (Jul. 2, 2025), https://huggingface.co/blog/jdelavande/text-to-video-energy-cost (figures doubled to account for secondary costs).

[103] *Id*.; *see also* Zhuoyi Yang et al., *CogVideoX: Text-to-Video Diffusion Models with an Expert Transformer*, ARXIV (Mar. 26, 2025), https://arxiv.org/pdf/2408.06072. In Delavande et al., *Id*, CogVideoX 5B used 51.86Wh to produce a 6.13 second video with 50 diffusion steps, 8 FPS, and 480x720 resolution (when accounting for secondary costs). In the technical paper for the model, generating a 5 second, 768x1360 resolution video with 50 diffusion steps consumed up to 97.22Wh (based on running one H800 at full capacity for 500 seconds, accounting for secondary costs). Its creators explain that they ran a single H800 for 500 seconds to generate one five-second 768x1360

[104] Kiin Kim, Byeongjun Shin, Jinha Chung, & Minsoo Rhu, *The Cost of Dynamic Reasoning: Demystifying AI Agents and Test-Time Scaling from an AI Infrastructure Perspective*, ARXIV at 2 (Jun. 4, 2025), https://arxiv.org/pdf/2506.04301 (defining AI agents).

chat bot access to a customer's profile.[105] Potential applications of agents are broad, with some agents expected to run continuously as they perform a given task. Since agents perform so many calculations to complete a task, they may be many times more energy consumptive than non-agentic models. One study found that agent-augmented LLaMa-3.1-Instruct 70B used between 62.1 and 136.5 times as much energy to complete the same task as the base model (2.55Wh per query compared to 158.48 and 348.41Wh per query).[106] The technology is very new, meaning that only limited research has focused on its energy consumption, and efficiency improvements are likely. That said, if agents rapidly overtake traditional models, we may see massive growth in AI's energy load.

The common theme among software developments that drive AI energy intensity is the addition of more steps (reasoning, tool use, diffusion steps, etc.). These steps improve response quality in many instances but also add computational load. Reasoning models outperform traditional models on many tasks and more diffusion steps improve image quality.[107] It is unclear, though, whether these improvements justify the cost for every task, and their ubiquitous deployment may unnecessarily drive up the energy demand and cost of AI inference.[108] For example, writing assistants probably do not need reasoning for spellcheck or grammar suggestions, and generating an AI avatar probably does not need 150 diffusion steps to be acceptable. Greater transparency about energy use may induce developers to critically assess the use cases of their models before defaulting to higher energy designs.

AI Environmental Footprint Calculators

AI footprint calculators offer a valuable approach to disclosure as they are designed to be accessible to a broad audience. By comparing multiple models, footprint calculators can serve as a one-stop resource for

---

[105] *See* Matt Renner & Matt A.V. Chaban, *601 real-world gen AI use cases from the world's leading organizations*, GOOGLE BLOG (Apr. 9, 2025), https://cloud.google.com/transform/101-real-world-generative-ai-use-cases-from-industry-leaders (discussing 601 use cases of AI agents).

[106] Kim, *supra* note 104, at 13.

[107] *See* Zhong-Zhi Li et al., *From System 1 to System 2: A Survey of Reasoning Large Language Models*, ARXIV (Feb. 24, 2025), https://arxiv.org/html/2502.17419v1 (on reasoning); and Yongshi Ye et al., *How Well Do Large Reasoning Models Translate? A Comprehensive Evaluation for Multi-Domain Machine Translation*, ARXIV (May 26, 2025), https://arxiv.org/html/2505.19987v1 (on reasoning); Chen Hou, Guoqiang Wei, & Zhibo Chen, *High-Fidelity Diffusion-based Image Editing*, ARXIV (Jan. 4, 2024), https://arxiv.org/html/2312.15707v3 (on image generation).

