EFFICIENCY VS DEMAND IN AI ELECTRICITY: IMPLICATIONS FOR POST-AGI SCALING

Doyi Kim, Jiseok Ahn, Haewon McJeon, Changick Kim*
Korea Advanced Institute of Science and Technology (KAIST)
{doyi.kim, jiseok.ahn, hmcjeon, changick}@kaist.ac.kr

ABSTRACT

As AI capabilities and deployment accelerate toward a post-AGI era, concerns are growing about electricity demand and carbon emissions from AI computing, yet AI is rarely represented explicitly in long-term energy, economy, and climate scenario models. In such a setting, digital infrastructure scaling may be constrained by power system dynamics. We introduce an AI computing sector into the Global Change Analysis Model (GCAM) and run U.S. scenarios that couple AI service growth with changes in compute energy intensity over time and economic drivers. We find that service growth does not translate linearly into electricity demand. Outcomes depend on efficiency trajectories and demand responsiveness. With sustained efficiency improvements, AI electricity demand remains moderated. With slower or saturating gains, demand driven by income growth dominates by the middle of the century. Sensitivity analyses show weak responsiveness to price signals but strong dependence on income growth, implying limited leverage from price-based mechanisms alone. Rather than offering a single forecast, we identify conditions under which efficiency-led versus demand-led regimes emerge, providing a compact template for scenarios of long-term AI electricity demand and their implications for emissions in the power sector.

1 INTRODUCTION

Artificial intelligence (AI) has rapidly transitioned from theoretics to a general-purpose technology embedded across economic sectors, driving rapid growth in large-scale data centers for training and inference. A recent report (IEA (2025)) shows that data centers accounted for approximately 1.5% of global electricity consumption in 2024, with usage growing at a 12% annual rate over the past five years, raising concerns about grid reliability, energy affordability, and carbon emissions.

At the same time, the AI stack has experienced rapid improvements in efficiency. Historical trends in GPU performance for leading AI accelerators suggest that computational efficiency has improved by roughly 1.3× per year (Pilz et al. (2025)), creating a countervailing dynamic in which growing AI service demand may be partially offset by declining energy per unit of compute. This tension has generated substantial uncertainty in projections of AI-driven electricity demand, with some studies warning of large increases in emissions (Aljbour et al. (2024); Brijs et al. (2025); Goldman Sachs Research (2024); Shehabi et al. (2024)) and others arguing demand may remain manageable in the near term (Nadel (2025); Koomey et al. (2025); Masanet et al. (2020)). Continued improvements in specialized accelerators and data-center operations further complicate long-run projections.

Building on these near-term perspectives, we extend the analysis to mid-century and identify the conditions, especially efficiency persistence and demand elasticities, under which projections shift between efficiency-dominant and demand-dominant regimes. A macro-scale energy–economy model makes this interaction explicit by propagating these assumptions through power-sector transitions and emissions. In a potential post-AGI setting where AI services become pervasive across the economy, these long-run sensitivities matter because power-system dynamics may become a first-order constraint on digital infrastructure scaling. Accordingly, we embed AI computing in GCAM and run mid-century scenarios to map regime thresholds for AI electricity demand.

*Corresponding author.

