Reducing the Carbon Impact of Generative AI Inference (today and in 2035)

Andrew A. Chien*
University of Chicago &
Argonne National Laboratory
Chicago, IL, USA
achien@cs.uchicago.edu

Liuzixuan Lin*
University of Chicago
Chicago, IL, USA
lzixuan@uchicago.edu

Hai Nguyen*
University of Chicago
Chicago, IL, USA
ndhai@cs.uchicago.edu

Varsha Rao*
University of Chicago
Chicago, IL, USA
varsharao@uchicago.edu

Tristan Sharma*
University of Chicago
Chicago, IL, USA
tristansharma@uchicago.edu

Rajini Wijayawardana*
University of Chicago
Chicago, IL, USA
rajini@uchicago.edu

ABSTRACT

Generative AI, exemplified in ChatGPT, Dall-E 2, and Stable Diffusion, are exciting new applications consuming growing quantities of computing. We study the compute, energy, and carbon impacts of generative AI inference. Using ChatGPT as an exemplar, we create a workload model and compare request direction approaches (Local, Balance, CarbonMin), assessing their power use and carbon impacts.

Our workload model shows that for ChatGPT-like services, inference dominates emissions, in one year producing 25x the carbon-emissions of training GPT-3. The workload model characterizes user experience, and experiments show that carbon emissions-aware algorithms (CarbonMin) can both maintain user experience and reduce carbon emissions dramatically (35%). We also consider a future scenario (2035 workload and power grids), and show that CarbonMin can reduce emissions by 56%. In both cases, the key is intelligent direction of requests to locations with low-carbon power. Combined with hardware technology advances, CarbonMin can keep emissions increase to only 20% compared to 2022 levels for 55x greater workload. Finally we consider datacenter headroom to increase effectiveness of shifting. With headroom, CarbonMin reduces 2035 emissions by 71%.

CCS CONCEPTS

• **Social and professional topics → Sustainability**; • **Computing methodologies → Artificial intelligence**; • **Applied computing → Data centers**.

KEYWORDS

Generative AI, Sustainability, Carbon emissions, Large language models, Geographic shifting

*Authors contributed equally to this research.

HotCarbon '23, July 9, 2023, Boston, MA, USA

ACM ISBN 979-8-4007-0242-6/23/07.
https://doi.org/10.1145/3604930.3605705

ACM Reference Format:
Andrew A. Chien, Liuzixuan Lin, Hai Nguyen, Varsha Rao, Tristan Sharma, and Rajini Wijayawardana. 2023. Reducing the Carbon Impact of Generative AI Inference (today and in 2035). In *2nd Workshop on Sustainable Computer Systems (HotCarbon '23), July 9, 2023, Boston, MA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3604930.3605705

1 INTRODUCTION

Generative AI for text (e.g. ChatGPT [43]), images (e.g. DALL-E 2 [44]), and other media has growing creative, informational, and commercial applications. One representative application, ChatGPT, leads the explosive growth of generative AI, hitting 100 million monthly active users in January 2023 as the fastest growing application [23]. After OpenAI partnered with Microsoft, global tech companies (Google, Meta, Baidu, Alibaba, etc.) have announced a slew of generative AI applications [10, 12, 24, 47]. Many expect that generative AI applications will proliferate in daily life and commerce [25, 37].

Behind the intelligently generated content are large machine-learning models. GPT-3 [14], the model used in ChatGPT today, is a representative large language model (LLM) with 175 billion parameters. The cycle of developing a generative AI model can be divided into training and inference. Once trained, a model can serve many user requests. Much previous work has focused on the carbon impact of model training [13, 45, 49, 54, 58], and we believe inference (operation) can also be problematic, particularly with rapid user growth and integration into everyday applications. For example, generative AI-backed search can cost 5 times more compute per request [53], requiring billions of dollars of computing infrastructure [56], and increasing associated embodied and operational carbon emissions.

The carbon impact of an application depends on its workload characteristics, such as compute per request, latency requirement, and location of users. Given its rising popularity and usage, we use ChatGPT as exemplar for Generative AI. Thereby, we model and characterize the compute, energy, and carbon emissions of generative AI, and explore how to reduce its carbon impact. Specific contributions include:

- A ChatGPT-like application with estimated use of 11 million requests/hour produces emissions of 12.8k metric ton

CO_2/year, 25 times the emissions for training GPT-3. Inference is critical to environmental and power cost.
- We show that it's possible to perform geographic shifting on this user-responsive workload and maintain similar user experience. Further, CarbonMin, an algorithm that directs requests to low-carbon regions, reduces carbon emissions by 35% in today's power grids.
- Looking forward (2035), considering usage growth (55x) and power grid decarbonization (3x lower average carbon intensity), CarbonMin reduces emissions by 56%, but with usage growth this results in 1.2x emissions vs 2022 levels. Benefit is limited by datacenter capacity.
- Increasing datacenter headroom, enables *CarbonMin* to achieve 71% reduction for 1x headroom for 2035, a remarkable 73x reduction in per-inference emissions.

