

Reflection #8. Emergy Required for AI Use.

In Reflection #7, I explored the emergy required to develop and train a large artificial intelligence (AI) model. But training is only the beginning. Once developed, a model may be used millions or billions of times, each response requiring energy and the support of an extensive infrastructure of computers, buildings, equipment, labor, and services.

In this Reflection, I examine the emergy required for AI use, beginning with a large AI data center and then estimating the emergy required to produce the information it generates.

Emergy Evaluation of 100-MW AI Data Center

The rapid growth of artificial intelligence is bringing with it an equally rapid expansion of the computing infrastructure needed to support it. Many existing data centers require less than 10 MW of electrical capacity, while newer facilities designed for AI may require hundreds of megawatts. To get some idea of the resources required to support AI use, I used a 100-MW data center. While on the large side, it's not exceptional among the facilities now being developed. The evaluation draws on data for a much larger facility (Michael and Cottier, 2026), for which relatively complete information was available, with resource inputs scaled to represent a 100-MW facility.

The 100-MW facility evaluated here consumes approximately 712 million kWh of electricity annually, equivalent to the electricity use of roughly 68,000 American households (EIA, n.d.). It also consumes approximately 381,000 cubic meters of water each year, equivalent to the domestic water use of about 1,400 households (Dieter et al., 2018). Operating a large AI data center requires much more than electricity to run its computers. There are buildings and equipment to maintain, cooling systems to operate, and the labor and services needed to keep the facility running. When buildings, equipment, labor, and services are included, the total annual emergy is estimated at 2.54×10^{21} sej.

Table 1. Emergy evaluation of a 100-MW AI data center

Item	Units	Quantity	UEV (sej/unit)	Emergy (E18 sej/yr)	Percent of total
Operating energy					
Electricity ^{1,2}	kWh/yr	7.12E+08	9.28E+11	660.70	26.0%
Water ^{1,2}	m ³ /yr	3.81E+05	2.00E+12	0.76	0.0%
Building and equipment					
Building, 20-year life ^{2,3}	m ² /yr	1.55E+03	1.40E+16	21.68	0.9%
Digital equipment materials, 6-year life ^{2,3}	kg/yr	1.64E+05	Composite	12.44	0.5%
Transformers, 30-year life ^{2,3}	set/yr	3.30E-03	1.12E+20	0.37	0.0%
Backup generators, 25-year life ^{2,3}	set/yr	4.00E-03	2.18E+20	0.87	0.0%
Semiconductor fabrication, 6- year approximation ^{2,4}	kwh/yr	1.46E+07	9.28E+11	13.55	0.5%
Subtotal				710.37	28.0%
Labor and services					

Operating labor ^{1,2,5}	hr/yr	1.04E+05	1.10E+14	11.44	0.5%
Operating services, excluding payroll ^{1,2,5}	\$/yr	8.63E+07	3.14E+12	270.95	10.7%
Digital equipment capital services, 6-year life ^{1,2,5}	\$/yr	4.35E+08	3.14E+12	1366.59	53.8%
Facility capital services, 20-year life ^{1,2,5}	\$/yr	5.80E+07	3.14E+12	182.07	7.2%
Subtotal				1831.05	72.0%
TOTAL				2541.43	

Footnotes to Table 1.

1. Data center configuration and operating inputs. The analysis is based on the 1-GW AI data center cost model developed by Michael and Cottier (2026) for Epoch AI, scaled proportionally to a 100-MW facility. All physical inputs, labor requirements, and monetary expenditures are assumed to scale with facility capacity. The model provides estimates of electricity consumption, water requirements, staffing, capital expenditures, and operating costs. The facility uses NVIDIA GB200 NVL72 server systems. Average facility electricity consumption is calculated using 71.25% IT utilization, a power usage effectiveness (PUE) of 1.14, and 8,766 operating hours per year. The resulting evaluation represents a hypothetical large AI data center rather than measurements from an existing facility.