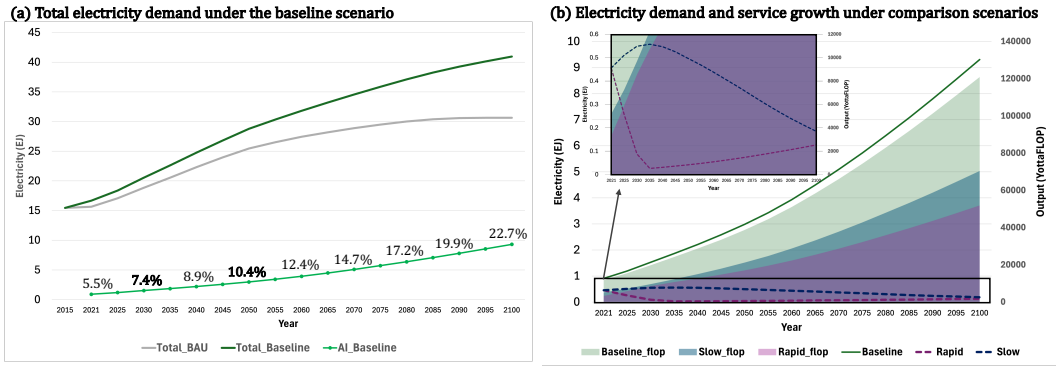

Figure 1: **AI-driven electricity demand under alternative efficiency assumptions.** (a) Total U.S. electricity demand under the business-as-usual(BAU) and baseline scenario that adds an AI data-center load with fixed compute energy intensity γ. (b) AI data-center electricity demand (left axis, solid lines) and AI service output (right axis, shaded area) under alternative efficiency trajectories (*Rapid* and *Slow*). Service output grows rapidly in all scenarios, but electricity demand diverges depending on the assumed pace of efficiency improvement.

2 Experimental Set up

We extend the Global Change Analysis Model (GCAM, Calvin et al. (2019)) by adding an explicit AI computing sector within the industrial module. GCAM is a partial-equilibrium, market-based macro-scale energy-economy model that links socioeconomic drivers, energy supply and demand, land use, and emissions. However, it does not explicitly represent AI-related computing demand, motivating a dedicated sectoral representation.

We measure AI service output as floating-point operations (FLOP), a proxy for computational workload (Pilz et al. (2025)). AI electricity demand is modeled as the product of service output $S(t)$ (FLOP) and a system-level compute energy intensity coefficient $\gamma(t)$ (Joule/FLOP)[1], such that

$$E(t) = S(t) \times \gamma(t),$$

where $\gamma(t)$ captures chip-level efficiency as well as data center overheads (cooling, networking, storage, and power usage effectiveness). In GCAM, $S(t)$ is determined endogenously from demand elasticities and macroeconomic drivers. We consider a baseline scenario with a fixed γ (7.7×10^{-5} EJ/YF) and two alternative efficiency trajectories: *Rapid* (doubling compute performance roughly every 2.34 years until 2035) and *Slow* (roughly 180% per decade). We also run sensitivity scenarios that vary price elasticity (-0.2, -0.7, -1.2) and income elasticity (1.6, 2.5, 3.5) around the baseline (Table 1). Full parameter and implementation details are provided in Appendix A.

Table 1: Scenarios used in this study

Scenario	BAU	Coefficient change			Sensitivity test (low/base/high)					
		Baseline	Rapid	Slow	PE			IE		
Coefficient (EJ/YF)*	–	7.7×10^{-5} (fixed)	Up to 2035, datacenter performance has doubled roughly every 2.34 years.	NVIDIA's chip performance has increased by roughly 180% per decade.	7.7×10^{-5}(fixed)					
Price-elasticity	–	-0.7			-0.2	-0.7	-1.2	-0.7		
Income-elasticity	–	2.5			2.5			1.6	2.5	3.5

* All scenarios were calculated based on the 2023 coefficient value of 3.85x10-5 EJ/YottaFlop. Although this coefficient was derived from the literature, we applied a scale factor of 2 across all scenarios to align the resulting energy input(EJ) with observed statistics

[1]We express $S(t)$ in YottaFLOP (YF, 1 YF $= 10^{24}$ FLOP), so $\gamma(t)$ is reported in EJ/YF (equivalently J/FLOP).

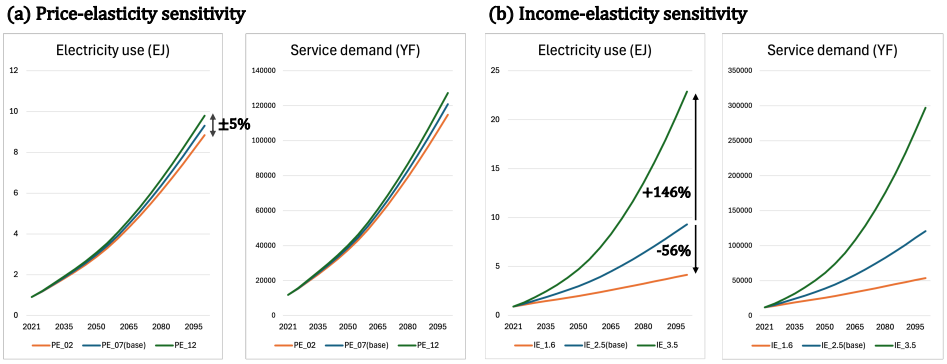

Figure 2: **Sensitivity of AI computing electricity demand to price and income elasticitie.** (a) Electricity demand and AI service output under alternative price elasticity assumptions (All other parameters are the same as baseline). (b) Electricity demand and AI service output under alternative income elasticity assumptions.