2 PROBLEM

Generative AI has set record for fastest growing technology. Along with its usage, its compute requirements and carbon emission are growing too [35, 48].In order to understand the carbon impact of generative model inference and find carbon reduction solutions, the following research questions need to be answered:

- What is generative AI inference's workload and user response requirements?
- What is its carbon emissions impact today? and how might it grow?
- Can inference serving be directed to reduce carbon impact today? in the future?

3 APPROACH

Carbon impact of computing power consumption depends heavily on where and when it happens, as grid carbon emissions are highly-variable across power grids/locations and the course of hours, days, weeks, and even seasons [2, 15, 17, 21, 33, 36, 57].

We propose to shift workload geographically to reduce carbon emissions [27, 31, 34, 40, 41, 60, 62, 63], but this is only possible when applications are flexible; more precisely latency-tolerant. Generally, user-facing inference has been considered inflexible and thus not suitable for shifting. Our study covers:

(1) Characterization of ChatGPT-like workloads' load pattern, predictability, and user-response requirements.
(2) A variety of request direction algorithms, using the workload, with focus on hardware utilization and carbon-emissions reduction.
(3) Evaluation of request direction algorithm effectiveness and ability to maintain user-responsiveness.
(4) A projection of these schemes in the future (2035 workloads and power grids), and even future datacenter capacity (1x headroom).

3.1 Characterizing Compute Load

We create a model of the ChatGPT workload and its service QoS. The ChatGPT load is predominantly human-generated and therefore follows a diurnal structure. Based on 1.6 billion visits in March 2023 [50], the assumption of 5 queries per visit produces 0.27 billions requests/day (*ChatGPT-RR* in Table 1). We distribute this load over 8 exemplar cities, based on documented ChatGPT usage rates [50] and national population, skewing for waking hours [28, 52]. Figure 1 shows resulting load, aligned with the US Pacific timezone. Note that the load is dominated by USA (39%) and European Countries (35%), reflecting their higher ChatGPT usage.

Workload Model	Inference Cost (GPU-hrs)	Training Cost (GPU-hrs)	Inference/ Training
ChatGPT-RR	55,966,667	2,236,467	25x
Google-RR	3,099,154,167	2,236,467	1386x

Table 1: Annual Compute for Inference and Model Training, various workload models (A100 GPU-hrs).

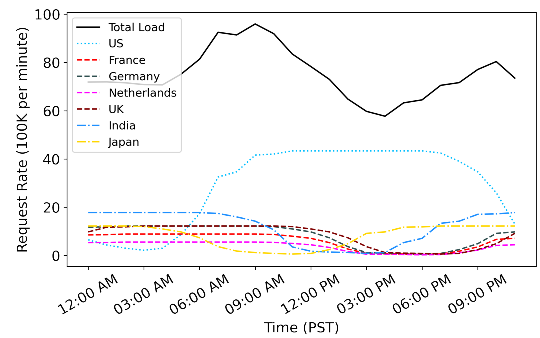


Figure 1: Present Load: Diurnal structure from waking hours, weighted by ChatGPT use and population [28, 52]

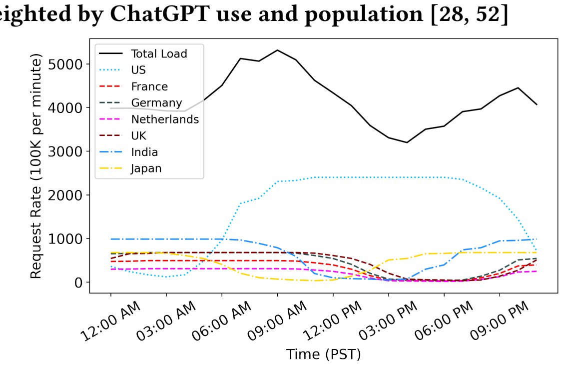
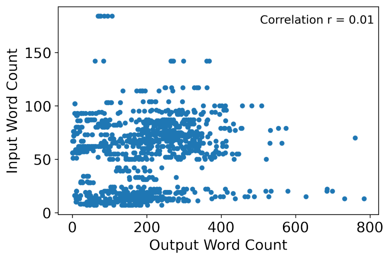


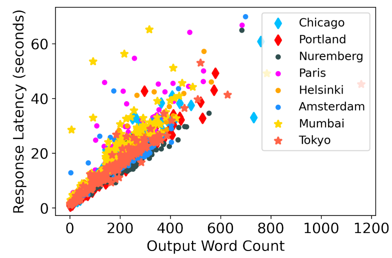
Figure 2: Future load based on Google visits [51] and diurnal structure from waking hours

To project future load, we scale usage up to match Google search request rates (88.6 billion/month), using 5 queries per visit [51] (*Google-RR in Table 1*). Load increases significantly, but 24-hour shape is similar (Figure 2).

