Building Green Generative AI: An Ecosystem-Wide Approach to Environmental Sustainability

Fabian Helms, Kay Hönemann, Manuel Wiesche

TU Dortmund University, Department of Business & Economics, Dortmund, Germany
fabian.helms@tu-dortmund.de, kay.hoenemann@tu-dortmund.de,
manuel.wiesche@tu-dortmund.de

Generative AI's (GenAI) rapid growth raises environmental concerns due to high energy consumption. Despite accelerating technological advancements, understanding how different stakeholders in the GenAI ecosystem can contribute to environmental sustainability remains limited. We address this gap with a taxonomy of actions for environmentally sustainable GenAI ecosystems. Our taxonomy, developed through a design science approach combining literature review and case analysis, categorizes environmental sustainability interventions across resources, models, and usage. We identify key stakeholders (hardware manufacturers, cloud providers, model developers, application providers) and map their roles in implementing these actions. The taxonomy reveals trade-offs between performance, cost, and environmental sustainability, highlighting the need for context-specific strategies. Through an illustrative vignette, we demonstrate how GenAI application providers can systematically implement sustainability measures. We provide a framework for researchers and practitioners to develop environmentally responsible GenAI solutions, fostering coordinated action to ensure GenAI benefits without compromising environmental well-being.

DOI
https://doi.org/10.18690/um.fov.4.2025.38

ISBN
978-961-286-998-4

Keywords:
generative AI, sustainability, environment, design science, taxonomy, energy

University of Maribor Press

1 Introduction

By 2026, data centers and AI systems will consume approximately 1000 TWh of electricity annually, equivalent to Japan's total consumption (IEA, 2024), making energy the "primary bottleneck" for AI development (Lacey & Phillips, 2024). This surge, driven by increasingly complex AI models and popular GenAI tools like ChatGPT, strains energy infrastructure while increasing carbon emissions and water usage (Zuccon et al., 2023), creating tension between innovation and environmental sustainability.

While GenAI capabilities advance rapidly, understanding how different ecosystem stakeholders[1] can improve environmental sustainability remains limited. Current research primarily addresses technical optimizations (Bai et al., 2024; Jiang et al., 2024; Yu et al., 2024), yet environmentally sustainable deployment requires coordinated action across stakeholders. The complex relationships and roles within the GenAI ecosystem and their environmental sustainability opportunities are not yet clearly mapped.

To address this gap, we developed a taxonomy using Nickerson et al.'s (2013) design science-informed approach, combining conceptual-to-empirical and empirical-to-conceptual strategies. Taxonomies aim to conceptualize objects within a domain of interest to assist researchers and practitioners in deepening their understanding. Through a structured literature review (vom Brocke et al., 2015; Webster & Watson, 2002) and analysis of real-world cases across the ecosystem, we created a comprehensive framework that categorizes stakeholder actions and responsibilities to facilitate targeted environmental sustainability strategies throughout the GenAI ecosystem.

2 Background

Generative AI (GenAI), a technology capable of creating content like text, images, and code, relies on a complex ecosystem of interconnected components (Banh & Strobel, 2023). We can view this ecosystem as four interacting layers: the hardware layer (specialized chips like Nvidia GPUs); the cloud provider layer (Infrastructure-

[1] Stakeholders are non-human entities in the GenAI ecosystem, mostly organizations.

as-a-service offerings, e.g., AWS, Microsoft Azure, Google Cloud); the models layer (foundation models like GPT-4o to specialized models); and the applications layer (user-facing tools like ChatGPT and Midjourney). These layers interact dynamically, with each stage consuming energy as users engage with applications that utilize models hosted on cloud platforms running on specialized hardware.

Therefore, the GenAI capabilities come with a significant environmental cost, primarily through energy consumption projected to reach 1000 TWh by 2026 (IEA, 2024). This energy demand is driven by several factors: the computationally intensive training of large models, the ongoing energy required for inference, and the energy used for data storage and transfer. It is estimated that GPT-3 consumed 1,287MWh for training, while total inference demands required higher energy consumption (de Vries, 2023). Carbon intensity varies with electricity sources, which can be reduced by switching to sustainable energy sources or offsetting carbon emissions through certificates (Schwartz et al., 2020). Another aspect is the choice of models and their size, which significantly impact the energy consumption for training and inference (Argerich & Patino-Martinez, 2024; Everman et al., 2023). GenAI's rapid growth, with OpenAI's ChatGPT reaching 300 million weekly users by December 2024, further intensifies these concerns (Roth, 2024). Beyond energy, water usage is also a significant concern (Zuccon et al., 2023). Current research efforts in environmental GenAI sustainability have mostly focused on specific parts of the GenAI ecosystem (Verdecchia et al., 2023), such as model training (McDonald et al., 2022), inference (Samsi et al., 2023), or benchmarking (Hodak et al., 2024). A holistic, ecosystem-level perspective is largely missing.