2. Unit emergy values. All unit emergy values (UEVs), including those for electricity, water, construction materials, labor, and services, are taken from the author's datasets deposited in Zenodo (Brown, 2025a, 2025b). These datasets support the forthcoming book *Emergy: The Coupling of Human and Natural Systems* and use the Biosphere Dynamic Empower (BDE) baseline of 40.0×10^{24} sej/yr. Composite UEVs are calculated from the constituent material UEVs.

3. Buildings and equipment. The building UEV is 1.4×10^{16} sej/m², applied to an estimated floor area of 30,970 m² and annualized over 20 years. Server rack materials are estimated using a mass of 1,472 kg per rack and a provisional composition of 50% steel, 10% copper, 10% aluminum, and 30% other materials and electronics. The latter fraction is assigned a screening UEV of 9.77×10^{12} sej/kg, based on the median of ten polymer UEVs. Rack materials are annualized over six years. Transformer and backup generator material quantities are engineering estimates annualized over 30 and 25 years, respectively. These compositions and lifetimes are assumptions of the present analysis.

4. Semiconductor fabrication. Semiconductor fabrication is evaluated from estimated electricity consumption during wafer processing. Nagapurkar and Das (2022) report cumulative energy demands of approximately 9–38 MJ/cm² for integrated circuit manufacturing, with approximately 75% attributed to on-site energy. Using the upper estimate, a 300-mm wafer area of 706.9 cm², 15,685 estimated wafer equivalents, and a six-year equipment lifetime gives approximately 1.46×10^7 kWh/yr, assuming on-site energy is electricity. Krishnan et al. (2008) provide supporting information on semiconductor manufacturing inputs. Silicon, metals, plastics, and process chemicals are excluded as separate physical inputs. A screening calculation suggests that including these additional inputs could increase total emergy excluding services by approximately 2%. This is an assumption-based estimate, not a measured materials inventory.

5. Labor and services. Operating labor is based on 50 full-time employees working 2,080 hours annually. Payroll is excluded from operating services because labor is evaluated directly. Capital expenditures are annualized using a six-year lifetime for digital equipment and a 20-year lifetime for facility infrastructure. Monetary expenditures are converted using an emprice of 3.14×10^{12} sej/\$. The original capital and operating expenditures are from Michael and Cottier (2026); the equipment lifetimes are assumptions adopted for this evaluation. Services are recorded separately from physical inputs.

6. Building-footprint aerial empower intensity (AEI). Operating AEI is calculated by dividing the annual emergy of electricity and water consumed by the data center (6.615×10^{20} sej/yr) by the estimated building footprint (30,970 m²), yielding 2.14×10^{16} sej/m²/yr. The calculation assumes single-story construction, with total floor area approximating the ground-level footprint. Capital inputs, labor, and services are excluded.

Although water consumption has become a prominent concern in discussions of AI data centers, it accounts for only 0.12% of the combined energy of electricity and water in the evaluation. The household comparisons offer a similar perspective: water consumption is equivalent to that of approximately 1,400 households, compared with 68,000 households for electricity.

Unit Energy Value of AI-Generated Information

Once an AI model has been developed and trained, it can be used repeatedly to generate information. The energy invested in model development can therefore be distributed over the information produced during the model's useful life. The energy required to operate the data center, however, continues as long as the AI model is being used. Consequently, the unit energy value (UEV) of AI-generated information depends not only on the initial investment in developing the model but also on how intensively the model is used. Table 2 gives the estimated UEVs of AI-generated information at different levels of sustained data center throughput.