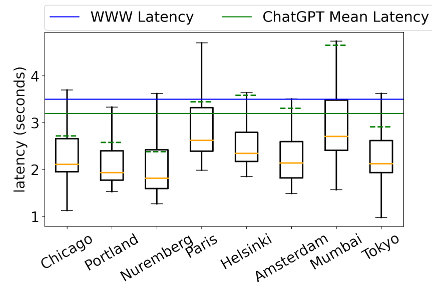
We estimate the total compute of ChatGPT inference based on averages of output word count (Table 1), for both *ChatGPT-RR* and future *Google-RR* load scenarios. Inference cost is based on the model in Section 4.1. For comparison we also include published training cost estimates [32] for GPT-3 scaled for A100 GPUs. The ratios for annual inference to model training cost are 25x and 1386x respectively.



(a) Input and output complexity by word count; correlation coefficient is 0.01



(b) ChatGPT: Response latency vs output word count for diverse geographic location



(c) ChatGPT User Latency (secs) for first 25 words vs. location (Orange line within the box plot is median and Green dashed line is average/mean)

Figure 3: Workload Characterization for ChatGPT

N. America	California, US; Texas, US; Iowa, US
Europe	London, UK; Frankfurt, DE;Ireland
Asia	Tokyo, Japan; Pune, India

Table 2: Compute region locations span three continents.

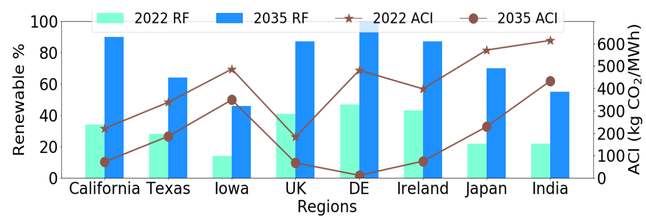


Figure 4: Regional Renewable Fraction (RF) and Average Carbon Intensity (ACI): 2022 vs. 2035.

Our study uses varied input prompts, captured from prompt engineering examples and video tutorials on Youtube. Figure 3a shows there is little correlation between prompt and output length, and because compute for the GPT models, computation cost is proportional to the number of output tokens (length). Thus, request cost or latency is difficult to predict from the prompt, making intelligent direction difficult.

3.2 Characterizing User Response Requirements

We characterize user response requirements by measuring ChatGPT response latency (request to full response) from global locations (Figure 3b), finding that response latency is weakly correlated with client location. Response latency is strongly correlated with output complexity (and thus weakly correlated with request size in Figure 3a). For long responses, users may read before output is complete, so we model the first 25 words latency, using the average latency per word. Latency varies by location, but the overall average latency is 3.14 seconds (Figure 3c). For reference, we show the suggested WWW page load latency for good interactive experience [29]. From the ChatGPT results, we conclude that response latency depends on output complexity and varies by location.

4 EVALUATION

4.1 Resource and Carbon Emissions Model

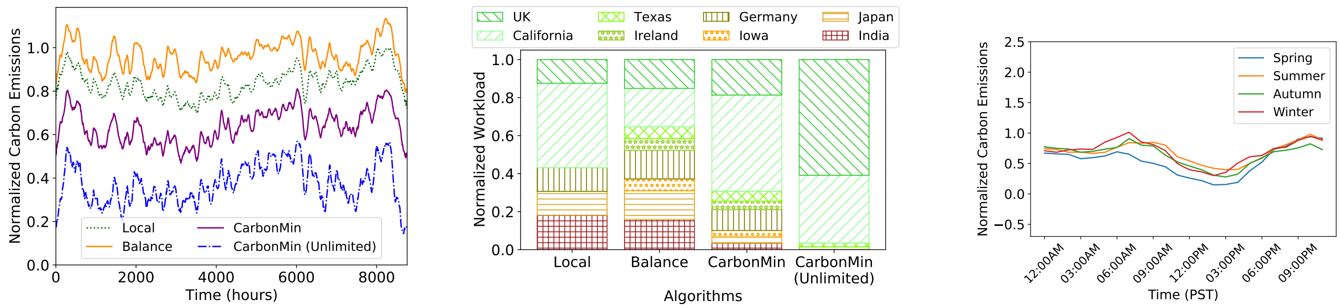
Generative AI serving is done from N cloud regions, in distinct geographies and powered by different power grids. We model for $N = 8$ locations [5] as in Table 2. Considering the average utilization of production (inflexible) workloads [19, 55, 61], we assume an average of 30% resource capacity at each datacenter, to be exploitable for serving ChatGPT inference (flexible load). This available resource capacity is denoted by $U_n(t)$ for region n. Shifting is limited by resource availability, so we consider headroom capacity, H_n, to increase shifting effectiveness[18].