3 Research Design and Methodology

We developed a taxonomy following the method proposed by Nickerson et al. (2013), which has become the de facto standard for taxonomy development in Information Systems (IS) research (Szopinski et al., 2019). This approach enables systematic classification of dimensions relevant to environmental GenAI sustainability, valuable for structuring complex domains while revealing relationships between elements and their theoretical foundations (Bailey, 1994; Schöbel et al., 2020). We adapted the methodology of Kundisch et al. (2022) to ensure rigorous development of our taxonomy.

Our taxonomy aims to equip GenAI ecosystem stakeholders with a tool for developing environmentally sustainable solutions based on real-world cases, benefiting both researchers and practitioners seeking guidance on environmental sustainability and cost-influencing activities. Thus, as a first step, it is necessary to identify a meta-characteristic, which serves as the overarching purpose of the taxonomy. In our case, the meta-characteristic stated as follows: "Features and properties of GenAI ecosystems that impact environmental sustainability. Then, the next step in a taxonomy development process is to define the ending conditions that end the iterative taxonomy building process. Nickerson et. al (2013) proposed five subjective ending conditions and eight objective ending conditions, which we adopted in our building process. We employed a hybrid approach combining conceptual-to-empirical, where characteristics and dimensions are derived from literature, researcher's existing knowledge, and individual judgment, and empirical-to-conceptual, where real-world objects are analyzed and grouped based on their shared characteristics. (Nickerson et al., 2013). Based on this approach we developed meta-dimensions (MDs), dimensions, characteristics, and their relevant stakeholders to formulate our taxonomy.

First, we conducted a structured literature review following vom Brocke et al. (2015) and Webster & Watson (2002), searching for sustainability-related GenAI publications across recognized databases (Scopus, AIS eLibrary, arXiv) with a post-2021 timeframe. We included arXiv to capture the most recent publications in this rapidly evolving field. We limited the timeframe to post-2021, as the launch of ChatGPT in 2022 marked the primary onset of GenAI ecosystem development and research. Our search strategy encompassed titles, abstracts, and keywords using the search string: ("sustainab*" OR "climat*" OR "energ*" OR "environmental") AND ("generative AI" OR "generative artificial intelligence" OR "genai" OR "llm" OR "large language model" OR "text-to-image"). After screening 1,671 articles, we identified 53 relevant publications, with four additional papers from forward and backward searching. We focused on articles addressing the direct environmental sustainability aspects of GenAI operations. Articles discussing the application of GenAI for sustainability in other fields, such as energy research (Kench et al., 2024) or enhancing climate literacy (Atkins et al., 2024), were excluded.

Table 1: Meta-Dimensions (MDs) and Dimensions

Meta-Dimensions (MDs)	Dimensions	Sources
Resources	Compute Resources	(Liu & Yin, 2024)
	Energy Resources	(Dodge et al., 2022)
Models	Model Size	(Argerich & Patino-Martinez, 2024)
	Training & Fine-tuning	(Albalak et al., 2024; Bai et al., 2024)
Usage	Inference	(Stojkovic et al., 2024)
	Input/Output	(Husom et al., 2024)
	User Interface	(Ren et al., 2023)

Source: Own

In the second iteration, to further ground our proposed taxonomy with practical relevance, we followed an empirical-to-conceptual approach by analyzing real-world cases from 20 key ecosystem stakeholders identified in an industry report, including hardware manufacturers, cloud computing providers, model providers, and application providers (Artificial Analysis, 2024). For those 20 stakeholders we conducted an internet search and analyzed the relevant documentations and applications for further information for the taxonomy building process. We identified no new characteristics in the second iteration. The taxonomy was finalized after verifying all ending conditions were met (Nickerson et al., 2013). After reaching saturation, we abstracted from the identified dimensions and characteristics to group them under meta-dimensions, which are presented in Table 1 (Nickerson et al., 2013; Strobel et al., 2024).

