Are LLMs More Energy Efficient Than Humans?

Dan Kuznetsov

May 2026

LLM-assisted and unassisted knowledge work consume energy in the same order of magnitude, with the LLM approach typically more efficient once broader human energy use is counted.

Contents

1 Introduction

Few people would now refuse a capable AI assistant for cognitive work. Large language models (LLMs) are increasingly used not just to answer questions but to carry out extended, multi-step tasks. Running these workflows is energy demanding. A natural question follows: which is more energy efficient at solving a given problem — an LLM or a human?

Adoption of LLMs at scale carries an energy cost that needs to be accounted for in any serious discussion of sustainability. The systems are rapidly becoming more capable, and the cost of compute per task has been growing even faster than the capability gains it buys [1]. If LLMs are substantially less energy efficient than humans, that materially affects the case for scaling LLM-based systems. If they are comparable, or more efficient, the picture changes considerably — not least because LLMs can perform work that is not feasible for humans at the same scale.

LLMs and humans turn out to be within the same order of magnitude in energy efficiency for knowledge work tasks. More importantly, energy efficiency may not be the most useful lens through which to evaluate LLM adoption — the capability expansion argument can be more consequential.

2 Defining the comparison

Not surprisingly, the hardest part of pulling this piece together was finding actual data on energy consumption and settling on meaningful metrics.

People and companies more deeply involved in the LLM business certainly have more precise figures. LLM energy consumption is also an emergent property of complex systems, and it is likely treated as a commercially sensitive matter that providers have little incentive to publish. I constrain myself to publicly available data and am happy with order-of-magnitude accuracy in estimating energy consumption for comparable tasks.

Defining a comparable metric for humans is not easy either. Should we include only the energy associated with the task itself, or the broader supply chain that keeps a knowledge worker (human or otherwise) fed, housed, and equipped? As a simple reference point, in most

comparisons I use the energy value of an average daily food intake, with reference to total daily per-capita consumption where it matters.

3 How much energy do LLMs consume?

3.1 tokens

LLMs process text as tokens — common character sequences found in a body of text [2, 3]. A short conversational query might span 100 to 300 tokens; an extended technical exchange can run to many thousands. Token count is the natural unit for measuring LLM workload, and energy per token the natural efficiency metric.

According to a post by Sam Altman [4], the average query uses about $0.34\,\mathrm{W\,h} \approx 1.2\,\mathrm{kJ}$. The number of tokens in a typical query varies widely; assuming a range of 200 to 1000 gives an energy cost of 1 to 6 J/token.

3.2 processing throughput

Energy consumption per token depends significantly on model size and hardware utilisation. A 2026 benchmark for the NVIDIA H100 [5] reports peak inference throughput of 3500 to 4000 token/s for Llama 70B and 5000 to 6000 token/s for Llama 13B.

At 700 W GPU power, peak throughput corresponds to about 0.2 J/token for the 70B model. In practice, processing a million tokens per day occupies the GPU for 2 to 3 h once model loading, batching, and queuing overheads are accounted for, which raises the figure to 5 to 8 J/token. Doubling for facility overhead (cooling, networking, storage) gives a realistic range of 10 to 15 J/token for a 70B-scale model.

That is broadly consistent with the figures reported in the 2025 MIT Technology Review article [6]: approximately 100 J for a "typical" query using Llama 3.1 8B, and approximately 7000 J for the larger Llama 3.1 405B. For a 100 to 1000 token query that corresponds to 0.1 to 1 J/token and 7 to 70 J/token respectively.

As a further cross-check, the MELODI framework paper [7] reports energy consumption ranging from 1×10^{-7} to 15×10^{-7} kW h/token for 2–8 billion parameter models and 0.9×10^{-5} to 1.4×10^{-5} kW h/token for 70 billion parameter models — corresponding to 0.4 to 5 J/token and 30 to 50 J/token respectively.

Across these three independent sources, the picture is consistent (Figure 1): energy per token rises from under 5 J/token for small models to several tens of joules per token for frontier-scale models. 10 J/token is a reasonable estimate for a large frontier-class model, and that is the figure I use in the case studies below.

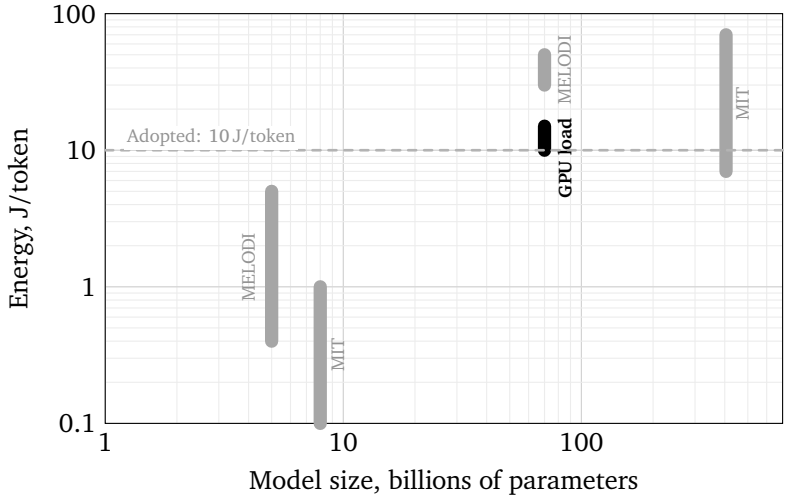

FIGURE 1: Energy per token across three independent sources, by model size. Vertical bars span the range reported by each source: MELODI [7], MIT Technology Review [6], and the throughput-based estimate from this article (black bar, "GPU load"). The dashed line marks the central estimate used in the case studies.