For each compute region, we model the carbon emissions from inference serving $C_n(t)$ as the sum of operational carbon $C_n^{Op}(t)$ and embodied carbon $C_n^{Em}(t)$. The operational carbon emissions are calculated as the product of energy consumed $E_n(t)$ for serving the inference workload and the average carbon intensity $ACI_n(t)$. We use hourly Average Carbon Intensity (ACI) values from RiPiT [2] and Electricity Maps [36] (Figure 4).

We assume that each region uses Azure ND A100 v4-series instances for serving inference requests. Since amount of served data is small (a few hundreds of words) compared with the model size (billions of weights), we assume that the physical instances' GPU and CPU are the main sources of energy consumption. Therefore, we calculate the energy consumed for serving inference requests as

$$E_n(t) = I_n(t) \cdot f \cdot TDP \cdot PUE \quad (1)$$

where $I_n(t)$ is the number of requests processed in the compute region n. Our ChatGPT inference workload uses the diurnal model presented in Section 3 with an average latency of 21.7 seconds/request. Using Azure ND A100 v4-series [11], we model $TDP = 0.428$ kW per GPU (1/8 of 3.43 kW for the instance). Region power utilization efficiency (PUE) is 1.1 [6]. f, the computation (GPU-seconds) per



(a) Normalized Carbon Emissions vs. Algorithm (1-year)

(b) Load Distribution across regions (darker green reflects lower carbon intensity)

(c) Daily carbon variation for *CarbonMin*, comparing seasons. (Pacific time)

Figure 5: Evaluation Results (2022): *CarbonMin* and *CarbonMin (Unlimited)* consistently achieve lower Carbon emissions than the average annual carbon emission (a) by serving ChatGPT requests at low-carbon regions (b) which determine the daily emissions across seasons (c).

request is modeled conservatively as follows:

$$f = \frac{OI \cdot IW \cdot WC}{C} \quad (2)$$

where $OI = 0.35$ is TFLOPS per inference assuming GPT-3 model (around 175 billion weights) processed with BF16 operations. $IW = 5$ is the number of inferences per output word (assumed window/sampling of 5 for each output word), WC is the output word count (measured average of 185 output words/request), and $C = 156$ TFLOPS is the GPU capacity assuming 50% efficiency [1], and yields the average $f = 2.07$ GPU-sec/request.

The embodied emission is the total emissions of the Azure ND A100 v4-series instances apportioned over service time T share of the hardware overall lifetime LT (3 years) [26]:

$$C_n^{Em}(t) = \frac{T}{LT} E_{hw} = (U_n(t) + d \cdot H_n) \frac{avgRuntime \cdot I_n(t)}{LT} E_{hw} \quad (3)$$

where d is the fractional embodied emission of headroom's additional hardware and E_{hw} is per-GPU emission calculated as 1/8 of estimated per-instance emissions:

$$E_{hw} = \frac{1}{8}(PF + E_{GPU} + E_{CPU} + E_{DRAM} + E_{SSD} + E_{HDD}) \quad (4)$$

where PF is IC packaging Carbon footprint while E_{GPU}, E_{CPU}, E_{DRAM}, E_{SSD}, and E_{HDD} are GPU, CPU, memory, and storage emissions, respectively. We estimate these emissions based on previous reports [26] and instance hardware specifications [1, 3, 11], yielding $E_{hw} = 318$ $kgCO_2$ per GPU.

4.2 Request Direction Algorithms

When a compute region receives a user's request, it can be processed locally or sent to another compute region. The target region for shifting must respect the capacity constraint:

$$\forall_{t,n \le N} : \sum_{i \ne n} s_{in}(t) \cdot f \le U_n(t) + H_n \quad (5)$$

where $s_{in}(t)$ is number of requests shifted from region i to n.

Requests are directed based on varied optimization criteria. We evaluate three direction algorithms: (i) ***Local***: requests are processed at the region that they are received at; (ii) ***Balance***: requests are directed proportional to available region resources, producing equal utilization at all sites; and (iii) ***CarbonMin***: requests are directed to the region with lowest carbon intensity, minimizing carbon emissions. These three algorithms are subject to the capacity constraint. We also consider (iv) ***CarbonMin (Unlimited)*** which eliminates the capacity constraint. In all cases, the request latency is request computation time plus round trip network latency. We model network latency using the median round-trip network latency for each pair of Azure's regions [4].