3 Results

3.1 Baseline Impacts in GCAM

Under the U.S. baseline scenario, annual electricity demand from AI data centers reaches approximately 1.5 EJ (about 420 TWh) by 2030 and 3 EJ (about 830 TWh) by 2050, accounting for roughly 10% of total U.S. electricity use in 2050 (broadly consistent with recent U.S. estimates (Brijs et al. (2025); Aljbour et al. (2024))). These near-term magnitudes are also comparable to recent short-horizon estimates (e.g., Kumar & Manning (2025)), providing an external anchor for our baseline calibration.

Incorporating time-varying efficiency, as in the *Rapid* and *Slow* efficiency scenarios, substantially moderates electricity demand relative to the baseline (Fig. 1(b)). The *Rapid* case drives demand to implausibly low levels and is therefore treated as an upper-bound stress test for regime boundaries, while the *Slow* case still reduces demand but by a more limited margin. Put simply, electricity demand reflects how much compute is used and how much energy each unit of compute requires; efficiency trajectories determine whether the latter can offset the former.

3.2 Demand Responsiveness of AI Computing

To quantify this, we conduct a set of sensitivity experiments to isolate the role of demand responsiveness in determining how rapidly AI service growth translates into electricity use. In Fig. 2(a), varying the price elasticity has a relatively modest effect on both electricity use and service output. Across the tested range, changes in electricity demand remain within ±5% relative to the baseline. This price inelasticity is consistent with the operational properties of AI stacks, including high fixed capital costs and continuous operation.

In comparison, Fig. 2(b) shows that demand is highly sensitive to assumptions about income elasticity. Increasing the income elasticity of AI computing demand leads to large and nonlinear increases in both demand and output. Under higher income elasticity assumptions (IE_3.5 scenario), mid-century electricity demand can increase about 150% relative to the baseline elasticity case, suggesting income-driven expansion as a dominant driver of long-term electricity use. As income rises, AI adoption expands not only in intensity but also in scope, penetrating new applications and sectors. This behavior differs markedly from that of traditional consumer services and underscores the importance of modeling AI demand as an intermediate input within energy-economy scenario models.

3.3 Efficiency Trajectories and Regime Thresholds

Next, we examine how this demand behavior interacts with alternative efficiency trajectories. We ask when demand outweighs efficiency improvements in long-term electricity use. In a post-AGI

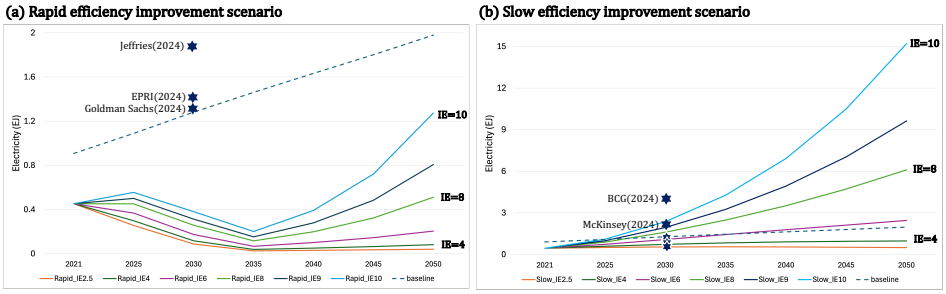

Figure 3: **Income elasticity thresholds under alternative efficiency trajectories.** Electricity demand responses to varying income elasticity assumptions under (a) *Rapid* and (b) *Slow* efficiency improvement scenarios. In (a), demand growth dominates only under very high income elasticities, whereas in (b) it does so at more moderate values. Blue stars indicate 2030 U.S. electricity demand estimates from the IEA (2025) (Table 3.2), highlighting the consistency with external projections.

setting where AI services diffuse broadly across sectors, these regime boundaries help identify when power-system constraints are likely to become a first-order limiter of compute scaling.