[108] Reasoning is unnecessary for many simple tasks and the number of diffusion steps reaches a point where improvements are unnoticeable.

those interested in understanding the energy and environmental costs of AI inference. It is easy to assume that footprint calculators are just for retail users of AI, but they have the potential to educate AI firms, the company, government, and other managers who select AI firms for use by their employees, the employees of those organizations, household users, and policymakers. The accuracy of these footprint calculators is essential, though, as footprint calculators that are flawed or based on incorrect assumptions can lead users to make poor decisions about the best AI firms, models, or queries for their needs. This section examines four footprint calculators to explore the differences in their estimates of AI inference's energy consumption, the implications of these discrepancies, and why they may occur. [109]

Table 1: Characteristics of selected AI footprint calculators.[110]

AEFC	Ecologits[111]	AI Energy Score[112]	ML.ENERGY [113]	ChatUI[114]
Disclosure type	Individual select	Multi compare	Multi compare	Live tracker
Disclosure units	Energy (Wh) to complete a task (e.g. write a 50-token Tweet)	GPU energy (Wh) per 1,000 responses	Energy (J) per response	Total energy (Wh) and duration (sec)
Function type	Wh/Token	Wh/response	Wh/response	Wh/second[115]
Scope	Compute and cooling	GPU energy	GPU energy	GPU energy

[109] This section does not cover LLM Perf Leaderboard which is popular but has limited model coverage. *See* HUGGING FACE BLOG, *LLM-Perf Leaderboard,* https://huggingface.co/spaces/optimum/llm-perf-leaderboard (last visited Aug. 12, 2025).

[110] It is impossible to completely control for response length (i.e. how many tokens make up one "response") and each footprint calculator is slightly different in this regard.

[111] HUGGING FACE BLOG, *supra* note 38.

[112] HUGGING FACE BLOG, *AI Energy Score Leaderboard*, https://huggingface.co/spaces/AIEnergyScore/Leaderboard (last visited Aug. 12, 2025).

[113] ML.ENERGY LEADERBOARD, *How much energy do GenAI models consume? LLM chatbot response generation*, https://ml.energy/leaderboard/ (last updated May 12, 2025).

[114] Delavande, *supra* note 37.

[115] Qwen3 8B uses real energy consumption from NVIDIA NVML but all others are a function of estimated average power and response time. Since total energy is a linear time function, we can convert from seconds to tokens/responses by tokenizing outputs then dividing total energy by number of tokens.

Source database(s)	N/A[116]	WikiText, OSCAR, UltraChat[117]	ShareGPT[118]	N/A
Task Coverage	Text	Multimodal	Multimodal	Text
Hardware	NVIDIA A100 80GB[119]	NVIDIA RTX 4090 (<20B) NVIDIA H100 (>20B)[120]	NVIDIA H100 80GB[121]	NVML API, average power ≈70W

Design choices in the footprint calculators make a direct comparison among the footprint calculators challenging (see Table 1 for an overview of key characteristics of each footprint calculator). This Article opts for a per-token approach to standardize estimates. It also compares similar models since all four footprint calculators only share one model in common (LLaMa 3.1 8B),[122] although they may cover different generations of the same model (e.g., LLaMa 3.1 70B and LLaMa 3.3 70B). Comparisons therefore will not be perfect but will identify major differences in the estimates.

Differences between the footprint calculators are immediately clear from the LLaMa 3.1 8B estimates. The highest estimate was 58 times greater than the lowest, and both leaderboards estimated that the models were at least 30 times less consumptive than Ecologits or ChatUI.[123] When converted to a standard 400-token response, ChatUI Energy is most consistent with studies of similarly sized models from the same generation, but it is still several times higher than estimates from the literature.[124]

---

[116] Energy is benchmarked using LLM-Perf Leaderboard based on model size and number of output tokens.

[117] Total input tokens (T) for text generation = 369,139 for 1,000 data points. It is impossible to know average tokens per response, but, based on the study design, we can assume it is around 369T. *See* HUGGING FACE BLOG, *AI Energy Score: FAQ*, https://huggingface.github.io/AIEnergyScore/#documentation (last visited Aug. 12, 2025).

[118] Average output tokens specified by model when selecting "show more technical details." Range between 375-515 tokens for all models.

[119] https://ecologits.ai/latest/methodology/llm_inference/#on-hardware

[120] HUGGING FACE BLOG, *supra* note 111.

[121] Users can also select A100. Using energy from H100 because it is a newer chip.