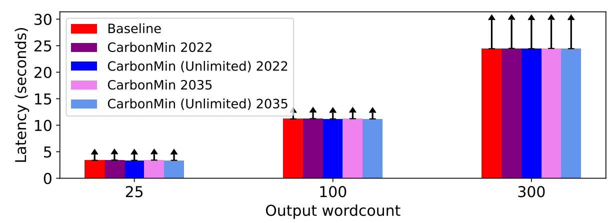
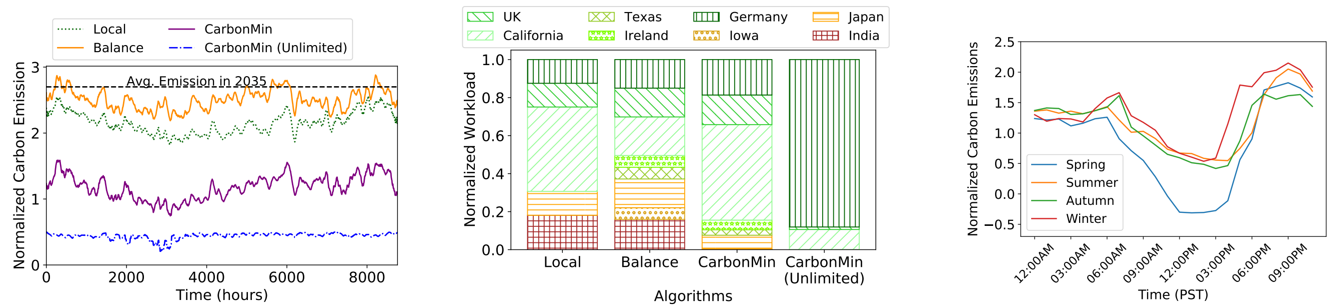


Figure 6: The QoS is maintained – Average latency of request direction algorithms at varying output wordcounts, extending upto the P90 latency, remains the same.

4.3 Results: Today

Figure 5a shows hourly carbon emissions of ChatGPT (i.e., the *ChatGPT-RR* workload) in 2022, normalized to the average carbon emission of all regions, without additional headroom (i.e., $H_n = 0$). While *Local* and *Balance* remain close to average carbon emissions, *CarbonMin* consistently reduces the emission by 35%. Eliminating the capacity constraint yields 63% carbon reduction.

Figure 5b shows the distribution of requests by service site. *Local* and *Balance* serve large fractions of requests at high-carbon sites. In contrast, *CarbonMin* directs a higher fraction of requests to lower emission locations (California, UK, Germany and Ireland). *CarbonMin (Unlimited)*, does even better by shifting most of load to a much greener location (UK). In Figure 5c, we consider the seasonal



(a) Normalized Carbon Emissions vs. Algorithm (1-year)

(b) Load Distribution across regions (darker green reflects lower carbon intensity)

(c) Daily carbon variation for *CarbonMin*, comparing seasons. (Pacific time)

Figure 7: Evaluation Results (2035): *CarbonMin* keeps Carbon emission increasing by only 20% compared to 2022 despite 55x growth in load (a) due to greener grids that make more low-carbon resources available for shifting (b) with greater variability (c).

variation in *CarbonMin* benefit over the course of a day. Carbon reduction is dominated by shifting to solar power generation in California with greater benefit midday, but with varying degree across seasons.

We analyze whether the request direction algorithms can maintain the expected inference QoS (previously characterized in Figure 3c). In Figure 6 (left), the average user-response latency of 25 words is only 1.46% larger than the baseline. The difference is even smaller for 100 word and 300 word outputs (0.44% and 0.20% respectively). These small increases reflect the average Azure region-region roundtrip latency for this workload distribution of 49 milliseconds. Average and P90 user-response latency for *CarbonMin* and *CarbonMin(Unlimited)* algorithms is essentially unchanged, now and in 2035.

4.4 Results: 2035

We evaluate the potential carbon impact of ChatGPT in the future (2035), focusing the overall outcome of grid decarbonization (more renewables so more low-carbon periods) and potential huge load growth. To model future grid ACI, we obtained detailed generation history, and scaled up wind and solar power generation to match public renewable fraction (RF) policy goals [22, 39, 42, 46] or where such was not available, we used a linear extrapolation [8, 9] (Figure 4). More formally, wind and solar generation are scaled by $(2035\ RF/2022\ RF)$ and non-renewable generation is scaled down by $[(1 - 2035\ RF)/(1 - 2022\ RF)]$, producing a 2–3x lower average carbon intensity (ACI) in 2035 for most regions. We project ChatGPT load based on today's Google search activity (i.e., *Google-RR*) as discussed in Section 3. Compute resources are scaled up to match the higher usage, and we model hardware energy efficiency improvements of 10x by 2035, an optimistic view of industry progress [30, 38].

Figure 7a presents ChatGPT carbon emissions in 2035, normalized to the annual global average in 2022. Increased renewable power and advanced computing technology produce a net 2.6x increase (dashed line), despite 55x load increase. With the boom in Generative AI [16], the situation is now far different than recent

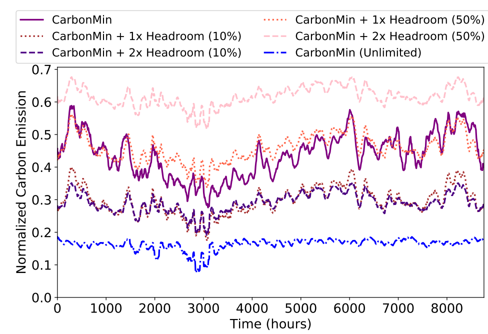


Figure 8: Annual Carbon Emissions normalized to 2035 using *CarbonMin* varying headroom capacity (CarbonMin + Nx Headroom (d%)); N is the increased capacity and d is the embodied emission factor)

reports [45]. In Figure 7b, we see how each algorithm distributes requests; *CarbonMin* effectively selects Germany, UK, California, and Ireland, reducing carbon emissions to 1.2x compared to 2022 levels. However, capacity constraints limit benefits. With unlimited capacity, load distribution changes drastically with 88% load in Germany, which aims 100% renewables by 2035.

