Investigating Power Consumption Flexibility of AI Data Centers for Demand Response Participation

Fatih Acun
Electrical and Computer Engineering
Boston University
Boston, MA, USA
acun@bu.edu

Can Hankendi
Electrical and Computer Engineering
Boston University
Boston, MA, USA
hankendi@bu.edu

Ethan Levine
Electrical and Computer Engineering
Boston University
Boston, MA, USA
elevine@bu.edu

Hudson Reynolds
Electrical and Computer Engineering
Boston University
Boston, MA, USA
hudsonre@bu.edu

Joshua Bardwick
Electrical and Computer Engineering
Boston University
Boston, MA, USA
jbardwic@bu.edu

Ayse K. Coskun
Electrical and Computer Engineering
Boston University
Boston, MA, USA
acoskun@bu.edu

Abstract

The recent advances in artificial intelligence (AI) have driven a sharp rise in data center electricity consumption, raising concerns about power grid reliability. To mitigate the stress on power systems, Independent System Operators (ISOs) use demand-side flexibility through demand response (DR) programs. Modern data centers, using mechanisms like hardware power capping, can dynamically adjust their power usage and serve as flexible grid resources. In this study, we assess the potential of AI data centers to participate in DR programs while meeting the quality-of-service (QoS) requirements of their workloads. By analyzing the distinct characteristics of training and inference workloads in real-world DR programs, our findings show that AI data centers can offer between 18% and 55% flexibility relative to their average power consumption, highlighting their potential for supporting a more sustainable power grid.

CCS Concepts

• **Social and professional topics** → **Sustainability**; • **Hardware** → **Enterprise level and data centers power issues**.

Keywords

Data Center Sustainability, Power Management, Demand Response

ACM Reference Format:
Fatih Acun, Can Hankendi, Ethan Levine, Hudson Reynolds, Joshua Bardwick, and Ayse K. Coskun. 2026. Investigating Power Consumption Flexibility of AI Data Centers for Demand Response Participation. In *The 17th ACM International Conference on Future and Sustainable Energy Systems (E-Energy '26), June 22–25, 2026, Banff, AB, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3744255.3798112

1 Introduction

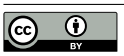
Data centers form the critical infrastructure powering ever-growing artificial intelligence (AI) workloads, including serving and training massive Large Language Models (LLMs). Driven by the energy-intensive nature of LLM training and inference workloads that utilize Graphics Processing Units (GPUs), the power demand of data centers continues to rise sharply. Recent estimates suggest U.S. AI-related data center capacity is reaching 5 GW by 2025 and exceeding 50 GW by 2030 [4]. This rapid growth further threatens grid stability, as the scale and power consumption patterns of data centers place exceptional demands on power grids [10].

To maintain safe and reliable grid operation, ISOs utilize demand-side flexibility by calling on consumers to adjust usage through demand response (DR) programs targeting different events and customer types [1]. Depending on the objective of the DR program, participants may be required to reduce their power consumption during peak demand periods or increase their usage when additional load is needed to absorb excess power generation. Some examples of DR programs include dynamic pricing [18], regulation service reserves (RSR) [13], and emergency demand response (EDR) [2, 11].

Data centers are strong candidates for DR since they consume large amounts of power and can rapidly adjust demand using (1) hardware power capping and (2) power-aware scheduling that shifts workloads to meet power objectives. Prior research investigates data center DR participation by using workload shifting in conjunction with local generation to adjust the power consumption [9]. QoS-aware DR participation methods address performance constraints of data center jobs by simultaneously following the power targets of DR programs [3, 19]. Although data center DR is gaining attention, most studies focus on CPU-based resources; integrating GPU-based AI workloads into DR programs remains underexplored.

In this paper, we assess how effectively AI data centers can participate in DR programs by combining real hardware power-performance profiles with a scalable data center DR simulation framework. Our key contributions are as follows:

- Demonstrating the flexibility of AI data centers to participate in EDR and RSR programs while meeting the QoS requirements of AI workloads.
- Examining the varying implications of DR program power constraints on AI inference and training jobs.
- Characterizing the power–performance behavior of popular AI workloads on GPU systems.