Table 2. Estimated UEV of AI-generated information at different rates of model utilization

Sustained throughput ^{1,2}	Responses/day ³ (billions)	Development ⁴ (sej/bit)	AI Use ⁵ (sej/bit)	Total ⁶ (sej/bit)
25%	0.515	5.04E+05	5.36E+06	5.87E+06
50%	1.03	2.52E+05	2.68E+06	2.94E+06
75%	1.545	1.68E+05	1.79E+06	1.96E+06
100%	2.06	1.26E+05	1.34E+06	1.47E+06

Notes to Table 2

1. AI Hardware and Output Capacity. The 100-MW facility contains approximately 48,000 NVIDIA Blackwell GPUs in 667 GB200 NVL72 racks (Table 1). Throughput is estimated at 149 output tokens/s/GPU, based on the MLPerf Inference v5.1 server-mode benchmark for Llama 3.1 405B reported by Nebius (2025).

2. Sustained throughput. The 25%, 50%, and 75% scenarios represent assumed fractions of benchmark throughput sustained during continuous operation. They are sensitivity assumptions, not measured industry averages. The 100% case is a benchmark reference.

3. Information output. Each AI response is assumed to contain 300 output tokens, equivalent to approximately 1,200 characters or 240 words. Following the assumptions used in Reflection #7, raw Shannon information is estimated at 10.5 bits per word, giving 2,520 raw Shannon bits per response.

4. Model development. Development energy is estimated at 4.78×10^{20} sej from Reflection #7 and allocated over the model's assumed two-year service life. This GPT-scale estimate is used as a proxy for the model represented by the inference benchmark.

5. AI Use energy. Annual facility energy is 2.54×10^{21} sej/yr (Table 1), including electricity, water, infrastructure, equipment, labor, and services. The full annual energy is allocated to inference output at each assumed throughput level.

6. Total UEV. Total energy per output bit is the sum of allocated model-development and inference energy. Values refer to estimated raw Shannon bits, not digitally encoded bits.

The interesting thing about the results in Table 2 is how much the UEV of AI-generated information depends on the rate at which the data center produces information. At 25% of its

estimated processing capacity, the 100-MW facility generates approximately 515 million responses per day, with a UEV of 5.87×10^6 sej per raw Shannon bit. At 75% of capacity, the number of responses increases to approximately 1.55 billion per day, while the UEV falls to 1.96×10^6 sej/bit. What surprised me is that model development accounts for only about 9% of the total UEV. Most of the emergy is associated with the continuing operation of the data center, particularly the equipment and services required to maintain its computing capacity.

The UEVs of AI-generated information are several orders of magnitude lower than those estimated for human information creation in Reflection #7. Human speech, for example, has an estimated UEV of 1.17×10^9 sej per raw Shannon bit, while a scientific journal article is estimated at 1.32×10^{13} sej/bit. At 50% of estimated data center processing capacity, the UEV of AI-generated information is approximately 2.94×10^6 sej/bit. These differences are striking, but they require some caution in interpretation. Both humans and AI draw on an enormous accumulation of human knowledge. A scientist, for example, relies not only on personal education and experience but also on discoveries and information developed by countless others. Similarly, an AI model draws on knowledge incorporated during its training. Neither emergy evaluation includes the original investment required to create this accumulated knowledge. Also, producing information is not necessarily the same as creating new knowledge. The relatively low UEV of AI-generated text therefore tells us something about the emergy required to produce information, but not much about its originality or usefulness.

AI Investment and the Socioeconomic Hierarchy

I've been trying to understand what happens to the socioeconomic hierarchy when such enormous amounts of energy and economic resources are directed toward AI in such a short period of time. It's estimated that data centers consumed approximately 4.4% of U.S. electricity in 2023, with their share projected to reach approximately 10–15% by 2030 (Shehabi et al., 2024; IEA, 2025; Smith et al., 2026). U.S. private investment in AI reached approximately \$286 billion in 2025, which is equivalent to over 6% of all U.S. private nonresidential fixed investment and exceeds total federal, state, and local government investment in research and development by about 20% (Stanford HAI, 2026; U.S. Bureau of Economic Analysis, 2026).