4 Results

The final taxonomy and their relevant stakeholders are presented in Table 2. Dimensions and characteristics are presented in the following.

Table 2: Taxonomy of Environmental Sustainability Actions in GenAI Ecosystems

MDs	Dimensions	Characteristics	Stakeholders			
Resources	Compute Resources	Hardware Choice	Hardware	Cloud	Model	Application
		Partitioning	Hardware	Cloud		
		Power Capping	Hardware	Cloud		
	Energy Resources	Tracking Emissions		Cloud	Model	Application
		Time & Location Shifting		Cloud	Model	Application
Models	Model Size	Model Compression			Model	Application
		Model Choice			Model	Application
		Specialized Models			Model	Application
	Training & Finetuning	Algorithms			Model	Application
		Data Management			Model	Application
Usage	Inference	Batching				Application
		Caching				Application
	Input/Output	Input Optimization				Application
		Output Optimization				Application
	User Interface	Energy Usage Display				Application

Source: own

4.1 Resources

4.1.1 Compute Resources

Hardware update strategies significantly reduce energy usage, as shown by the transition from NVIDIA T4 to A100 GPUs, which cuts carbon emissions by 83% for equivalent workloads (Liu & Yin, 2024). Novel architectures like memristor crossbar chips achieve 69% energy reduction compared to traditional systems (Wang et al., 2024). Several companies offer application specific integrated circuit (ASIC) chips specialized for AI computational efforts at greater energy efficiency, e.g. AWS Trainium and Inferentia or Google Tensor Processing Units (TPU) (Jouppi et al., 2023).

GPU partitioning effectively manages smaller workloads, reducing energy demand by up to 33% and enabling 55% faster fine-tuning while using 42% less energy, though with 2-9.5x slower computation (Amazon Web Services, 2025).

The power consumption of GPUs can be capped to reduce energy consumption and temperature of the chips, which can extend the lifespan of GPUs (Zhao et al., 2023). Power capping reduces GPU energy consumption during inference by 23.21% with only 6.7% increased inference time (Samsi et al., 2023). Similarly, training a BERT model with a 150W cap (vs. 250W) needs 108.5% of the time but only 87.7% of the energy (McDonald et al., 2022). A 20% lower frequency cap for the GPUs, which saves about 20% of power, can support a medium load without lowering the latency or throughput of inference (Stojkovic et al., 2024).

4.1.2 Energy Resources

Tools like FootPrinter help assess datacenter carbon footprints (Niewenhuis et al., 2024). Time and location shifting leverages varying renewable energy availability, reducing emissions for training GenAI models (Dodge et al., 2022; Jagannadharao et al., 2023). Routing inference requests to greener datacenters can reduce carbon emissions by 35% while maintaining acceptable latency (Chien et al., 2023). Leading cloud providers offer tools which allow users to track carbon emissions of their usage and inform users about the energy mix at their chosen locations.

4.2 Models

4.2.1 Model Size

Models with smaller number of parameters use less energy (Argerich & Patino-Martinez, 2024; Liu & Yin, 2024). Quantization techniques like GPT-Generated Unified Format (GGUF) reduce energy usage by decreasing numerical precision and thus computational requirements (Rajput & Sharma, 2024). A 4 bit quantized model uses less than half of the energy of a 16-bit model, while reducing inference latency 3x (Argerich & Patino-Martinez, 2024). Distillation is another approach to reduce model size by transferring knowledge from larger to smaller models for specific use cases (Alzoubi & Mishra, 2024). Some cloud providers offer services to distill models for customers' use cases.

Model selection should balance efficiency with performance requirements (Argerich & Patino-Martinez, 2024; Everman et al., 2023). Developers of end-user apps can also utilize techniques to choose the right model for the requested task by employing

model cascading or model routing. Model cascading starts the inference process with the smallest, most efficient model and only escalates requests to larger ones when necessary. This enables smaller models to answer simple tasks and only employs larger models for more complex tasks, reducing energy consumption (L. Chen et al., 2023; L. Chen & Varoquaux, 2024). Model routing assesses the complexity of the requested task to decide on the right model to answer successfully (L. Chen & Varoquaux, 2024). End-user applications like Perplexity and ChatGPT allow the user to choose a model for specific tasks. Perplexity also features an "auto mode" to select appropriate models based on users' prompts.