[122] ChatUI Energy uses Nous Research's fine tuned version of LLaMa models. *See* Ryan Teknium, Jeffrey Quesnelle, & Chen Guang, *Hermes 3 Technical Report*, ARXIV (Aug. 15, 2024), https://arxiv.org/pdf/2408.11857.

[123] Ecologits = 3.78E-3Wh/T, ChatUI = 1.98E-3Wh/T, AI Energy Score = 9.42E-5Wh/T, ML.ENERGY = 6.5E-5Wh/T. All figures except Ecologits doubled to account for secondary costs.

[124] Estimate = 0.79Wh/request. *See e.g.,* Wong; *supra* note 70; Husom, *supra* note 88.

The relative difference between the highest and lowest estimates for the larger LLaMa 3.1 70B model is smaller, but the gross difference is much larger.[125] A person who generates 1,000 tokens of text each day for a year with LLaMa 3.1 70B could believe that they are using anywhere from 0.14 to 4.6 kWh depending on the footprint calculator they reference. For the even larger 3.1 405B model, the same person could believe that they are using anywhere from 1.5 to 87 kWh per year.[126] When scaled to the grid or interconnection level, these discrepancies can lead decision-makers to vastly different conclusions about what investments might be warranted. Model providers can help those responsible for managing energy supply informed choices by disclosing their marginal energy demand.

*Table 2: Comparison of estimates of selected models in Wh/400T. * Indicates figures have been doubled to account for secondary costs.*

	EcoLogits	AI Energy Score*	ML.ENERGY*	ChatUI*
Phi 3 Mini	1.23	N/A	2.57E-2	N/A[127]
LLaMa 3.1 8B	1.51	3.76E-2	2.60E-2	0.79
Gemma 2/3 27B	2.57	N/A	7.09E-2	0.32
Mistral 8x7Bv0.1[128]	1.73	1.33	9.45E-2	N/A
LLaMa 3.1/3.3 70B	5.02	3.73	0.15	0.32

Comparing ChatUI's estimates for LLaMa 3.1 8B and LLaMa 3.3 70B shows how design choices can influence the estimates generated by a footprint calculator. ChatUI's equation using average power draw and duration make the estimated energy per token for LLaMa 3.3 70B less than 3.1 8B because of 3.3's superior token throughput.[129] In part, this reflects Meta's intention to make 3.3 a more efficient option that can operate with fewer calculations (and therefore produce tokens faster),[130] but the estimate

[125] ML.ENERGY Leaderboard had the lowest again at 3.77E-4Wh/T and Ecologits had the highest at 1.26E-2Wh/T.

[126] 0.239Wh/T for Ecologits, 4.14E-3 for ML.ENERGY Leaderboard. Neither of the other footprint calculators cover 3.1 405B.

[127] Phi 3.5 Mini is listed but when prompted returns an error saying the model does not exist.

[128] Assumes that not all parameters are active. It is possible to change this in "Expert Mode," but the default is presented here.

[129] 8.06E-4Wh/T for 3.3. Both models are assumed to draw 1.53E-2Wh/sec, but 3.3 is able to produce more than twice as many tokens per second.

[130] Ahmad Al-Dahle (@Ahmad_Al_Dahle), X (Dec. 6, 2024, 11:30 AM), https://x.com/Ahmad_Al_Dahle/status/1865071436630778109.

relies on the improbable assumption that both models have the same average power draw, undercutting the reliability of the estimate.

On the other hand, Ecologits' reliance on model size may overlook key factors like computational intensity.[131] EcoLogits estimates GPU energy by averaging the number of active parameters, GPUs needed to service a model based on its active parameters, average output tokens based on model size, and tokens per second (i.e. generation latency) of a variety of models. Then, EcoLogits adds this to an estimate of server energy *without* GPUs and multiplies the figure by a PUE of 1.2. The assumptions that different models of all different sizes are similar enough to accurately predict a trend and that every model with the same number of active parameters will consume the same amount of energy to fulfill a task are likely untrue. That said, these assumptions, accurate or not, allow EcoLogits to predict the energy demand of models that individuals cannot run locally (i.e. proprietary models like ChatGPT), which allows EcoLogits to generate estimates for the most popular models.