Figure 7c shows seasonal variation in carbon emissions within a day, using the *CarbonMin* algorithm. The higher levels of renewable generation in 2035 cause greater variability. We see carbon emissions decrease earlier in the day, due to attractive European regions. Note that in Spring (blue line), ACI can actually be negative (when CAISO exports surplus solar generation to other grids).

We consider adding headroom with used computers [20] (see Figure 8), for several scenarios. Because they are used, the headroom computers can have lower embodied emissions (10%, 50%) vs. the primary resources. With 1x headroom emissions reduction increases from 56% to 71%. Further headroom (2x) gives little benefit.

The sensitivity to embodied emissions is illustrated by the 50% embodied case, where adding 1x headroom yields no benefit. Careful headroom design is needed to maximize load-shifting benefits.

In 2035, the grids are projected to have dramatically higher renewable fraction, lowering their overall annual Average Carbon Intensity by 2-3x. However, the *Google-RR* load posits a 55-fold increase. So, grid improvements, technology improvements contain the carbon emissions increase to 2.6-fold. Using CarbonMin can further reduce the increased carbon emissions to only 1.2x compared to 2022 levels. In short, carbon optimal request routing algorithms can be an important way to reduce emissions.

5 SUMMARY AND FUTURE

We have estimated the carbon cost of serving a generative AI model, showing that its emissions can be reduced with intelligent request direction algorithms, tied to power grid carbon information. More importantly, this optimization is possible with user-response latencies. In the future, the benefits of this approach are even greater.

Future research directions include broader characterization of generative AI workloads, new datacenter design for sustainability such as 100% power supply from renewables [7, 59] and adding headroom capacity [18], and updated studies as the growth structure of generative AI and power grid decarbonization develops.

ACKNOWLEDGMENT

We thank the anonymous reviewers for their insightful reviews. This work is supported in part by NSF Grants CMMI-1832230, OAC-2019506, CNS-1901466, and the VMware University Research Fund. We also thank Pierre Segonne from Electricity Maps for providing power grid data, and the Large-scale Sustainable Systems Group members for their support of this work!

REFERENCES

[1] 2021. Nvidia A100 datasheet. https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-us-nvidia-1758950-r4-web.pdf.

[2] 2021. Right Place, Right Time (RiPiT) Carbon Emissions Service. http://ripit.uchicago.edu. University of Chicago.

[3] 2023. AMD EPYC 7642 Specifications. https://www.techpowerup.com/cpu-specs/epyc-7642.c2247.

[4] 2023. Azure network round-trip latency statistics. https://learn.microsoft.com/en-us/azure/networking/azure-network-latency.

[5] 2023. Choose the right azure region for you: Microsoft Azure. https://azure.microsoft.com/en-us/explore/global-infrastructure/geographies/#geographies.

[6] 2023. Efficiency – Datacenters – Google. https://www.google.com/about/datacenters/efficiency/.

[7] Bilge Acun, Benjamin Lee, Fiodar Kazhamiaka, Kiwan Maeng, Manoj Chakkaravarthy, Udit Gupta, David Brooks, and Carole-Jean Wu. 2022. Carbon explorer: A holistic approach for designing carbon aware datacenters. *arXiv preprint arXiv:2201.10036* (2022).

[8] U.S. Energy Information Administration. 2022. Short-term Energy Outlook Data Browser. Retrieved September 15, 2022 from https://www.eia.gov/outlooks/steo/data/browser/

[9] International Energy Agency. 2023. International Energy Agency. "https://www.iea.org/.

[10] Alibaba. 2023. Tongyi Qianwen. https://tongyi.aliyun.com/.

[11] Microsoft Azure. 2023. ND A100 v4-series. "https://learn.microsoft.com/en-us/azure/virtual-machines/nda100-v4-series.

[12] Baidu. 2023. Wenxin Yiyan. https://yiyan.baidu.com/welcome.

[13] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[15] Andrew A Chien. 2018. *Characterizing opportunity power in the California independent system operator (CAISO) in years 2015-2017.* Technical Report. TR-2018-07.

[16] Andrew A. Chien. 2023. GenAI: Giga$$$, TeraWattHours and GigaTons of CO2. *Communications of the Association for Computing Machinery* (August 2023).

[17] Andrew A Chien, Fan Yang, and Chaojie Zhang. 2018. Characterizing curtailed and uneconomic renewable power in the mid-continent independent system operator. *AIMS Energy* 6, 2 (2018).