Developers have to decide whether to use a base model or fine-tune a model to their specific use case, requiring additional energy usage. Fine-tuning should be used judiciously, as specialized models for medical and financial domains often don't outperform base models (Jeong et al., 2024; X. Li et al., 2023). Specialized models still outperform large general models in classification tasks such as sentiment, approval/disapproval, emotions, and party positions (Bucher & Martini, 2024). While large general models can achieve many of the tasks of specialized models in a zero-shot manner with no further fine-tuning, the energy impact of larger models is greater than those of smaller models (Luccioni et al., 2024).

4.2.2 Training & Fine-tuning

Algorithms for fine-tuning models can be optimized to reduce the computational need of fine-tuning models. Fine-tuning optimization frameworks like GreenTrainer achieve 64% reduction in computation by adaptively selecting tensors based on importance and backpropagation cost, demonstrating energy-efficient training without performance sacrifice (Huang et al., 2024). Additional efficiency approaches during model training include data parallelism (dividing datasets across nodes), model parallelism (distributing model layers across nodes), and mixed-precision training (reducing floating-point types) (Bai et al., 2024).

For LLMs trained on massive text corpora, improving dataset efficiency significantly reduces energy demands. The quality of large text datasets vary which calls for careful selection of data to train a capable model and reduce energy consumption (Albalak et al., 2024). Data pruning estimates the importance of data points in the

dataset to prioritize those data points during training (D. Chen et al., 2024; Zhuang et al., 2023).

4.3 Usage

4.3.1 Inference

Batch processing improves GPU utilization and energy efficiency, particularly for smaller models (Argerich & Patino-Martinez, 2024). Reducing batch size during low demand can save up to 15% of energy consumption (Stojkovic et al., 2024). Most API providers offer services to process requests in batches to save cost, compute, and thus energy usage.

Caching requested answers in databases reduces energy usage by retrieving stored responses for similar prompts rather than re-running inference, particularly beneficial for LLM-powered search engines (Betti et al., 2024). Most API providers offer prompt caching capabilities that store parts of prompts or documents so that in future requests models can access this cached information without reprocessing.

4.3.2 Input/Output

For multimodal LLMs, selecting only important image tokens can reduce token count by 79%, decreasing processing time by 67% and memory usage by 30% (Betti et al., 2024). Similarly, simplifying text prompts and concatenating multiple queries into single prompts reduces computational demands (L. Chen et al., 2023; Husom et al., 2024).

Image generation is more energy-intensive than text generation (Luccioni et al., 2024). When generating images with diffusion models, reducing the number of inference steps and image resolution lowers the energy usage (Seyfarth et al., 2025). Early detection of unsuccessful image generations can reduce inference time by 12% (Betti et al., 2024). For text generation, energy usage scales with the length of the generated text (Husom et al., 2024). Instructing larger models to provide concise answers can reduce carbon footprint by 40%, though with reduced accuracy for complex reasoning tasks (B. Li et al., 2024). This can be realized through most API providers by adapting the system message to generate shorter answers or defining a

maximum number of tokes to return. End-user applications such as ChatGPT and Claude also allow users to define system preferences or styles which could include an instruction to generate shorter answers.

4.3.3 User Interface

A way to save energy through reducing the use of GenAI tools is by presenting the user with information about the environmental sustainability of the continued use of the tool. Displaying energy usage information through visualizations effectively promotes sustainable usage, with users responding positively to transparency about environmental impact (Ren et al., 2023).

5 Application

We illustrate the taxonomy's practical application with a vignette describing choices and events (Miles et al., 2020). Consider "Imaginarium," a hypothetical text-to-image generation service similar to DALL-E or Midjourney. Using the taxonomy, Imaginarium can systematically analyze its operations to identify environmental sustainability opportunities across different ecosystem layers.

At the resource level (cloud provider), Imaginarium could select providers offering energy-efficient hardware (newer GPUs, TPUs) and utilize tools to track emissions. They could strategically choose data centers with more renewable energy and explore time/location shifting for training and inference.

At the model level (model/API providers, application developers), Imaginarium could offer a smaller, efficient diffusion model, potentially accepting a minor image quality trade-off for substantial energy reductions. Techniques like quantization could further minimize the model's memory and computational demands.

Within the usage layer (application developers, end-users), Imaginarium could offer batch inference for predictable loads or cache generated images for frequent prompts. Features encouraging smaller images or fewer inference steps could be implemented, with transparent feedback on energy implications. This example shows how applying the taxonomy helps Imaginarium identify concrete environmental sustainability actions across multiple GenAI ecosystem layers.

