WHITE PAPER

The GenAI Era: A Path to Sustainable Innovation

Understanding the societal impact of GenAI and Large Language Models and how Chata.ai's solution aims to reduce operational inference costs by more than 97%, aligning advanced tech with environmental sustainability.

CHATA.AI

Table of Contents

In the dawn of the Generative Artificial Intelligence (GenAI) era, society finds itself at a crossroads, gazing into a future where the promise of GenAI and Large Language Models (LLMs) holds the potential to redefine human-machine interaction. These technologies, powered by NVIDIA A100 Graphics Processing Units (GPUs) and their ilk, have demonstrated an unprecedented ability to generate text, create art, solve complex problems, and even simulate human-like conversations. The allure of such capabilities has led to a surge in demand, with visions of millions, if not billions, of active users engaging with these platforms daily. However, this dream is shadowed by a pressing reality: the environmental cost of our digital ambitions. In this white paper, we will explore the anticipated obstacles faced by society in its current trajectory and will unveil our innovative solution designed to significantly reduce operational inference costs by more than 97% and cut carbon emissions by over 95%.

Environmental Implications of Digital Ambitions

The narrative thus far with LLMs, reminiscent of Maslow's Hammer, or the law of the instrument, reflects a society captivated by the newest tool at its disposal, attempting to apply it to every conceivable task, without a full consideration of the broader implications. GenAI and LLMs, for all their prowess, have become the proverbial hammer to which every challenge appears as a nail. Yet, this approach is neither sustainable nor practical, given the environmental toll and the inherent limitations of these technologies when applied indiscriminately.

Consider the scale: supporting 10 million active users on highly optimized implementations of these AI marvels would require an estimated 1,000,000 NVIDIA A100 GPUs. The electrical consumption for such an endeavor? A staggering 9,600,000 kilowatt-hours per day. This figure is not just a number but translates into a carbon footprint of approximately 3,840 metric tons of CO2 every single day. To put this into perspective, it's akin to the yearly emissions from electricity use of over 233,333 average households. And this is for just 10 million users—a fraction of the global population aspiring to leverage these technologies.

36.5%

projected annual compounded growth rate for GenAI from 2025-2030

Active Users	Carbon Emissions per year (metric tons)	Annual Emissions Equivalent
10,000	13,885	3,019 cars on the road
100,000	138,846	7,156,701 sq meters of commercial building space
1,000,000	1,388,460	1.54 M metric tons of cement production
10,000,000	13,884,600	Annual GHG emissions of mid-size US City
100,000,000	138,846,000	15% of total emissions from commercial airline industry

The recent NVIDIA announcement with the new Blackwell chips promise a 25% decrease in power consumption and heralds a significant stride towards efficiency. Yet, this progress, while commendable, casts only a dim light on the broader challenge that looms over our collective endeavors—the relentless demand on our power grids for the foreseeable future.

Even though we only focus on LLMs in this discussion, we cannot forget that high performance hardware is also required for the Metaverse which arguably has more momentum behind it than GenAI. While it's optimistic to anticipate substantial energy savings from these innovations such as the Blackwell chips, practical outcomes may fall short.

However, let's posit that the energy savings are as projected, hardware production aligns with demand, and similar advancements in energy efficiency occur biennially. Even with these favorable conditions, the electrical infrastructure remains ill-equipped to support the escalating demands from AI, cryptocurrency mining, and the Metaverse, not to mention the impending requirements of public electric vehicle (EV) mandates. With the US Energy Information Administration (EIA) only projecting an annual growth rate of just below 1% in electricity consumption for the next 25 years, it is easy to see how there can be misalignment with the projected annual compounded growth rate of 36.5% for GenAI from 2025-2030.

The United States has experienced numerous instances of brownouts and blackouts, with the EIA reporting numerous such events in recent years, underscoring the grid's challenges in sustaining the existing demand. The widespread adoption of AI technologies without corresponding upgrades to our electrical infrastructure threatens to exacerbate this situation, potentially leading to more frequent and widespread power outages.

Towards a Sustainable Technological Ecosystem

Acknowledging this, there emerges a pressing need to diversify our technological toolkit. Not every digital task necessitates the computational and energy-intensive might of GenAI or LLMs. There are swathes of problems that could be addressed more efficiently, both in terms of computational resources and energy consumption, by simpler or more specialized technologies. For instance, routine queries or data processing tasks might be better served by less processing heavy models. Similarly, emerging AI technologies that prioritize efficiency and reduce carbon footprints, such as those employing more advanced forms of model pruning or leveraging quantum computing for specific types of calculations, offer promising alternatives.

This shift in approach calls for a broader societal understanding and acceptance of the principle that while GenAI and LLMs are powerful tools, they are not universal solutions. As we stand on the brink of this AI revolution, the path forward is clear. We must embrace a multi-faceted approach to technological development, one that balances the incredible potential of GenAI and LLMs with the imperative of environmental sustainability. This journey requires major changes to our electricity infrastructure, a re-evaluation of our carbon emissions policies, and, most importantly, a collective reimagining of how and when we choose to deploy our most advanced tools. In doing so, we ensure that the promise of AI enriches not just our lives but the health of our planet as well.