AI Energy Score and ML.ENERGY Leaderboard may misrepresent actual usage patterns since they gather estimates from a random assortment of queries. A 2025 study by a team led by Marta Adamska found that the "semantic meaning of the prompt" plays a role in the associated energy consumption.[132] For instance, prompts that include the keyword "analyze" were found to noticeably increase energy consumption relative to prompts that used the keyword "classify" because they prime the model to respond differently. If the dataset does not reflect the ways individuals tend to use AI tools, including the words and their implied meanings, it may cause the estimates to err.[133]

In sum, footprint calculators are important tools, but they suffer from data availability problems and from the inability to validate their estimates by comparing them to actual measurements of emissions or energy consumption. They produce results that differ from each other based on the same inputs and from the conclusions in the academic

[131] For instance, it is unlikely that LLaMa 2 7B and Mistral 7B v0.3 consume the same energy per token but Ecologits assumes as much.
[132] Marta Adamska et al., *Green Prompting*, ARXIV (Mar. 9, 2025), https://arxiv.org/html/2503.10666v1.
[133] ML.ENERGY Leaderboard used real user prompts to try to control for usage patterns, but the dataset likely suffers from selection bias and has since gone out of service, so it is impossible to know how well their dataset predicts real-world patterns. *See* SHAREGPT, *ShareGPT*, www.sharegpt.com (last visited Aug. 12, 2024).

literature because they must make assumptions without a standard method for estimating AI energy use. Ideally, disclosure from model providers would make footprint calculators obsolete, but in the meantime a viable alternative is to develop a standardized methodology that accounts for as many relevant factors as possible. A standardized methodology will help avoid inaccurate or misleading information in these public-facing tools.

IV. As Compared to What?

Proponents of AI proliferation often compare its energy use to its alternatives. If one AI prompt can get a user to the same place as 30 Google searches, it is likely that AI is the more efficient option.[134] Likewise, if AI-assisted coding can produce the same code in a fraction of the time, it may yield net energy savings by reducing the amount of time a programmer's screen is on,[135] the hours spent in an air-conditioned office, and the coffee brewed to complete a sprint.[136] AI inference is relatively efficient by all accounts so it may be able to complete tasks using far less energy than its alternatives.

Applying AI to optimize inefficient systems and accelerate research into low-energy solutions may also yield significant energy savings.[137] Companies like Google and NextEra have developed AI tools to eliminate unnecessary consumption and boost clean energy generation.[138] For

---

[134] *See* Matthew Elmore, *The Hidden Costs of ChatGPT: A Call for Greater Transparency*, 23 AM. J. BIOETHICS 47, 48 (2023), https://doi.org/10.1080/15265161.2023.2250335 (concluding that a single AI prompt has the carbon footprint of 5 Google searches, thus suggesting that AI may be a six times more efficient option).

[135] Computer monitors typically use at least 10W so reducing the screen time of a monitor by just a few minutes may balance out the energy from an AI query.

[136] Enrico Gianfranco Campari & Beatrice Fraboni, *Power and Energy Consumption Reduction in Coffee Brewing* (Mar. 28, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5197204. One at-home experiment comparing energy consumption of coffee machines found that they used between 5.23-26.43Wh per cup.

[137] Nicholas Stern et al, *Green and intelligent: the role of AI in the climate transition*, 4 NPJ CLIMATE ACTION 1 (Jun. 23, 2025), https://www.nature.com/articles/s44168-025-00252-3; *see also* Daniel Slate et al., *Adoption of Artificial Intelligence by Electric Utilities*, 45.1 SSRN JOURNAL 1 (2024), https://www.ssrn.com/abstract=4847872; Stein, *supra* note 5, at 902 ("Google and DeepMind applied machine learning to wind power capacity in the United States 'to better predict wind power output thirty-six hours ahead of actual generation,' assisting in optimal hourly day-ahead delivery commitments.").