6 Discussion

We presented a taxonomy of stakeholders and activities in the GenAI ecosystem specifically designed to identify environmental sustainability opportunities. This holistic framework is crucial for understanding GenAI's complex environmental footprint. We demonstrate that responsibility for environmental GenAI sustainability cannot be assigned to a single stakeholder or confined to a specific layer within the ecosystem. It needs to be a coordinated effort, encompassing hardware manufacturers, cloud providers, model developers, application developers, and even end-users. Given the great importance of platforms for the development, access, and deployment of GenAI capabilities, the interconnectedness of these platforms and their diverse stakeholders means that actions and challenges at one level impact others, necessitating coordinated approaches for environmental sustainability (Heimburg et al., 2025).

We show inherent trade-offs between performance, cost, and environmental sustainability. For instance, larger models offer higher performance but consume significantly more energy (Argerich & Patino-Martinez, 2024; Liu & Yin, 2024). Conversely, model compression can reduce energy consumption, but may slightly decrease performance (Rajput & Sharma, 2024). The optimal approach is highly context-dependent, contingent upon the specific application, its performance requirements, and the acceptable environmental impact.

We identify numerous opportunities for enhancing efficiency at various stages, from hardware choice and data center operations to model optimization and application-level choices. These efficiency improvements not only mitigate environmental impact but can also lead to significant cost savings, creating a mutually beneficial scenario for both environmental sustainability and economic viability.

We show that the ecosystem is closely connected and measures to achieve environmental sustainability rely on different stakeholders to be achieved successfully. Transparent reporting standards of the environmental effects of GenAI operations are therefore crucial. Hardware manufacturers and cloud providers should offer standardized reports on the energy usage and environmental sustainability of their offerings. Model providers can then reliably calculate and report environmental factors for training and inference. This ecosystem-wide

transparency enables application providers and end-users to make informed choices about their GenAI usage and its environmental impact. Hugging Face, a platform for hosting GenAI models, has created a promising initiative to increase transparency in the reporting of environmental impact of different GenAI models (Hugging Face, 2025).

While we provide a comprehensive taxonomy, we acknowledge limitations and outline promising directions for future research. The GenAI ecosystem is characterized by rapid and continuous evolution. New models, hardware advancements, and innovative techniques are constantly emerging. Consequently, the taxonomy will require regular updates and refinements to maintain its relevance and accuracy in this dynamic landscape. A recent example is the release of reasoning models such as OpenAI o1 (OpenAI, 2024) and DeepSeek R1 (DeepSeek-AI et al., 2025), which, by design, produce a greater amount of tokens during inference and thus have higher energy demands. Furthermore, the taxonomy identifies opportunities but doesn't precisely quantify environmental impact for each action, focusing primarily on technical and operational aspects.

Future research should explore the role of economic incentives, such as carbon pricing mechanisms, and policy regulations to encourage sustainable practices across the ecosystem. Understanding user behavior is another critical area for future investigation. Further research is needed on how user interface design, information presentation, and user education can influence user choices and promote environmentally sustainable usage patterns.

In conclusion, we present a taxonomy promoting a holistic, environmentally sustainable approach to GenAI development and deployment. By promoting coordinated action across the entire ecosystem, we contribute to the responsible and environmentally sustainable development of this transformative technology. The application vignette demonstrates its practical value, offering a roadmap for companies improving their environmental sustainability. The goal is to ensure that the substantial benefits of GenAI can be realized without compromising long-term environmental well-being.

References

Albalak, A., Elazar, Y., Xie, S. M., Longpre, S., Lambert, N., Wang, X., Muennighoff, N., Hou, B., Pan, L., Jeong, H., Raffel, C., Chang, S., Hashimoto, T., & Wang, W. Y. (2024). A Survey on Data Selection for Language Models (No. arXiv:2402.16827). arXiv. http://arxiv.org/abs/2402.16827

Alzoubi, Y. I., & Mishra, A. (2024). Green artificial intelligence initiatives: Potentials and challenges. Journal of Cleaner Production, 468, 143090. https://doi.org/10.1016/j.jclepro.2024.143090

Amazon Web Services. (2025). AI Accelerator—AWS Trainium—AWS. Amazon Web Services, Inc. https://aws.amazon.com/ai/machine-learning/trainium/