In Fig. 3(a), the *Rapid* trajectory delays demand-driven increases in electricity use until efficiency gains plateau, whereas the *Slow* trajectory allows income-driven growth to dominate earlier despite continued improvements. As a result, long-term projections depend not only on the pace of efficiency gains, but also on their persistence over time. Differences in efficiency trajectories, rapid early improvements followed by saturation versus slower, sustained improvements can shift the demand-dominant threshold by decades. This highlights the need to jointly consider demand response and the timing/durability of efficiency gains when representing AI compute in macro-scale energy and emissions scenario models.

4 Discussion and Limitations

The results suggest that rising AI demand does not translate mechanically into proportional increases in electricity use. Instead, long-term outcomes depend on the interaction between service growth and system-level efficiency. Under scenarios with sustained efficiency improvements, electricity demand growth can be substantially moderated, whereas slower or saturating efficiency gains allow income-driven demand to dominate over longer time horizons. In a potential post-AGI setting where AI services diffuse broadly across sectors, these regime boundaries matter because power-system dynamics may become a first-order constraint on the pace and carbon intensity of digital infrastructure scaling.

These findings may appear different from some recent reports that project rapid growth in data center electricity demand. We do not interpret this divergence as a contradiction, but as reflecting differences in scope and assumptions—particularly how quickly compute energy intensity improves and whether efficiency gains persist or plateau. By embedding AI-driven electricity demand within a macro-scale energy–economy framework and extending scenarios to mid-century, our analysis makes these interactions explicit and identifies conditions under which projections can diverge.

Sensitivity experiments for the U.S. indicate weak responsiveness to price signals but strong dependence on income growth, implying limited leverage from price-based mechanisms alone and a first-order role for efficiency progress and the scale of service expansion. Because these elasticities are calibrated and evaluated for the U.S. setting, the relative importance of price versus income effects should be validated across other countries with different market structures, electricity pricing regimes, and digital adoption pathways.

Several limitations apply. AI service output is determined endogenously within GCAM from assumed demand elasticities and macroeconomic drivers, rather than imposed as an arbitrary time series. The AI computing sector introduced here is a simplified representation and does not capture heterogeneity in workloads, infrastructure design, or deployment strategies. Moreover, we focus on the joint dynamics of service growth and system-level efficiency; location-specific grid constraints and generation-mix heterogeneity that shape siting decisions and emissions are left for future work.

ACKNOWLEDGMENTS

We thank Professor Sonny Kim and Artem Vasilev for helpful discussions and guidance on the scenario setup used in this study.

REFERENCES

J. Aljbour, T. Wilson, and P. Patel. Powering intelligence: Analyzing artificial intelligence and data center energy consumption. White Paper 3002028905, EPRI, 2024. URL https://www.wpr.org/wp-content/uploads/2024/06/3002028905_Powering-Intelligence_-Analyzing-Artificial-Intelligence-and-Data-Center-Energy-Consumption.pdf.

Eddy Bekkers, Lee Humphreys, Hryhorii Kalachyhin, Karolina Wilczynska, and Danchen Zhao. Through the looking glass: Artificial intelligence, international trade, and economic growth in the long run. Staff Working Paper ERSD-2025-09, World Trade Organization, Economic Research and Statistics Division, October 2025. URL https://www.wto.org/english/res_e/reser_e/ersd202509_e.pdf.

Abdessalam Benhari, Denis Trystram, Fanny Dufossé, Yves Denneulin, and Frédéric Desprez. Green hpc: An analysis of the domain based on top500. *arXiv preprint arXiv:2403.17466*, 2024.

Tom Brijs, Jonas Geerinck, Vivian Lee, Midas De Bondt, Bart Van Praet, Melissa Ruseler, Margot Meurrens, and Emmy Vanherpe. The power of compute: The effects of data center growth on belgium's energy system. Technical report, Boston Consulting Group, May 2025. URL https://web-assets.bcg.com/f9/d3/f35435ce484e93d0fe1e9dfdc70b/the-power-of-compute-the-effects-of-data-center-growth-on-belgiums-energy-system-r.pdf.

Katherine Calvin, Pralit Patel, Leon Clarke, Ghassem Asrar, Ben Bond-Lamberty, Ryna Yiyun Cui, Alan Di Vittorio, Kalyn Dorheim, Jae Edmonds, Corinne Hartin, et al. Gcam v5. 1: representing the linkages between energy, water, land, climate, and economic systems. *Geoscientific Model Development*, 12(2):677–698, 2019.

Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857, 2023.