[138] Google's AlphEvolve does this by automatically improving complex algorithms and has helped Google recover 0.7% of its global compute. *See* ALPHAEVOLVE, *AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms*, GOOGLE DEEPMIND (May 14, 2025), https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/. Google DeepMind's prediction software helped Google better predict wind conditions and thereby increase generation. *See* Carl Elkin & Sims Witherspoon, *Machine learning can boost the value of wind energy*, GOOGLE DEEPMIND (Feb. 26, 2019), https://deepmind.google/discover/blog/machine-learning-can-boost-the-value-of-wind-energy/. NextEra boasts that its technologies help customers reduce energy costs and improve their energy mix. *See*

instance, researchers at the Paul Sherrer Institute have used AI to rapidly conceive of thousands of cleaner concrete recipes.[139] Using AI, therefore, may not just be about its immediate tradeoffs but also its potential to accelerate progress in other areas. It is possible that seemingly unrelated developments today will reveal new solutions to difficult climate problems in the same way that semiconductors have become an essential part of solar energy despite originally serving a very different purpose.[140]

Although it is true that AI is a relatively efficient operation, especially when considering its complexity, the cumulative energy consumption of billions of inferences is already significant and expected to increase. The marginal cost of a single inference may also increase if model complexity continues to outpace efficiency gains. When considering the electricity mix supplying data centers, energy demand is even more concerning.[141]

The notion that AI will solve our climate woes is also contingent on the priorities of AI developers. Although some organizations are applying AI in ways that benefit the environment, most of the industry appears to be more focused on supercharging power and performance. The predictions of AI boosters like Sam Altman and Eric Schmidt will only come to pass if high level decision makers choose to focus on reducing energy demand and its environmental impacts, an outcome that is in tension with the projections of massive electricity demand growth that will outstrip the supply of electricity from the nuclear and renewable energy sectors.

The problem domains to which AI is applied also have profound implications for the impact of AI tools for addressing climate change. The oil and gas industry is rapidly adopting AI tools to allow it to extract more

---

NEXTERA RESOURCES 360, *Case Studies*, https://www.nexteraenergyresources.com/nextera360/resource-center/case-studies.html (last visited Aug. 12, 2025).

[139] Romana Boiger et al., *Machine learning-accelerated discovery of green cement recipes*, 58 SPRINGER NATURE 172 (Apr. 25, 2025), https://link.springer.com/article/10.1617/s11527-025-02684-z.

[140] *Id*. (explaining how AI has led to insights about protein folding and crystal construction that have major implications for alternative protein sources and renewable energy, respectively).

[141] Guidi et al., *supra* note 14 (showing data centers supplied by electricity that is 48% more carbon intensive); *see* Pham Nhat Linh Chi, Le Thi Minh Hang & Nguyen Viet Vuong, *The Negative Impacts of AI on the Environment and Legal Regulation*, 11 INT'L J.L. 124 (2025), https://www.lawjournals.org/assets/archives/2025/vol11issue1/11016.pdf (finding that the majority of electricity supply data centers currently comes from traditional fossil fuels).

oil and gas more cheaply. In turn, this could delay the transition to clean energy sources.[142]

The implications of AI deployment on the environmental footprint of industry are impossible to predict. It is possible that, like past improvements in industrial productivity, AI deployment will induce companies to build and consume more.[143] There is no 1:1 tradeoff between less efficient human workers and more efficient AI tools because people whose jobs are replaced or improved by AI will continue to emit carbon for some purpose, it may just be attributed differently.[144] It is hard to reconcile projections of surging energy demand attributable to AI and the industry's rapid buildout of energy infrastructure with a vision of the technology as net energy reducing.

Although AI has ample potential to reduce energy consumption and accelerate progress towards low-energy solutions, its actual impact remains uncertain. As AI becomes more widespread, providing information about its energy use and environmental effects need not become a barrier to its rapid development and can help reduce costs and environmental impacts.[145] This greater transparency will not arise from federal government pressure in the near term, but as the footprint calculator comparison above suggests, some AI tools and firms may compare favorably to others on these metrics. That difference in energy use and environmental impacts, if widely known, could provide these firms with economic incentives to disclose electricity use and environmental impacts and to market their competitive advantage to companies and individuals who use their services and have preferences for efficient, low

---

[142] Sheila Dang & Georgina McCartney, *AI leading to faster, cheaper oil production, executives say*, REUTERS (Mar. 13, 2025), https://www.reuters.com/business/energy/ceraweek-ai-leading-faster-cheaper-oil-production-executives-say-2025-03-13/.