Argerich, M. F., & Patino-Martinez, M. (2024). Measuring and Improving the Energy Efficiency of Large Language Models Inference. IEEE Access, 12, 80194–80207. Scopus. https://doi.org/10.1109/ACCESS.2024.3409745

Artificial Analysis. (2024). Artificial Analysis AI Review—2024 Highlights. https://artificialanalysis.ai/downloads/ai-review/2024/Artificial-Analysis-AI-Review-2024-Highlights.pdf

Atkins, C., Girgente, G., Shirzaei, M., & Kim, J. (2024). Generative AI tools can enhance climate literacy but must be checked for biases and inaccuracies. Communications Earth and Environment, 5(1). Scopus. https://doi.org/10.1038/s43247-024-01392-w

Bai, G., Chai, Z., Ling, C., Wang, S., Lu, J., Zhang, N., Shi, T., Yu, Z., Zhu, M., Zhang, Y., Yang, C., Cheng, Y., & Zhao, L. (2024). Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models (No. arXiv:2401.00625; Version 3). arXiv. https://doi.org/10.48550/arXiv.2401.00625

Bailey, K. D. (1994). Typologies and Taxonomies: An Introduction to Classification Techniques. SAGE Publications. https://books.google.de/books?id=1TaYulGjhLYC

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. Electronic Markets, 33(1), 63. https://doi.org/10.1007/s12525-023-00680-1

Betti, F., Baraldi, L., Baraldi, L., Cucchiara, R., & Sebe, N. (2024). Optimizing Resource Consumption in Diffusion Models through Hallucination Early Detection (No. arXiv:2409.10597; Version 1). arXiv. https://doi.org/10.48550/arXiv.2409.10597

Bucher, M. J. J., & Martini, M. (2024). Fine-Tuned "Small" LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification (No. arXiv:2406.08660). arXiv. https://doi.org/10.48550/arXiv.2406.08660

Chen, D., Huang, Y., Ma, Z., Chen, H., Pan, X., Ge, C., Gao, D., Xie, Y., Liu, Z., Gao, J., Li, Y., Ding, B., & Zhou, J. (2024). Data-Juicer: A One-Stop Data Processing System for Large Language Models. Companion of the 2024 International Conference on Management of Data, 120–134. https://doi.org/10.1145/3626246.3653385

Chen, L., & Varoquaux, G. (2024). What is the Role of Small Models in the LLM Era: A Survey (No. arXiv:2409.06857; Version 3). arXiv. https://doi.org/10.48550/arXiv.2409.06857

Chen, L., Zaharia, M., & Zou, J. (2023). FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance (No. arXiv:2305.05176; Version 1). arXiv. https://doi.org/10.48550/arXiv.2305.05176

Chien, A. A., Lin, L., Nguyen, H., Rao, V., Sharma, T., & Wijayawardana, R. (2023). Reducing the Carbon Impact of Generative AI Inference (today and in 2035). 2nd Workshop on Sustainable Computer Systems, HotCarbon 2023. Scopus. https://doi.org/10.1145/3604930.3605705

de Vries, A. (2023). The growing energy footprint of artificial intelligence. Joule, 7(10), 2191–2194. https://doi.org/10.1016/j.joule.2023.09.004

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (No. arXiv:2501.12948). arXiv. https://doi.org/10.48550/arXiv.2501.12948

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A. S., Smith, N. A., DeCario, N., & Buchanan, W. (2022). Measuring the Carbon Intensity of AI in Cloud Instances. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, 1877–1894. https://doi.org/10.1145/3531146.3533234

Everman, B., Villwock, T., Chen, D., Soto, N., Zhang, O., & Zong, Z. (2023). Evaluating the Carbon Impact of Large Language Models at the Inference Stage. 150–157. Scopus. https://doi.org/10.1109/IPCCC59175.2023.10253886

Heimburg, V., Schreieck, M., & Wiesche, M. (2025). Complementor Value Co-Creation in Generative AI Platform Ecosystems. Journal of Management Information Systems, 42, 2.

Hodak, M., Ellison, D., Van Buren, C., Jiang, X., & Dholakia, A. (2024). Benchmarking Large Language Models: Opportunities and Challenges. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 14247 LNCS, 77–89. Scopus. https://doi.org/10.1007/978-3-031-68031-1_6

Huang, K., Yin, H., Huang, H., & Gao, W. (2024). TOWARDS GREEN AI IN FINE-TUNING LARGE LANGUAGE MODELS VIA ADAPTIVE BACKPROPAGATION. 12th International Conference on Learning Representations, ICLR 2024. Scopus.