To get a better picture of what might happen when so much investment is directed toward the upper levels of the socioeconomic hierarchy, I constructed a simple diagram (Figure 1). The idea comes from simulations I did for my dissertation in 1980, where I explored what happened when different levels of a hierarchy were disturbed. In those simulations, increasing the input to an upper level caused empower at the supporting levels to decline as resources were drawn toward the upper level. The rapid expansion of AI investment may have a similar effect, at least initially, as electricity, materials, labor, and infrastructure are drawn away from other uses. Over time, the new information-processing capacity may stimulate growth and reorganization throughout the hierarchy. Whether this happens depends on the availability of additional resources and whether the resulting reorganization increases the empower of the system as a whole.

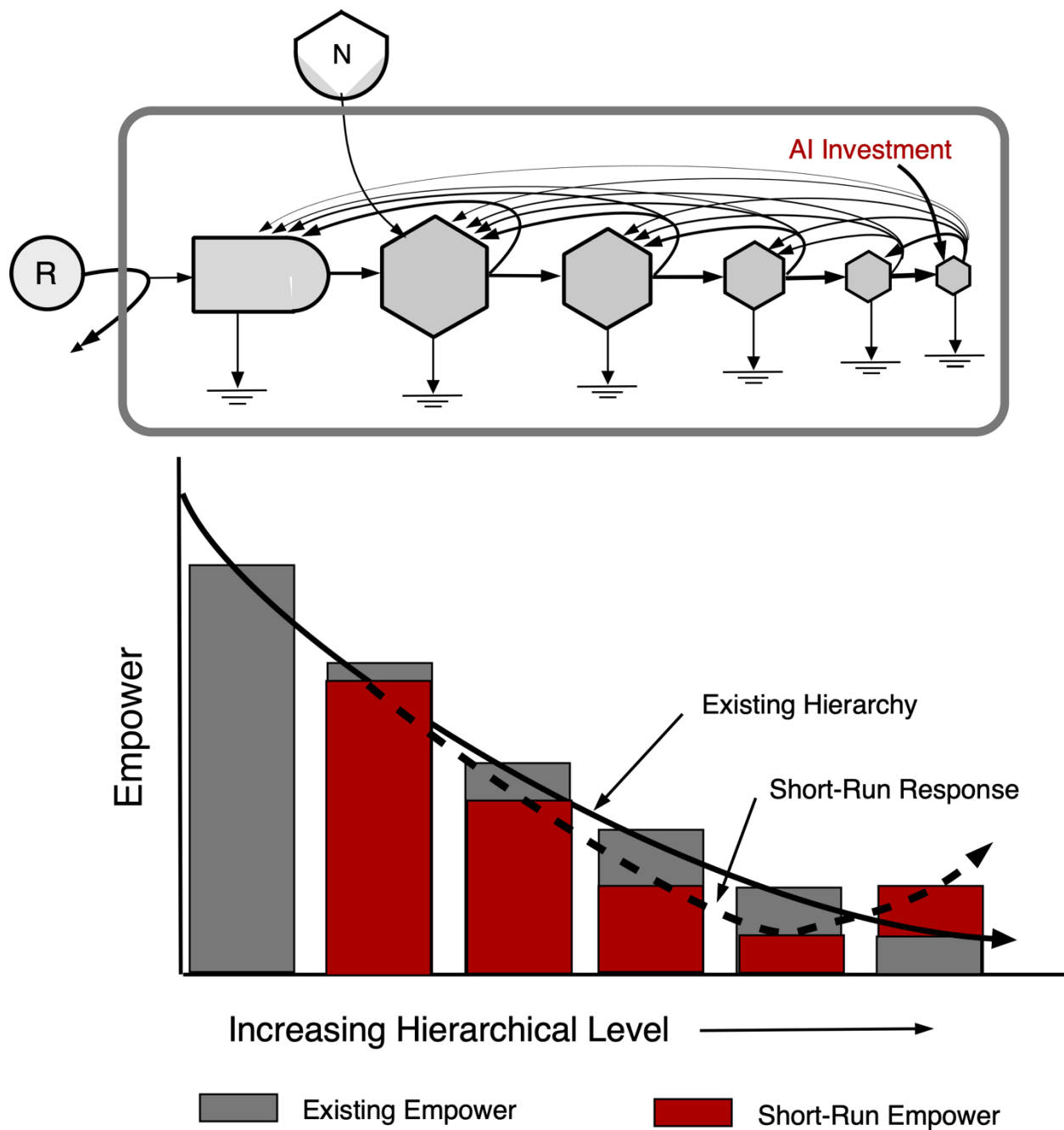


Figure 1. A possible consequence of increased investment in AI is a reduction in resources available to support other levels of the socioeconomic hierarchy. Economists refer to this displacement of investment in other sectors as “crowding out.”

Conclusions

The emergy required to support AI is substantial. The 100-MW data center evaluated here requires approximately 2.54×10^{21} sej annually, equivalent to the total annual emergy use of a small American town of about 12,000 people. Yet at 50% of its estimated processing capacity, the facility can generate more than a billion responses per day. When this emergy is distributed

over the information produced, the resulting UEV is several orders of magnitude lower than that estimated for human information creation.

One finding that surprised me is the relatively small contribution of water. Water consumption has become a major public concern as data centers expand, yet in this evaluation it accounts for only 0.12% of the combined energy of electricity and water. Of course, water availability can be critical in places where supplies are limited, but these results suggest that much of the public attention given to water may be out of proportion to its contribution to the overall resource requirements of AI data centers. Electricity, equipment, and services are far more significant.