[143] Total Factor Productivity by Major Industries, *see* U.S. BUREAU OF LABOR STATISTICS, *Productivity Tables* https://www.bls.gov/productivity/tables/home.htm (last visited Aug. 12, 2024). Since 2017, total factor productivity has increased 6% while total inputs have increased 13%.

[144] It is also possible that an AI tool that is not instructed to prioritize efficiency will embed inefficiencies into its solutions. *See* Md Arman Islam et al., *Evaluating the Energy-Efficiency of the Code Generated by LLMs*, ARXIV (May 23, 2025),https://arxiv.org/html/2505.20324v1 (finding that LLM-generated code is less efficient than human-generated and that there are significant differences in code generated by different LLMs).

[145] For example, many companies that perform sustainability audits find costly inefficiencies. It is possible that AI developers who track and report emissions will find similar algorithmic or hardware inefficiencies. *See* Marinica, *supra* note 33, 92 ("[W]e appreciate that Al can contribute to the fight against climate change and the consolidation of climate forecasts, but at the same time it also involves costs for the planet, which can be offset or even reduced if a better understanding of the carbon footprint is considered while using Al.").

environmental impact AI models. Greater transparency is the first step in harnessing the full potential of AI.

Energy Disclosures to Achieve AI's Potential

The most important question may not be "to AI or not to AI," but rather when and how to use AI tools. For example, programmers can save money, time, and energy by opting for coding-specific tools instead of larger general-purpose models. Likewise, finetuning (i.e. training a model on a specific dataset to improve performance on a given task) and post-training quantization (i.e. reducing precision of model weights) boost efficiency without seriously affecting response quality.[146] Smaller, more tailored models are often more effective than the largest, most powerful option.

Understanding the tradeoffs among tools can help users make the most of AI without unnecessarily increasing energy consumption. Without information, though, AI users lack the ability to make smart decisions for themselves, for their companies or other organizations, or for society. Footprint calculators can be especially valuable in this regard because they enable comparison of various options in a single space. Rather than defaulting to more expensive flagship models, users can compare the costs and qualities of multiple options before selecting the right tool for their task. Doing so requires accurate estimates of energy costs that are not currently available, and model developers can help AI reach its potential by releasing firsthand estimates of relative costs.

Optimizing model decisions should be acceptable to data center operators and model developers, too. Data centers can reduce energy costs while still handling the same number of queries if users select more efficient or more flexible options. They also may be able to avoid paying for expensive infrastructure projects designed to handle larger loads. These savings can then be passed on to model developers who agree to design their applications to favor less consumptive versions. Informed users can choose specialized low-cost models that reduce compute, memory, and

[146] Samar Pratap, *The fine art of fine-tuning: A structured review of advanced LLM fine-tuning techniques*, 11 NAT. LANG. PROCESSING J. 1 (Jun. 2025), https://doi.org/10.1016/j.nlp.2025.100144 (fine-tuning); Han Liu, *SEPTQ: A Simple and Effective Post-Training Quantization Paradigm for Large Language Models*, ASS'N FOR COMPUTING MACHINERY (Jul. 20, 2025), https://doi.org/10.1145/3690624.3709287 (quantization).

energy costs for data centers without negatively affecting their experience – a desirable outcome for all parties involved.

There is a risk that immediate disclosure could incidentally increase overall AI energy demand as users consume more of the relatively less consumptive option. Evidence for this emerges from research on the Jevons Paradox, which suggests that in some cases efficiency improvements can drive users to consume more, not less, because each marginal unit of consumption costs less.[147] In a market moving as quickly as AI, it is possible that demand will counteract efficiency improvements and increase total energy demand.[148] It is also possible that pro-environmental AI use will crowd out other pro-environmental behaviors if users feel that efficient AI use justifies other, relatively more influential actions.[149] As the 2025 study by Luccioni and colleagues suggests, this means the AI industry should take a more nuanced and wider reaching approach to calculating environmental costs to avoid negative rebound effects.[150]