Hugging Face. (2025). AI Energy Score. AI Energy Score. https://huggingface.github.io/AIEnergyScore/

Husom, E. J., Goknil, A., Shar, L. K., & Sen, S. (2024). The Price of Prompting: Profiling Energy Use in Large Language Models Inference (No. arXiv:2407.16893; Version 1). arXiv. https://doi.org/10.48550/arXiv.2407.16893

IEA. (2024, January 24). Electricity 2024. IEA. https://www.iea.org/reports/electricity-2024

Jagannadharao, A., Beckage, N., Nafus, D., & Chamberlin, S. (2023). Timeshifting strategies for carbon-efficient long-running large language model training. Innovations in Systems and Software Engineering. Scopus. https://doi.org/10.1007/s11334-023-00546-x

Jeong, D. P., Garg, S., Lipton, Z. C., & Oberst, M. (2024). Medical Adaptation of Large Language and Vision-Language Models: Are We Making Progress? (No. arXiv:2411.04118). arXiv. https://doi.org/10.48550/arXiv.2411.04118

Jiang, P., Sonne, C., Li, W., You, F., & You, S. (2024). Preventing the Immense Increase in the Life-Cycle Energy and Carbon Footprints of LLM-Powered Intelligent Chatbots. Engineering, 40, 202–210. https://doi.org/10.1016/j.eng.2024.04.002

Jouppi, N. P., Kurian, G., Li, S., Ma, P., Nagarajan, R., Nai, L., Patil, N., Subramanian, S., Swing, A., Towles, B., Young, C., Zhou, X., Zhou, Z., & Patterson, D. (2023). TPU v4: An Optically Reconfigurable Supercomputer for Machine Learning with Hardware Support for Embeddings (No. arXiv:2304.01433). arXiv. https://doi.org/10.48550/arXiv.2304.01433

Kench, S., Squires, I., Dahari, A., Brosa Planella, F., Roberts, S. A., & Cooper, S. J. (2024). Li-ion battery design through microstructural optimization using generative AI. Matter, 7(12), 4260–4269. Scopus. https://doi.org/10.1016/j.matt.2024.08.014

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An Update for Taxonomy Designers. Business & Information Systems Engineering, 64(4), 421–439. https://doi.org/10.1007/s12599-021-00723-x

Lacey, S., & Phillips, N. (2024, May 22). Energy is now the 'primary bottleneck' for AI. Latitude Media. https://www.latitudemedia.com/news/energy-is-now-the-primary-bottleneck-for-ai

Li, B., Jiang, Y., Gadepally, V., & Tiwari, D. (2024). Toward Sustainable GenAI using Generation Directives for Carbon-Friendly Large Language Model Inference (No. arXiv:2403.12900; Version 1). arXiv. https://doi.org/10.48550/arXiv.2403.12900

Li, X., Chan, S., Zhu, X., Pei, Y., Ma, Z., Liu, X., & Shah, S. (2023). Are ChatGPT and GPT-4 General-Purpose Solvers for Financial Text Analytics? A Study on Several Typical Tasks (No. arXiv:2305.05862). arXiv. http://arxiv.org/abs/2305.05862

Liu, V., & Yin, Y. (2024). Green AI: Exploring carbon footprints, mitigation strategies, and trade offs in large language model training. Discover Artificial Intelligence, 4(1). Scopus. https://doi.org/10.1007/s44163-024-00149-w

Luccioni, S., Jernite, Y., & Strubell, E. (2024). Power Hungry Processing: Watts Driving the Cost of AI Deployment? Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 85–99. https://doi.org/10.1145/3630106.3658542

McDonald, J., Li, B., Frey, N., Tiwari, D., Gadepally, V., & Samsi, S. (2022). Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models. 1962–1970. Scopus.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2020). Qualitative data analysis: A methods sourcebook (Fourth edition). SAGE.