In reflecting on the enormous energy investments made throughout human history in developing knowledge and technology, I was reminded of an old criticism of energy accounting: that we fail to include these historical investments in our evaluations. Yet if we attempted to carry forward all the energy invested in developing human knowledge and technology, the accumulated investment could dwarf contemporary resource inputs, making comparisons among products almost meaningless. And how could we possibly decide what portion of that history belongs to a scientific article, an AI model, or any other product? Knowledge and technology have developed in such interconnected ways that assigning these historical investments to individual products would be nearly impossible. The knowledge and technologies inherited from previous generations form the foundation of contemporary production, not a separate energy contribution to be added to every product.

Perhaps the most interesting question concerns what we will do with AI's remarkable ability to produce information at such a low energy cost. It can summarize and organize incredibly large amounts of information in seconds, often accomplishing what would take a human hours or days. But it can also write poetry, compose music, create images, and produce other forms of artistic expression. It may even produce works that people find as appealing as those created by human artists. What does this mean for writers, musicians, artists, and others whose work has traditionally depended on human creativity? If AI can produce similar results with so much less energy, how should we think about the value of human creative activity?

I find myself looking at both sides of this question. AI may replace some of the work people now do, but it may also give us capabilities we never had before. Someone with little musical training might compose a song, while a scientist might explore relationships in data that would otherwise take months to examine. Will AI free people to pursue new ideas and creative activities, or will it diminish the opportunities and rewards associated with work that has long been an important part of human life? I suspect we will see some of both.

And then there is the larger question of what this enormous investment in AI will mean for the socioeconomic system as a whole. AI may increase productivity, accelerate scientific discovery, and open possibilities we can scarcely imagine today. But the energy and resources required to develop and operate AI must come from somewhere. As suggested by Figure 1, concentrating resources at the upper levels of a hierarchy may initially reduce empower at the supporting levels. Whether AI eventually stimulates growth and reorganization throughout the hierarchy remains to be seen. From an energy perspective, the question is not simply how much information AI can produce, or how little energy is required to produce it, but whether the information and capabilities it provides ultimately increase the empower of the system as a whole.