We also note, though, that the Jevons effect is often far less powerful in the real world than in abstract theory. As energy-efficient LED lighting was starting to become widely available, some analysts predicted that it would lead, through a Jevons effect, to rapidly rising household energy consumption, as the decreasing cost of lighting led people to install more and brighter lights. In fact, the opposite happened, and as more than a billion energy-efficient light bulbs were installed in US homes, per-capita residential electricity consumption dropped considerably.[151]

The wide variations in footprint calculator outputs when evaluating the same queries could indicate that some footprint calculators are not well designed, but it also could indicate that those who construct footprint

---

[147] Mario Giampietro & Kozo Mayumi, *Unraveling the Complexity of the Jevons Paradox: The Link Between Innovation, Efficiency, and Sustainability*, 6 FRON. ENERGY RES. 26 (Apr. 3, 2018), https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2018.00026/full.

[148] Robery Diab, *Jevons Paradox Makes Regulating AI Sustainability Imperative*, TECH POLICY PRESS (Mar. 11, 2025), https://www.techpolicy.press/jevons-paradox-makes-regulating-ai-sustainability-imperative/.

[149] Grant. D. Jacobsen, Matthew J. Kotchen, & Michael P. Vandenbergh, *The behavioral response to voluntary provision of an environmental public good: Evidence from residential electricity demand*, 56 EUR. ECON. REV. 946 (Jul. 2012), https://www.sciencedirect.com/science/article/pii/S0014292112000268.

[150] Alexandra Sasha Luccioni, Emma Strubell, & Kate Crawford, *From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate*, ARXIV (JUN. 13, 2025), https://arxiv.org/pdf/2501.16548.

[151] Jonathan M. Gilligan & Michael P. Vandenbergh, *A framework for assessing the impact of private climate governance*, 60 ENERGY RSCH. & SOC. SCI. 101400, 2-3 (Feb. 2020), http://www.sciencedirect.com/science/article/pii/S2214629619310370.

calculators lack reliable information on AI energy use and environmental impacts. Although both may be true, the limited public disclosures from AI firms about their models' energy use and environmental impacts suggest that the latter is the more likely explanation. Public surveys indicate that protecting the environment is important to a large majority of the American population, including consumers, employees, and managers of companies and other organizations. In the absence of reliable information, AI users cannot act on their preferences to reduce the environmental impacts of their AI use. Public or private standards for AI energy and environmental disclosure could address this problem and enable footprint calculators to provide the public with accurate estimates. Doing so also can align the needs of users, model developers, and data center operators.

V. Conclusion

Ensuring that information is available to the public has been an important foundation for environmental protection in the United States for at least fifty years. Major statutes ranging from the Clean Air Act, the Clean Water Act, and the Toxic Release Inventory provisions of Emergency Planning and Community Response-to-Know Act all rely on public disclosure of information to enable the public, corporate managers, and government policymakers to make informed decisions about environmental protection. Environmental footprint calculators are a principal means of conveying information about the use of AI. Yet this Article has compared the outputs of four AI-focused footprint calculators and concluded that they produce estimates that vary by as much as 58 times when assessing the same use of AI. This wide disparity in output despite the same input indicates that reliable information is not available to the public about the energy and environmental impacts of AI.

This lack of information would not be important if projected AI energy use were not driving significant growth in electricity demand and environmental impacts, but the available studies suggest that AI is responsible for rapidly growing electricity demand and environmental harms. To address this information gap, this Article argues that public or private actors should induce public disclosures from AI firms and data center operators. Governments are unlikely to play this role in the near term, but the analysis in this paper suggests that wide variations are likely

to exist among AI firms, models, and queries, and thus that some firms may have incentives for additional disclosure.[152] In turn, this disclosure can drive changes in the timing, type, and amount of demand in ways that can reduce the energy costs and environmental effects of AI. This disclosure can be done without slowing or shifting the development of this promising new technology. As its proponents argue, AI has remarkable potential to contribute to economic success and environmental protection, but greater transparency will be necessary to accelerate the net contributions of AI.

[152] *See, e.g.*, Vandenbergh, supra note 31, at 137-50 (discussing private environmental standards developed by collaborations among advocacy groups and corporations).