[18] Andrew A Chien, Chaojie Zhang, Liuzixuan Lin, and Varsha Rao. 2022. Beyond PUE: Flexible Datacenters Empowering the Cloud to Decarbonize. *USENIX Hot Carbon Workshop* (2022).

[19] Eli Cortez, Anand Bonde, Alexandre Muzio, Mark Russinovich, Marcus Fontoura, and Ricardo Bianchini. 2017. Resource Central: Understanding and Predicting Workloads for Improved Resource Management in Large Cloud Platforms. In *Proceedings of the International Symposium on Operating Systems Principles (SOSP)* (proceedings of the international symposium on operating systems principles (sosp) ed.). https://www.microsoft.com/en-us/research/publication/resource-central-understanding-predicting-workloads-improved-resource-management-large-cloud-platforms/

[20] Mark Dietrich, Robert Gardner, Andrew Chien, and Hakizumwami Birali Runesha. 2020. How to Extend the Productive Lifetime of Scientific Computing Equipment. Retrieved from http://people.cs.uchicago.edu/~aachien/lssg/research/zccloud/lifetime/bof-index.html.

[21] Jesse Dodge, Taylor Prewitt, Remi Tachet des Combes, Erika Odmark, Roy Schwartz, Emma Strubell, Alexandra Sasha Luccioni, Noah A Smith, Nicole DeCario, and Will Buchanan. 2022. Measuring the carbon intensity of ai in cloud instances. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1877–1894.

[22] William Driscoll. 2022. California law would target 90% renewable and zero-carbon electricity by 2035. https://pv-magazine-usa.com/2022/09/06/california-law-would-target-90-renewable-and-zero-carbon-electricity-by-2035/.

[23] Fabio Duarte. 2023. Number of ChatGPT Users. https://explodingtopics.com/blog/chatgpt-users.

[24] Google. 2023. Bard. https://bard.google.com/.

[25] Google. 2023. The next generation of AI for developers and Google Workspace. https://blog.google/technology/ai/ai-developers-google-cloud-workspace/.

[26] Udit Gupta, Mariam Elgamal, Gage Hills, Gu-Yeon Wei, Hsien-Hsin S. Lee, David Brooks, and Carole-Jean Wu. 2022. ACT: Designing Sustainable Computer Systems with an Architectural Carbon Modeling Tool. In *Proceedings of the 49th Annual International Symposium on Computer Architecture* (New York, New York) *(ISCA '22)*. Association for Computing Machinery, New York, NY, USA, 784–799. https://doi.org/10.1145/3470496.3527408

[27] V. Gupta, P. Shenoy, and R. K. Sitaraman. 2018. Efficient solar provisioning for net-zero Internet-scale distributed networks. In *2018 10th International Conference on Communication Systems Networks (COMSNETS)*. 372–379.

[28] Yuki Hiruta. 2019. Assessing climate sensitivity of hourly electricity demand in Japan. https://www-iam.nies.go.jp/aim/aim_workshop/aimws_25/presentation/4-04_Hiruta.pdf. (2019).

[29] Damien Jordan. 2018. Response time standards. https://www.websitepulse.com/blog/response-time-standards.

[30] Jonathan Koomey and Samuel Naffziger. 2016. Energy efficiency of computing: What's next? *In Electronic Design. November 28* (2016). http://electronicdesign.com/microprocessors/energy-efficiency-computing-what-s-next

[31] Kien Le, Ricardo Bianchini, Thu D. Nguyen, Ozlem Bilgir, and Margaret Martonosi. 2010. Capping the Brown Energy Consumption of Internet Services at Low Cost. In *Proceedings of the International Conference on Green Computing (GREENCOMP '10)*. IEEE Computer Society, USA, 3–14. https://doi.org/10.1109/GREENCOMP.2010.5598305

[32] Chuan Li. 2020. OpenAI's GPT-3 Language Model: A Technical Overview. https://lambdalabs.com/blog/demystifying-gpt-3.

[33] Liuzixuan Lin and Andrew A Chien. 2020. Characterizing stranded power in the ERCOT in years 2012-2019: A preliminary report. *University of Chicago CS Tech Report* (2020).

[34] Z. Liu, M. Lin, A. Wierman, S. Low, and L. L. H. Andrew. 2015. Greening Geographical Load Balancing. *IEEE/ACM Transactions on Networking* 23, 2 (2015), 657–671.

[35] Sasha Luccioni. 2023. The mounting human and environmental costs of generative AI. https://arstechnica.com/gadgets/2023/04/generative-ai-is-cool-but-lets-not-forget-its-human-and-environmental-costs/.

[36] Electricity Maps. 2023. Electricity Maps. https://app.electricitymaps.com/map.

[37] Microsoft. 2023. Microsoft AI. https://www.microsoft.com/en-us/ai.