References

- Brown, M. T. (2025a). Energy conversion factors derived from the EPA EmCF life cycle assessment database using a Biosphere Dynamic Empower (BDE) baseline [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.18003589>
- Brown, M. T. (2025b). Unit Energy Values (UEVs) for Energy, Materials, Products, and Environmental Flows [Data set]. Zenodo. <https://doi.org/10.5281/zenodo.18009214>
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33: 1877–1901.
- Dieter, C. A., Maupin, M. A., Caldwell, R. R., Harris, M. A., Ivahnenko, T. I., Lovelace, J. K., Barber, N. L., & Linsey, K. S. (2018). Estimated use of water in the United States in 2015. U.S. Geological Survey Circular 1441. <https://doi.org/10.3133/cir1441>
- International Energy Agency (IEA). (2025). *Energy and AI*. International Energy Agency, Paris. Published April 10, 2025.
- Krishnan, N., Boyd, S., Somani, A., Raoux, S., Clark, D., & Dornfeld, D. (2008). A hybrid life cycle inventory of nano-scale semiconductor manufacturing. *Environmental Science & Technology*, 42(8), 3069–3075. <https://doi.org/10.1021/es071174k>
- Michael, A., & Cottier, B. (2026). Servers account for 60% of the total cost of ownership of a one-gigawatt AI data center. Epoch AI. <https://epoch.ai/data-insights/ai-datacenter-cost-breakdown>
- Nagapurkar, P., & Das, S. (2022). Economic and embodied energy analysis of integrated circuit manufacturing processes. *Sustainable Computing: Informatics and Systems*, 35, 100771. <https://doi.org/10.1016/j.suscom.2022.100771>
- NVIDIA. (2024). NVIDIA GB200 NVL72 delivers trillion-parameter LLM training and real-time inference. <https://developer.nvidia.com/blog/nvidia-gb200-nvl72-delivers-trillion-parameter-llm-training-and-real-time-inference/>
- MLCommons. (2025). MLPerf Inference v5.1 benchmark results [Benchmark data]. <https://mlcommons.org/benchmarks/inference-datacenter/>
- Nebius. (2025, September 10). Nebius proves bare-metal-class performance for AI inference workloads in MLPerf Inference v5.1. <https://nebius.com/blog/posts/bare-metal-class-performance-mlperf-inference>
- Shehabi, A., Smith, S.J., Hubbard, A., Newkirk, A., Lei, N., Siddik, M.A.B., Holecek, B., Koomey, J.G., Masanet, E.R., and Sartor, D.A. (2024). *2024 United States Data Center Energy Usage Report*. Lawrence Berkeley National Laboratory, Berkeley, CA. LBNL-2001637. DOI: 10.71468/P1WC7Q.
- Smith, S.J., Hubbard, A., Newkirk, A., Ganeshalingam, M., Holecek, B., Sartor, D.A., Mills, M., and Shehabi, A. (2026). *United States Data Center Energy Usage Report: 2025 Update*. Lawrence Berkeley National Laboratory, Berkeley, CA. DOI: 10.71468/P1RP4F.
- Stanford Institute for Human-Centered Artificial Intelligence (HAI). (2026). *The 2026 AI Index Report*. Stanford University. <https://hai.stanford.edu/ai-index/2026-ai-index-report>

U.S. Bureau of Economic Analysis. (2026). *Table 5.9.5. Gross government fixed investment by type*. National Income and Product Accounts. U.S. Department of Commerce. <https://apps.bea.gov/iTable/?reqid=19&step=2&isuri=1&categories=survey#eyJhcHBpZCI6MTksInN0ZXBzIjpbMSwyLDNdLCJkYXRhIjpbWyJjYXRlZ29yaWVzIiwU3VydmV5Il0sWyJOSVBBX1RhYmxlX0xpc3QiLCI0MDAiXV19>

U.S. Energy Information Administration (EIA). n.d. Electricity use in homes. U.S. Department of Energy. <https://www.eia.gov/energyexplained/use-of-energy/electricity-use-in-homes.php>