Toward Green Agentic AI: Balancing Performance with Environmental Sustainability

Yuekai Wang

Rancho Bernardo High School, San Diego, CA, USA
kevinyuekai@gmail.com

Abstract

Agentic AI is increasingly used across domains due to its autonomous decision-making capabilities on multi-step tasks. However, more agent interactions increase electricity consumption, water usage, and carbon emissions. Using xAI's new supercomputer in Memphis as a case study, we examine the environmental implications of emerging large-scale AI infrastructure. We interviewed Dr. Veronica Bolon-Canedo (University of A Coruna), Dr. Edith Cohen (Google Research), and Dr. Jon Krohn (SuperDataScience Host) to gain insight into the trade-offs between performance and cost. We discuss potential solutions for minimizing environmental impact, such as using self-refining local LLMs, deploying task-specialized small models, and promoting public awareness. Our findings highlight the importance of integrating green principles into agentic AI development to align innovation with environmental responsibility.

1 About Agentic AI

Agentic artificial intelligence (AI) is a new kind of system designed to perform complex multi-step tasks independently, without explicit human input. Unlike traditional software, it can evaluate situations, make decisions, and adapt dynamically. For example, multiple agents may coordinate to find a movie ticket, adjusting to sold-out shows or new releases. Today, agentic AI powers customer service, scheduling, and social media systems, marking a shift from rigid programming to adaptable intelligence.

However, a striking example of how agentic AI can be problematic is the "Colossus" supercomputer data center recently constructed by Elon Musk's xAI in Memphis, Tennessee. Built in just four months, the facility is accused of bypassing several environmental regulations and began operating without a permit. Shortly afterward, local residents reported higher rates of asthma and respiratory illnesses. Colossus illustrates the tangible environmental and public health costs of sustaining large-scale computational systems required by agentic AI. As these autonomous technologies continue to expand, the need to balance innovation with environmental responsibility becomes increasingly urgent.

2 Problems with Agentic AI

Agentic AI leads to increased electricity consumption, water usage for cooling, and carbon emissions because each reasoning step triggers an additional call to an LLM operating in energy-intensive data centers. For example, when finding a simple movie ticket, three agents may reason seven times, resulting in seven LLM calls. This scales exponentially for more complex tasks, and according to MIT Technology Review, training a single AI model can emit as much carbon dioxide as five cars over their entire lifetimes.

The widespread adoption of agentic AI across industries reflects competitive pressure rather than necessity. As Dr. Veronica Bolon-Canedo, a computer scientist and AI researcher, observed

in an interview, “Everyone was trying to apply deep learning to every single problem, even when sometimes you don’t need it.” The same pattern now appears in the shift to agentic AI, where novelty often outweighs sustainability. She further notes that the resulting energy consumption creates ethical dilemmas, particularly when the benefits are minimal. When it comes to recommending a better Netflix movie, ”It’s not ethical to waste so much energy or money on things that are for entertainment.” These observations underscore the urgent need for environmental accountability in the deployment of agentic AI systems.

3 Potential Solutions

There remains an encouraging path toward sustainable agentic AI. Dr. Jon Krohn, a company founder and AI podcast host, suggested in an interview that “instead of calling some huge model in the cloud, like GPT-4o from OpenAI, you can actually have very small LLMs, like some small Llama model that fits on some small GPU, running on your desktop.” Through efficient deployment of compact models, comparable performance can often be achieved at a fraction of the energy cost.

Reducing computational demand can also be accomplished by specializing smaller agents for specific tasks. A specialized agent does not require a large, all-purpose model operating in a data center. In my research developing an agentic AI system to assist elderly individuals, this task-specific design approach has proven effective in maintaining performance while minimizing unnecessary computation.

Education and user awareness further play a crucial role in mitigating environmental impacts. As Dr. Bolon-Canedo emphasizes, “Most people don’t know that if they say ‘thank you’ and ChatGPT answers ‘you’re welcome,’ this has an impact.” Promoting sustainability therefore requires responsibility from all participants—developers, organizations, and users alike.

4 Finding the Balance

Smarter implementations of agentic AI can offer meaningful benefits when applied appropriately. Smaller, energy-efficient models may suffice for tasks such as assisting students with academic work, whereas more powerful systems are justified in high-stakes contexts like medical diagnosis. Responsible development means designing systems based on need rather than ego, balancing performance with sustainability.

Ultimately, greater scale does not always equate to greater value. As Google researcher Dr. Edith Cohen noted in an interview, AI is like planes: “They helped connect humanity. They were mostly helpful. They were also used to deliver an atomic bomb once.” Sustainable progress in agentic AI depends on maintaining this balance by maximizing innovation while upholding ethical and environmental responsibility.

5 Acknowledgments

I would like to thank Dr. Bolon-Canedo, Dr. Cohen, and Dr. Krohn for their valuable time and insights.