E-Energy '26, Banff, AB, Canada

ACM ISBN 979-8-4007-2011-6/2026/06
https://doi.org/10.1145/3744255.3798112

We evaluate data center DR participation using a diverse mix of AI inference and training workloads across varying utilization levels. Our results show AI data centers can provide substantial power consumption flexibility in RSR, ranging from 18% to 55% of their average power consumption, and up to 20% power reduction in EDR programs.

2 Background in Demand Response Programs

Regulation Service Reserves. In RSR programs, ISOs maintain grid balance by adjusting power supply and demand at a high resolution [13]. This coordination is achieved through regulation signals broadcast to participating demand-side resources. PJM provides two types of regulation signals, RegA (traditional) and RegD (dynamic), each updated every 2 seconds. RegA exhibits slower variations, whereas RegD features faster changes [14]. Load types that are capable of rapid and accurate power adjustments are best suited for tracking the RegD signal, while those with more limited ramping agility can still reliably follow the smoother RegA signal. To join RSR programs, participants need to provide their forecasts (e.g., in hour-ahead markets) for average power consumption $\bar{P}$ and reserve capacity R to the ISO prior to DR execution. The reserve capacity R specifies the magnitude by which a participant commits to modulate its power consumption above or below its average power $\bar{P}$. During real-time operation, participants must adhere to a power target, P_{target}, calculated as:

$$P_{\text{target}}(t) = \bar{P} + y(t)R \tag{1}$$

where $y(t) \in [-1, 1]$ denotes the regulation signal. To assess tracking accuracy, we calculate the tracking error $\epsilon(t)$ as the absolute deviation between P_{target} and the actual power consumption $P(t)$ at time t, normalized by the R as follows:

$$\epsilon(t) = |P(t) - P_{\text{target}}(t)| \;/\; R\,. \tag{2}$$

RSR participants must track the regulation signal closely by keeping their tracking errors within acceptable thresholds. In this work, we define the constraint for power tracking probabilistically as follows:

$$\text{Prob}[\epsilon(t) > 0.3] < 10\%, \tag{3}$$

which implies tracking error must remain below 0.3 for at least 90% of the time. Finally, the monetary cost of energy is calculated such that the participants are incentivized by the amount of provided R, and penalized by the average tracking error, $\bar{\epsilon}$, as follows:

$$M^{\text{RSR}} = \left(\Pi^P \bar{P} - \Pi^R R + \Pi^\epsilon R\bar{\epsilon}\right) \times T, \tag{4}$$

where T is the RSR duration, Π^P, Π^R, and Π^ϵ are price coefficients.

Emergency Demand Response. EDR programs aim to maintain grid reliability during periods when demand approaches or exceeds available supply or transmission capacity [2, 11]. In EDR, participants are required to curtail their power consumption relative to a baseline level, P_{base}, and are compensated based on the amount of the reduction they provide [11]. For data centers, we calculate P_{base} as the average power consumption during normal operation without any power constraints. The monetary cost of energy during an EDR event is then given by:

$$M^{\text{EDR}} = \left(\Pi^P \bar{P} - \Pi^I (P_{\text{base}} - \bar{P})\right) \times T, \tag{5}$$

where $\bar{P}$ denotes the average power consumption of the data center during EDR, Π^P and Π^I are price coefficients, T is the duration.

3 Modeling Demand Response Participation by AI Data Centers

To analyze how AI data centers participate in DR programs while satisfying workload performance requirements, we develop a simulation and optimization approach that integrates DR program specifications with workload-level power–performance characteristics. Our approach unifies calculating monetary costs from DR participation with QoS constraints of workloads, enabling a comprehensive representation of both economic incentives and operational limitations of data center DR participation.

3.1 A Generalized Cost Formulation for Data Center Demand Response

DR participation introduces multiple, potentially competing objectives for data centers, including following grid rules and minimizing energy-related monetary costs while limiting performance degradation of running jobs. To capture these trade-offs, we adopt and extend the cost model introduced in Adaptive Policy with QoS Assurance (AQA) [19]. While AQA focuses on RSR participation, we broaden the formulation to account for both EDR and RSR programs. Our generalized cost function is defined as:

$$C \quad = \quad M + \beta \sum_j \text{SoftPlus}\left(\rho\left(\Pr[Q^j > Q^j_{\text{th}}] - \delta^j\right)\right), \tag{6}$$