Rajeev K Goel, Edward T Hsieh, Michael A Nelson, and Rati Ram. Demand elasticities for internet services. *Applied Economics*, 38(9):975–980, 2006.

Goldman Sachs Research. Gs sustain: Generational growth: Ai, data centers and the coming us power demand surge. Technical report, Goldman Sachs, April 2024. URL https://www.goldmansachs.com/pdfs/insights/pages/generational-growth-ai-data-centers-and-the-coming-us-power-surge/report.pdf.

Marius Hobbhahn and Tamay Besiroglu. Predicting gpu performance, December 2022. URL https://epoch.ai/blog/predicting-gpu-performance. Epoch AI blog post.

IEA. Energy and ai. Technical report, International Energy Agency, Paris, 2025. URL https://www.iea.org/reports/energy-and-ai. Licence: CC BY 4.0.

Jonathan Koomey, Zachary Schmidt, and Tanya Das. Electricity demand growth and data centers: A guide for the perplexed. Technical report, Bipartisan Policy Center and Koomey Analytics, February 2025. URL https://bipartisanpolicy.org/wp-content/uploads/2025/02/BPC-Report-Electricity-Demand-Growth-and-Data-Centers-A-Guide-for-the-Perplexed.pdf.

Iyngkarran Kumar and Sam Manning. Trends in frontier ai model count: A forecast to 2028. *arXiv preprint arXiv:2504.16138*, 2025.

Yan Liu. Generative ai: Catalyst for growth or harbinger of premature de-professionalization? Policy Research Working Paper 10915, World Bank, Washington, DC, September 2024. URL https://documents1.worldbank.org/curated/en/099520009172451039/pdf/IDU1aa745fd01bcf014ac51b11d1e9f762ce51e5.pdf.

Eric Masanet, Arman Shehabi, Nuoa Lei, Sarah Smith, and Jonathan Koomey. Recalibrating global data center energy-use estimates. *Science*, 367(6481):984–986, 2020.

Steven Nadel. Opportunities to use energy efficiency and demand flexibility to reduce data center energy use and peak demand. White paper, ACEEE, Washington, DC, October 2025. URL https://www.aceee.org/sites/default/files/pdfs/opportunities_to_use_energy_efficiency_and_demand_flexibility_to_reduce_data_center_energy_use_and_peak_demand.pdf.

Konstantin F Pilz, James Sanders, Robi Rahman, and Lennart Heim. Trends in ai supercomputers. *arXiv preprint arXiv:2504.16026*, 2025.

Arman Shehabi, S. J. Smith, Alexandra Hubbard, Andrew Newkirk, Nan Lei, M. A. B. Siddik, Brendon Holecek, Jonathan Koomey, Eric Masanet, and Dale Sartor. 2024 united states data center energy usage report. Technical Report LBNL-2001637, Lawrence Berkeley National Laboratory, Berkeley, CA, 2024. URL https://eta.lbl.gov/publications/2024-lbnl-data-center-energy-usage-report.

V Shiva Sankari and R Kavitha. Bitcoin adoption and price elasticity of demand: cross-country insights. *Humanities and Social Sciences Communications*, 12(1):1–12, 2025.

Charles Edison Tripp, Jordan Perr-Sauer, Jamil Gafur, Amabarish Nag, Avi Purkayastha, Sagi Zisman, and Erik A Bensen. Measuring the energy consumption and efficiency of deep neural networks: An empirical analysis and design recommendations. *arXiv preprint arXiv:2403.08151*, 2024.

A APPENDIX

A.1 AI COMPUTING SECTOR IMPLEMENTATION

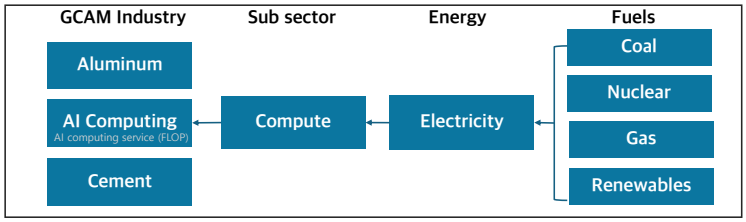

Figure 4: Schematic representation of the AI computing sector integrated into GCAM

We extend the Global Change Analysis Model (GCAM) by introducing an explicit AI computing sector within the industrial module. In its standard configuration, GCAM represents industrial energy demand through aggregated sectors such as cement, aluminum, and chemicals. These representations are insufficient to capture the distinctive characteristics of AI-related electricity demand, including computation-based service output, rapid efficiency change, and continuous operation. To address this gap, we introduce AI computing as a new industrial service sector that is structurally comparable to existing GCAM industries, but differs in its definition of output and its linkage to electricity demand.