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. European Journal of Information Systems, 22(3), 336–359. https://doi.org/10.1057/ejis.2012.26

Niewenhuis, D., Talluri, S., Iosup, A., & De Matteis, T. (2024). FootPrinter: Quantifying Data Center Carbon Footprint. 189–195. Scopus. https://doi.org/10.1145/3629527.3651419

OpenAI. (2024). Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/

Rajput, S., & Sharma, T. (2024). Benchmarking Emerging Deep Learning Quantization Methods for Energy Efficiency. 238–242. Scopus. https://doi.org/10.1109/ICSA-C63560.2024.00049

Ren, Y., Sivakumaran, A., Niemelä, E., & Jääskeläinen, P. (2023). How to Make AI Artists Feel Guilty in a Good Way? : Designing Integrated Sustainability Reflection Tools (SRTs) for Visual Generative AI. ICCC International Conference of Computational Creativity, Waterloo, Canada, 19-23 June, 2023. https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-327918

Roth, E. (2024, December 4). ChatGPT now has over 300 million weekly users. The Verge. https://www.theverge.com/2024/12/4/24313097/chatgpt-300-million-weekly-users

Samsi, S., Zhao, D., McDonald, J., Li, B., Michaleas, A., Jones, M., Bergeron, W., Kepner, J., Tiwari, D., & Gadepally, V. (2023). From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. 2023 IEEE High Performance Extreme Computing Conference, HPEC 2023. Scopus. https://doi.org/10.1109/HPEC58863.2023.10363447

Schöbel, S. M., Janson, A., & Söllner, M. (2020). Capturing the complexity of gamification elements: A holistic approach for analysing existing and deriving novel gamification designs. European Journal of Information Systems, 29(6), 641–668. https://doi.org/10.1080/0960085X.2020.1796531

Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63. https://doi.org/10.1145/3381831

Seyfarth, M., Dar, S. U. H., & Engelhardt, S. (2025). Latent Pollution Model: The Hidden Carbon Footprint in 3D Image Synthesis. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 15187 LNCS, 146–156. Scopus. https://doi.org/10.1007/978-3-031-73281-2_14

Stojkovic, J., Choukse, E., Zhang, C., Goiri, I., & Torrellas, J. (2024). Towards Greener LLMs: Bringing Energy-Efficiency to the Forefront of LLM Inference (No. arXiv:2403.20306; Version 1). arXiv. https://doi.org/10.48550/arXiv.2403.20306

Strobel, G., Banh, L., Möller, F., & Schoormann, T. (2024). Exploring Generative Artificial Intelligence: A Taxonomy and Types. Hawaii International Conference on System Sciences 2024 (HICSS-57). https://aisel.aisnet.org/hicss-57/in/platform_ecosystems/7

Szopinski, D., Schoormann, T., & Kundisch, D. (2019, June 13). Because your taxonomy is worth it: Towards a framework for taxonomy evaluation. ECIS.

Verdecchia, R., Sallou, J., & Cruz, L. (2023). A systematic review of Green AI. WIREs Data Mining and Knowledge Discovery, 13(4), e1507. https://doi.org/10.1002/widm.1507

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. Communications of the Association for Information Systems, 37(1). https://doi.org/10.17705/1CAIS.03709

Wang, Z., Luo, T., Liu, C., Liu, W., Goh, R. S. M., & Wong, W.-F. (2024). Enabling Energy-Efficient Deployment of Large Language Models on Memristor Crossbar: A Synergy of Large and Small. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1–17. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2024.3483654

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. MIS Quarterly, 26(2), xiii–xxiii.

Yu, Y., Wang, J., Liu, Y., Yu, P., Wang, D., Zheng, P., & Zhang, M. (2024). Revisit the environmental impact of artificial intelligence: The overlooked carbon emission source? Frontiers of Environmental Science and Engineering, 18(12). Scopus. https://doi.org/10.1007/s11783-024-1918-y

Zhao, D., Samsi, S., McDonald, J., Li, B., Bestor, D., Jones, M., Tiwari, D., & Gadepally, V. (2023). Sustainable Supercomputing for AI: GPU Power Capping at HPC Scale. 588–596. Scopus. https://doi.org/10.1145/3620678.3624793

Zhuang, B., Liu, J., Pan, Z., He, H., Weng, Y., & Shen, C. (2023). A survey on efficient training of transformers. Proceedings of the 32nd International Joint Conference on Artificial Intelligence, IJCAI 2023, 6823–6831. https://research.monash.edu/en/publications/a-survey-on-efficient-training-of-transformers

Zuccon, G., Scells, H., & Zhuang, S. (2023). Beyond CO2 Emissions: The Overlooked Impact of Water Consumption of Information Retrieval Models. 283–289. Scopus. https://doi.org/10.1145/3578337.3605121