$$\text{s.t.} \qquad \text{Prob}[Q^j > Q^j_{\text{th}}] \le \delta^j,\ \forall j \in [1, 2, \ldots, J], \tag{7}$$

where M refers to the monetary cost of the EDR or RSR program, Q^j and Q^j_{th} are the QoS degradation and thresholds for job type j, and J denotes the total number of job types. $\Pr[Q^j - Q^j_{\text{th}}]$ refers to the probability of QoS degradations exceeding the thresholds, and δ^j defines the constraint for the probability of QoS violations. Softplus is the activation function, $\ln(1 + e^x)$, that amplifies QoS constraint violations. We define the Q^j as a relative metric as follows:

$$Q^j = (T^j_{\text{so}} - T^j_{\text{min}}) \;/\; T^j_{\text{min}}, \tag{8}$$

where T^j_{min} shows the minimum execution time of job type j without any queueing time or enforced power caps, T^j_{so} is the sojourn time, representing the total time for queuing and execution.

AQA provides a gradient descent optimization approach to minimize the objective function. However, its gradient calculations depend on analytical estimations limited to RSR participation and include various empirical parameters that are sensitive to changes in the workload characteristics (e.g., power consumption and QoS thresholds). To avoid this sensitivity and broaden applicability, we employ simulated annealing (SA) [8], a simple but powerful optimization technique that eliminates the need for gradient information. SA explores the parameter space via randomized perturbations, while its temperature-driven cooling schedule (where temperature and cooling refer to internal parameters of an SA solver) gradually shifts the search from exploration to exploitation. With a properly designed cooling schedule, SA is theoretically guaranteed to converge to the global optimum [5].

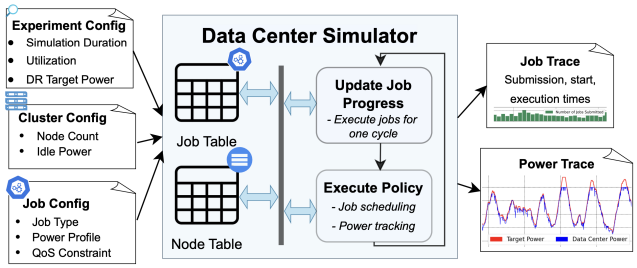


Figure 1: An illustration of our simulation-based approach for data center DR participation.

3.2 A Lightweight Data Center Simulator

We design and open-source a lightweight simulator, FlexDC-Sim, for modeling and evaluating the runtime control of data centers during DR participation.[1] Figure 1 presents the overall architecture of our simulator. At a high level, FlexDC-Sim models end-to-end data center operation, including job scheduling, job execution progress, and power management, while seamlessly integrating DR program requirements by converting them into explicit, time-varying power targets. To track these targets, the simulator applies mechanisms such as power capping and server idling or activation, and it emulates job power consumption and completion progress by using power–performance profiles measured on real hardware, enabling accurate reproduction of execution behavior under power constraints. The framework is generalizable and configurable such that users can specify the data center size, hardware characteristics (e.g., idle power), and power–performance profiles for arbitrary jobs. As long as the power–performance profiles are provided, the simulator supports CPU-based clusters as well as GPU systems.

Internally, FlexDC-Sim maintains two core data structures, a job table and a node table. The job table records job traces and QoS-related metrics, while the node table stores job execution progress on nodes and power consumption throughout the simulation. This design enables comprehensive reporting of performance and power consumption at the job, node, and cluster levels. We refer to the algorithmic procedure responsible for adjusting power consumption to follow DR power targets as the *runtime policy*. Although our simulator supports arbitrary runtime policy implementations, in this study, we adopt the policy from AQA [19]. AQA's runtime policy tracks the target power by first estimating the number of servers that should remain active or be idled, and then applying power caps to further reduce consumption. The power-capping methodology reduces the power of all running job types by the same percentage compared to their power consumption range and ensures that the cluster power aligns with the DR power target.