A Case Study in Innovation and Responsibility: Chata.ai

It is the harsh realization of the economic and environmental impacts of LLM use at scale that has motivated our R&D program to focus on highly scalable and highly economic technology. Copilots as well as new start-ups leveraging GenAI technologies are popping up everywhere in the tech scene.

Over 95%

Chata.ai's solution is designed to cut carbon emissions compared to technologies using GPUs

At Chata.ai we have a different approach. Our inference engine (which translates natural language to database query language) does not need A100 or similar GPU horsepower but can infer on Central Processing Units (CPUs). That may not mean a lot, but let's put this in perspective of economics and the environment. For the purposes of this analysis, we left out the training costs and footprint of the LLMs and just focused on inference.

Below is a table looking at the cost comparisons with varying parameter sizes of LLMs based on both real world and theoretical model sizes. We are basing the information on an optimized implementation even though many implementations of LLMs are not optimized.

Optimized LLM Implementations (1000 active* users)				
Model Size (Parameters)	Active Users per GPU (Optimized Estimation)	Estimated GPUs Needed (1,000 active users)	Estimated Monthly Cost** (1,000 active users)	Chata.ai*** as % of Total Estimated Monthly Cost
7B	40	25	$ 54,000	2.78%
13B	35	29	$ 62,640	2.39%
52B	30	33	$ 71,280	2.10%
70B	25	40	$ 86,400	1.74%
137B	22	45	$ 97,200	1.54%
175B	20	50	$ 108,000	1.39%
250B	18	56	$ 120,960	1.24%
500B	15	67	$ 144,720	1.04%
750B	12	83	$ 179,280	0.84%
1T (1000B)	10	100	$ 216,000	0.69%

**** Active user load is defined as 1 request every 30 sec.***
***** Cost is estimated at pre-purchased price of $3 USD per hour per GPU***
***** Chata.ai inference processing costs are approximately $1,500 per month for 1000 active users***

We next present a table that extends the cost analysis to look at electricity usage and carbon emissions again on optimized implementations.

Optimized LLM Implementations (1000 active* users)				
Model Size (Parameters)	Estimated GPUs Needed (1,000 active users)	Electricity Use (kWh/year)	Carbon Emissions (Metric Tons/year)	Chata.ai*** as % of Total Estimated Electricity & Carbon Emissions
7	25	87,600	35.04	5.37%
13	29	100,114	40.05	4.70%
52	33	116,800	46.72	4.03%
70	40	140,160	56.06	3.36%
137	45	159,273	63.71	2.95%
175	50	175,200	70.08	2.68%
250	56	194,667	77.87	2.42%
500	67	233,600	93.44	2.01%
750	83	292,000	116.8	1.61%
1000	100	350,400	140.16	1.34%

**** Active user load is defined as 1 request every 30 sec.***
***** Cost is estimated at pre-purchased price of $3 USD per hour per GPU***
***** Chata.ai electricity use ~4,704 kWh/year and Carbon Emissions ~1.88 Metric Tons/year for 1000 active users***

At Chata.ai, we stand at the crossroads of innovation and responsibility, deeply aware that our journey in conversational AI is but a thread in the vast tapestry of technological possibilities LLMs unveil. Our dedication transcends mere admiration for these models; it's a commitment to integrate them with an eye towards both economic and environmental stewardship. This ethos is why we've woven interoperability with LLMs into our fabric, ensuring that as we contribute to the evolution of conversational AI, we do so not only with readiness but with a profound respect for the broader impacts.

In this endeavor, we're not just preparing to meet the future — we're shaping it to be as sustainable as it is intelligent, embodying the true spirit of innovation that benefits all.

As we weave these advancements into the fabric of our digital society, the stark reality remains that even a substantial reduction in energy use does not alleviate the underlying stress on our electrical infrastructure. This conundrum serves as a poignant reminder that our pursuit of technological excellence must be matched with equal vigor towards reimagining and fortifying the very foundations that power our innovations. Only then can we truly harness the full potential of these breakthroughs, ensuring they contribute not just to our digital evolution, but to a sustainable coexistence with the planet that sustains us all.

Reflections and Future Directions

In a twist laden with irony, our exploration of the environmental impacts and technological trajectories of AI has been guided in part by the very subject at its heart: ChatGPT-4. This AI, a testament to the strides made in the field of LLMs, has not only aided in articulating our narrative but also in verifying the myriad of data scattered across the digital expanse. Through this lens, our journey underscores a profound reflection on the dual role AI plays—as both the harbinger of the future we strive towards and a mirror reflecting our current dilemmas. As we navigate these waters, ChatGPT-4 serves as a beacon, illuminating the path forward while reminding us of the intricate dance between technology's promise and its footprint on the world.

On behalf of the AI community, a special thanks to Hugging Face for their research and many references in the paper. It was inspirational in forming some key assumptions and helping to validate some of the information being output by ChatGPT-4 that were utilized in exploring the economic and environmental impacts of LLMs.