The AI computing sector is implemented as a new subsector under the GCAM industry module (Fig. 4). This subsector produces an abstract service, AI computation, that is demanded by the economy and supplied using electricity as the energy input. By embedding the sector within GCAM's market structure, AI electricity demand competes with other end-use sectors for electricity supply, allowing power-system responses and emissions to be determined endogenously.

Note that AI service output is not imposed as an arbitrary time series in GCAM. Instead, it is determined endogenously from the assumed demand elasticities and macroeconomic drivers through the model's market equilibrium. We use the baseline (fixed γ) as a reference case anchored to near-term U.S. estimates, and treat alternative efficiency paths as counterfactual scenarios to map long-run sensitivities.

Defining AI service output We define AI service output in terms of floating-point operations (FLOP). This choice allows AI demand to be linked directly to hardware efficiency trends and electricity consumption (Pilz et al. (2025); Desislavov et al. (2023); Tripp et al. (2024)). Service output is expressed at large numerical scales using PetaFLOP (10^{15}) and YottaFLOP (10^{24}) units to ensure numerical stability within the GCAM solution algorithm.

Linking computation to electricity demand Electricity demand from AI computing is derived by converting service output(S(t)) into energy use through a system-level efficiency coefficient. Total electricity demand E(t) is calculated as:

$$E(t) = S(t) \times \gamma(t)$$

where $\gamma(t)$ represents the electricity required per unit of computation (J/FLOP). This coefficient captures not only chip-level efficiency, but also system-level factors such as cooling, networking, storage, and power usage effectiveness.

A.2 PARAMETERIZATION OF AI COMPUTING DEMAND

To represent AI computing within GCAM, we parameterize demand using a small set of key assumptions governing service growth, price responsiveness, energy efficiency, and non-energy input costs. Rather than modifying GCAM's underlying functional forms, we retain the standard industrial sector formulation and focus on constraining these parameters using ranges informed by the existing literature. Exact parameter values used in each scenario are reported in the Table 1.

Income elasticity Income elasticity governs how AI computing service demand responds to economic growth. Given that AI services are primarily used for information processing, digital services, education, and entertainment, we adopt income elasticity values in the range of *1.6–2.0* based on estimates for comparable service-oriented sectors (Bekkers et al. (2025); Liu (2024)). To explore more aggressive AI adoption trajectories characterized by rapid service expansion, we additionally consider higher values up to *2.5*.

Price elasticity Price elasticity captures the sensitivity of AI computing demand to changes in service price. Consistent with prior studies of internet-based and digitally mediated services, we assume relatively low price responsiveness. Price elasticity values are set between -0.2 and -0.7, reflecting the small share of electricity and compute costs in the overall value of AI services and the limited role of price signals in constraining demand (Goel et al. (2006); Shiva Sankari & Kavitha (2025)).

Efficiency coefficient Energy efficiency links AI service output to electricity use and is expressed as energy consumption per unit of computation (EJ/YottaFLOP). Baseline efficiency values (3.83×10^{-5} $\mathrm{EJ/YottaFLOP}$) are derived from recent estimates of GPU and data center performance in 2023 (Benhari et al. (2024); Shehabi et al. (2024); Desislavov et al. (2023)), and alternative trajectories are constructed to represent constant efficiency, *rapid* and *slow* improvement scenarios. In the *rapid* improvement case, efficiency gains accelerate in the near term and saturate after 2035 (Hobbhahn & Besiroglu (2022)), whereas in the *slow* improvement case, efficiency improves steadily through the end of the century.

Non-energy cost Non-energy costs include capital investment, labor, and operational expenditures associated with AI data centers. Using published estimates of U.S. data center capital expenditures (Goldman Sachs Research (2024)), we derive an implied non-energy cost of approximately 0.0015 1975$/PetaFLOP of computational output. As a robustness check, we also consider alternative assumptions in which non-energy costs are set to two to four times energy costs, yielding comparable magnitudes.

Together, these parameters define the demand and cost structure of the AI computing sector while remaining consistent with the broader GCAM framework.