3.3 Job Types and Workload Traces

In our experimental setup, we consider AI training and inference jobs across computer vision and LLM applications to capture the diverse characteristics of modern data centers. For inference, we use batch-style jobs, which provide stable and repeatable measurements of power and performance over manageable execution windows. For training workloads, which can run for many hours, we profile shorter segments that are sufficient to capture their representative power–performance behavior without incurring prohibitive runtime overhead. To obtain power-performance properties, we perform an offline profiling study in which each application is executed under a range of power caps on Chameleon Cloud [7], using nodes equipped with four NVIDIA P100 GPUs. Chameleon's bare-metal access enables direct GPU power capping via the `nvidia-smi` power-limit interface [12], which typically requires admin privileges by default. Table 1 summarizes our power-performance profiling results, showing the minimum and maximum execution times of each job type when run under minimum and maximum power, $p_{\min}$ and $p_{\max}$. Using the profiled applications, we construct five workload traces with varying QoS thresholds and target data center utilizations, as listed in Table 2. We define utilization as the fraction of nodes actively running jobs, and job arrivals in each trace follow a Poisson process. To achieve a desired target utilization, we set the arrival rate for each job type j as follows:

$$\lambda_j = \eta\, N \,/\, (T^j_{\min} s^j J), \tag{9}$$

where η refers to the utilization, N is the total number of servers, and J is the total number of job types. s^j denotes the job size, representing the number of nodes allocated by each job submission of type j. To achieve the desired system utilization, the generator computes the arrival rate λ_j for each job type and produces the corresponding workload trace.

[1] GitHub repository: https://github.com/peaclab/flexdc-sim.

Table 1: Power-performance profilings of AI jobs.

Category	Job Type	$T_{\min}$(s)	$T_{\max}$(s)	$p_{\min}$(W)	$p_{\max}$(W)
Inference	Resnet-50 [6]	25	190	199	618
	GPT-2 [15]	28	180	198	732
	Llama-7B [17]	36	360	202	702
	Bloom-560M [16]	44	386	202	698
Training	Resnet-50	337	1321	199	755
	GPT-2	1185	4592	201	763
	Llama-7B	1242	4921	200	742
	Bloom-560M	1104	4202	202	726

Table 2: Inference and training workload traces with varying utilization, QoS thresholds, and GPU counts.

			QoS Threshold (Q^j_{th})			
Trace	Utilization	# GPUs	Resnet	GPT-2	Llama	Bloom
$W_{\text{inf-LU}}$	60%	4	5	4.5	4	3.5
$W_{\text{inf-HU}}$	80%	4	5	5	5	5
$W_{\text{train-LU}}$	60%	4	3	3	3	3
$W_{\text{train-HU}}$	80%	4	4	4	4	4
$W_{\text{train-multi}}$	70%	4/8*	3	4	4	3

*Resnet/Llama jobs use 4 GPUs; GPT-2 and Bloom use 8 GPUs.

4 Results

By including a wide range of experiments with varying workload properties, we evaluate data center participation in both RSR and EDR programs. Our results characterize the optimal power bids that quantify the flexibility of data centers, the monetary savings enabled by this flexibility, and the power tracking under DR participation.

Power Bids. We perform optimizations for data centers with 1,000 servers using each workload trace in Table 2. For RSR, we evaluate

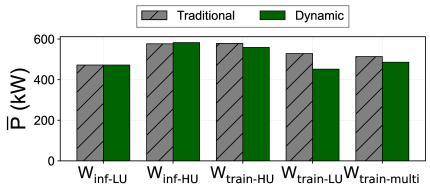


Figure 2: $\bar{P}$ bids for RSR program.

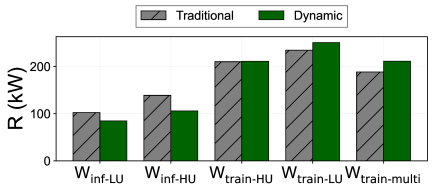


Figure 3: R bids for RSR program.

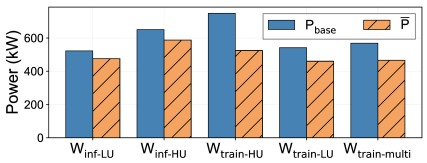


Figure 4: Power reduction in EDR.