[38] Christopher Mims. 2020. Huang's Law Is the New Moore's Law, and Explains Why Nvidia Wants Arm. https://www.wsj.com/articles/huangs-law-is-the-new-moores-law-and-explains-why-nvidia-wants-arm-11600488001.

[39] MISO. 2021. MISO Futures Report. https://cdn.misoenergy.org/MISO%20Futures%20Report538224.pdf.

[40] Hai Nguyen and Andrew Chien. 2023. Storm-RTS: Stream Processing with Stable Performance for Multi-cloud and Cloud-edge. In *2023 IEEE 16th International Conference on Cloud Computing (CLOUD)*.

[41] Hai Duc Nguyen, Chaojie Zhang, Zhujun Xiao, and Andrew A. Chien. 2019. Real-Time Serverless: Enabling Application Performance Guarantees. In *Proceedings of the 5th International Workshop on Serverless Computing* (Davis, CA, USA) *(WOSC '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3366623.3368133

[42] Ministry of New and Renewable Energy. 2022. Renewable Energy in India. https://pib.gov.in/FeaturesDeatils.aspx?NoteId=151141&ModuleId%20=%202.

[43] OpenAI. 2022. ChatGPT. https://chat.openai.com/.

[44] OpenAI. 2022. DALL-E 2. https://openai.com/product/dall-e-2.

[45] David Patterson, Joseph Gonzalez, Urs Hölzle, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David R So, Maud Texier, and Jeff Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink. *Computer* 55, 7 (2022), 18–28.

[46] Reuters. 2022. Germany aims to get 100% of energy from renewable sources by 2035. https://www.reuters.com/business/sustainable-business/germany-aims-get-100-energy-renewable-sources-by-2035-2022-02-28/.

[47] Emma Roth. 2023. Mark Zuckerberg says Meta now has a team building AI tools and 'personas'. https://www.theverge.com/2023/2/27/23617477/mark-zuckerberg-meta-ai-tools-personas.

[48] Josh Saul and Dina Bass. 2023. Artificial Intelligence Is Booming—So Is Its Carbon Footprint. https://www.bloomberg.com/news/articles/2023-03-09/how-much-energy-do-ai-and-chatgpt-use-no-one-knows-for-sure#xj4y7vzkg.

[49] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12 (nov 2020), 54–63. https://doi.org/10.1145/3381831

[50] Similarweb. 2023. Similarweb. https://www.similarweb.com/website/chat.openai.com.

[51] Similarweb. 2023. Similarweb. https://www.similarweb.com/website/google.com/#geography.

[52] Statista. 2023. Statista. https://www.dailymail.co.uk/sciencetech/article-3042230/Sleeping-habits-world-revealed-wakes-grumpy-China-best-quality-shut-eye-South-Africa-wakes-earliest.html.

[53] Chris Stokel-Walker. 2023. The Generative AI Race Has a Dirty Secret. https://www.wired.com/story/the-generative-ai-search-race-has-a-dirty-secret/.

[54] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3645–3650.

[55] Muhammad Tirmazi, Adam Barker, Nan Deng, Md E Haque, Zhijing Gene Qin, Steven Hand, Mor Harchol-Balter, and John Wilkes. 2020. Borg: the next generation. In *Proceedings of the fifteenth European conference on computer systems*. 1–14.

[56] Jonathan Vanian and Kif Leswing. 2023. ChatGPT and generative AI are booming, but the costs can be extraordinary. https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html.

[57] WattTime. 2023. WattTime. https://www.watttime.org.

[58] Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, et al. 2022. Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems* 4 (2022), 795–813.

[59] Fan Yang and Andrew A. Chien. 2017. Large-scale and Extreme-Scale Computing with Stranded Green Power: Opportunities and Costs. *IEEE Transactions on Parallel and Distributed Systems* 29, 5 (December 2017).

[60] Chaojie Zhang. 2023. *Eliminating the Capacity Variation Penalty for Cloud Resource Management.* Ph. D. Dissertation. The University of Chicago.

[61] Chaojie Zhang and Andrew A. Chien. 2021. Scheduling Challenges for Variable Capacity Resources. In *Job Scheduling Strategies for Parallel Processing: 24th International Workshop, JSSPP 2021, Virtual Event, May 21, 2021, Revised Selected Papers*. Springer-Verlag, Berlin, Heidelberg, 190–209. https://doi.org/10.1007/978-3-030-88224-2_10

[62] Jiajia Zheng, Andrew A. Chien, and Sangwon Suh. 2020. Mitigating Curtailment and Carbon Emissions through Load Migration between Data Centers. *Joule* (October 2020). https://doi.org/10.1016/j.joule.2020.08.001

[63] Zhi Zhou, Fangming Liu, Ruolan Zou, Jiangchuan Liu, Hong Xu, and Hai Jin. 2016. Carbon-Aware Online Control of Geo-Distributed Cloud Services. *IEEE Transactions on Parallel and Distributed Systems* 27, 9 (2016), 2506–2519.