participation under both the traditional and dynamic regulation signals, and the optimizer selects bids for average power $\bar{P}$ and reserve capacity R. Figures 2 and 3 show the RSR bids for all workload traces. The $\bar{P}$ bids reveal that average power requirements are driven primarily by data center utilization, such that higher utilization results in higher $\bar{P}$. Notably, $\bar{P}$ is largely insensitive to whether the workloads are inference or training. Likewise, the choice between traditional and dynamic RSR signals has only a marginal effect on average power requirements. For the R bids, we observe that train workloads offer substantially greater flexibility than inference workloads. We relate this to their longer and more malleable execution structure of training jobs. Although training workloads may appear to have stricter QoS constraints in Table 2, they effectively have more relaxed deadlines under our relative execution-time–based QoS definition in Eq. (8). For EDR, the optimizer determines only the average power bid $\bar{P}$, which directly reflects the achievable reduction from the baseline consumption P_{base}. Figure 4 compares P_{base} and $\bar{P}$ across all workloads. Similar to our observations for $\bar{P}$ in RSR, data centers with higher utilizations must commit to higher $\bar{P}$ values during EDR participation. Across all experiments in EDR and RSR, QoS requirements defined at 90^{th} percentile for each job type are satisfied with the provided bids.

Energy Cost Savings. Either by regulating their power consumption in RSR or providing reductions in EDR, data centers receive financial incentives from ISOs, yielding notable energy cost reductions. Figure 5 reports the percentage savings across all workload traces relative to the NoDR baseline. NoDR energy costs are calculated based on a data center's energy cost while executing without any power constraints. For the EDR program, cost savings range from 6% to 34%, reflecting the extent of power reduction from P_{base}, as previously shown in Figure 4. In the RSR program, savings are substantially higher, spanning 17% to 54%. Across both RSR signal types, training workloads consistently achieve greater cost reductions, driven by their large R bids.

Power Tracking. Power tracking is achieved using two control knobs, applied in order: (1) adjusting the number of active servers, and (2) applying power caps to running jobs as discussed in Section 3.2. Our results show that the effectiveness and usage of these knobs depend strongly on the workload characteristics. Figure 6 illustrates the power-tracking behavior for training and inference

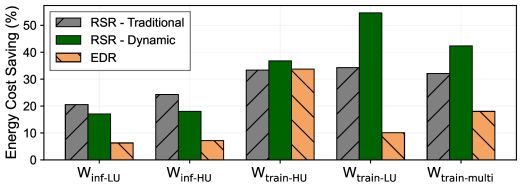


Figure 5: Energy cost savings with respect to NoDR costs.

data centers participating in the RSR program with a traditional signal for one hour. In the training data center, the long-running nature of training jobs prevents frequent scheduling turnover. Since preemption is not supported, power capping becomes the primary mechanism to reduce power. As seen in the first 500 seconds of Figure 6a, the number of active nodes stays steady, but power consumption reduces due to power caps. In contrast, the inference data center executes short jobs, enabling frequent scheduling decisions, and adjusting the number of active servers becomes the main control knob for tracking the power target as shown in Figure 6b.

5 Conclusion

The rapid expansion of generative AI has accelerated the build-out of large-scale data center infrastructure, making data centers one of the most substantial and concentrated loads on modern power grids. To overcome the emerging bottleneck of power availability for AI and to enable the sustainable growth of both computing infrastructure and power grids, AI data centers must be designed with flexibility as a core operational principle. In this work, we investigate the participation of AI data centers in two real-world DR programs and show that AI workloads can provide significant demand-side flexibility, between 18% and 55% of their average power consumption, while still meeting their QoS requirements during DR participation.

Acknowledgments

This work has been partially funded by the National Science Foundation DESC program under Award No. 2450111. We thank Dr. Daniel Wilson for his contributions to FlexDC-Sim's software architecture.

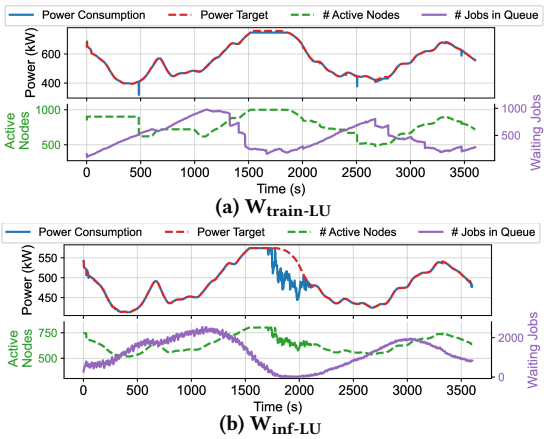


Figure 6: Power tracking of AI training and inference data centers for RSR participation with traditional signal type.

References

[1] International Energy Agency. 2023. Demand Response. Web page. https://www.iea.org/energy-system/energy-efficiency-and-demand/demand-response Accessed 2025-07-11.

[2] California Public Utilities Commission. 2024. Emergency Load Reduction Program. https://www.cpuc.ca.gov/industries-and-topics/electrical-energy/electric-costs/demand-response-dr/emergency-load-reduction-program. Accessed: 2024-06-29.

[3] Hao Chen, Yijia Zhang, Michael C. Caramanis, and Ayse K. Coskun. 2019. EnergyQARE: QoS-Aware Data Center Participation in Smart Grid Regulation Service Reserve Provision. *ACM Trans. Model. Perform. Eval. Comput. Syst.* 4, 1, Article 2 (Jan. 2019), 31 pages. doi:10.1145/3243172

[4] Electric Power Research Institute (EPRI). 2025. Scaling Intelligence: The Exponential Growth of AI's Power Needs. https://www.epri.com/research/products/000000003002033669

[5] V. Granville, M. Krivanek, and J.-P. Rasson. 1994. Simulated annealing: a proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16, 6 (1994), 652–656. doi:10.1109/34.295910

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, Las Vegas, NV, USA, 770–778. https://arxiv.org/abs/1512.03385

[7] Kate Keahey, Jason Anderson, Zhao Zhen, Pierre Riteau, Paul Ruth, Dan Stanzione, Mehmet Cevik, James Colleran, Haryadi S. Gunawi, Chris Hammock, Joe Mambretti, Andrew Barnes, Fred Halbach, Antonio Rocha, and Joe Stubbs. 2020. Lessons Learned from the Chameleon Testbed. In *Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20)*. USENIX Association. https://www.usenix.org/conference/atc20/presentation/keahey-lessons

[8] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220, 4598 (1983), 671–680. doi:10.1126/science.220.4598.671

[9] Zhenhua Liu, Adam Wierman, Yuan Chen, Benjamin Razon, and Niangjun Chen. 2013. Data center demand response: avoiding the coincident peak via workload shifting and local generation. *SIGMETRICS Perform. Eval. Rev.* 41, 1 (jun 2013), 341–342. doi:10.1145/2494232.2465740

[10] Tim McLaughlin. 2025. Big Tech's data center boom poses new risk to US grid operators. *Reuters* (2025). https://www.reuters.com/technology/big-techs-data-center-boom-poses-new-risk-us-grid-operators-2025-03-19/

[11] New York Independent System Operator (NYISO). 2023. *Emergency Demand Response Program (EDRP) Manual.* Technical Report. New York Independent System Operator. https://www.nyiso.com/documents/20142/2923301/edrp-mnl.pdf/8f34b039-de10-1726-bce4-0e02d8732a03?t=1747761514423 Accessed: 2025-08-27.

[12] NVIDIA Corporation. 2024. NVIDIA System Management Interface (nvidia-smi). https://developer.nvidia.com/nvidia-system-management-interface. Accessed: 2025-07-24.

[13] PJM Interconnection. 2022. PJM Manual 12: Balancing Operations, Revision 44. https://www.pjm.com/-/media/DotCom/documents/manuals/archive/m12/m12v44-balancing-operations-03-01-2022.pdf. Accessed: 2025-03-24.

[14] PJM Interconnection. 2025. Ancillary Services. https://www.pjm.com/markets-and-operations/ancillary-services.aspx. Accessed: 2025-08-27.

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog* 1, 8 (2019), 9. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[16] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100* (2022). https://arxiv.org/abs/2211.05100

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023). https://arxiv.org/abs/2302.13971

[18] Federal Energy Management Program U.S. Department of Energy. 2025. Demand Response and Time-Variable Pricing Programs. Web page. https://www.energy.gov/femp/demand-response-and-time-variable-pricing-programs Accessed 2025-11-19.

[19] Yijia Zhang, Daniel Curtis Wilson, Ioannis Ch. Paschalidis, and Ayse K. Coskun. 2022. HPC Data Center Participation in Demand Response: An Adaptive Policy With QoS Assurance. *IEEE Transactions on Sustainable Computing* 7, 1 (2022), 157–171. doi:10.1109/TSUSC.2021.3077254