4 How much energy do humans consume?

A standard daily food intake is approximately 10 000 kJ [8].

A person doing the kind of cognitive work typically delegated to LLMs, however, is not living off a self-sufficient plot of land without electricity. They live in a heated or cooled home, commute, use a computer, and depend on the broader infrastructure of modern life. Per-capita energy consumption varies widely around the world [9]; a reasonable working figure is about 20 000 kW h per capita per year, or $\approx$ 200 MJ per person per day. This is corroborated by the Energy Institute Statistical Review of World Energy 2025 [10], which reports an average world energy

supply of about 70 GJ per capita per year (≈ 190 MJ per person per day).

Total per-capita energy consumption is therefore approximately 20 times higher than food intake alone. This upper bound encompasses everything — heating, transport, the supply chain that feeds us — and presumably includes a small contribution from LLM queries themselves. For the case studies that follow, I use the food-only figure as a lower bound and the full per-capita figure as an upper bound. The truth for any given task falls somewhere between.

5 Case studies

One way to compare the energy use of problem-solving with and without LLM assistance is to look at substantial tasks — ones that consumes meaningful amounts of both time and tokens. Longer tasks are easier to measure and tend to average out performance fluctuations. Below I look at two examples: one a serious physics project carried out with Claude and documented on Anthropic's blog, the other my personal experience at a much smaller scale.

5.1 vibe-physics

In March 2026, Matthew Schwartz published a post on Anthropic's blog [11] about his experience writing a high-energy theoretical physics paper while outsourcing calculations and manuscript preparation to Claude. It is a good read for anyone in a broadly similar line of work.

The numbers in the post amount to about 36×10^6 token consumed and 50 to 60 h (≈ 7 d) of human supervision. That gives an LLM energy cost of

$$36 \times 10^6 \,\text{token} \times 10\,\text{J/token} = 3.6 \times 10^8\,\text{J}.$$

Adding 7 d of human supervision on a food-only basis,

$$7\,\text{d} \times 10\,000\,\text{kJ/d} = 7 \times 10^7\,\text{J},$$

brings the total LLM-assisted workflow to $\approx 4 \times 10^8$ J.

Schwartz estimates that without LLM assistance the same project would have taken 3 to 5 months. Four months at roughly 20 working days each gives 80 d of unassisted work. On the food-only basis,

$$80\,\text{d} \times 10\,000\,\text{kJ/d} = 8 \times 10^8\,\text{J},$$

about twice the energy of the LLM-assisted workflow. On the full per-capita basis, 80 days at 200 MJ/d amounts to 1.6×10^{10} J — roughly 40 times the LLM workflow. The true figure for any given task falls between these bounds, but under either accounting the LLM-assisted approach comes out ahead.

5.2 My recent experience

As a second example, a less ambitious project came up in my own work. I needed code for the numerical solution of a two-phase flow in porous media problem with particular boundary conditions and nonlinear coefficients. It looked straightforward but required some fiddling. I worked through it with Claude Opus.

The project consumed approximately 2×10^6 to 2.5×10^6 token, plus about one day of my own time for oversight and direction:

$$2 \times 10^6 \,\text{token} \times 10\,\text{J/token} = 2 \times 10^7\,\text{J},$$

to which one day of human supervision on a food-only basis (1×10^7 J) brings the total to 3×10^7 J.

Without LLM assistance I estimate the same work would have taken about three days, or 3×10^7 J on the food-only basis — essentially identical to the LLM-assisted workflow. On the full per-capita basis, three days at 200 MJ/d amounts to 6×10^8 J, about 20 times the LLM workflow. As in the previous example, the two approaches sit within the same order of magnitude on the lower bound, and the LLM workflow is more efficient on the upper.

Summary. Across both examples, LLM-assisted and unassisted workflows sit within the same order of magnitude on the food-only baseline, and the LLM-assisted approach is between 20× and 40× more efficient on the full per-capita baseline (Table 1).

6 Does this matter?

So, we can accept the working thesis that LLMs are at least as energy-efficient as humans for the kinds of tasks considered here. How that picture evolves is harder to predict: all else equal, larger and more capable models require more energy per token. The trajectory of energy per unit of LLM capability is at least as important to watch as the current snapshot. Does this matter?

TABLE 1: Energy comparison of LLM-assisted and unassisted workflows for the two case studies. Lower-bound human energy counts food intake only; upper bound counts total per-capita energy. The LLM-assisted workflow is competitive on the lower bound and substantially more efficient on the upper.

	Vibe-physics	Two-phase flow
LLM tokens	36×10^6	2×10^6
LLM energy	3.6×10^8 J	2×10^7 J
Human supervision (food)	7×10^7 J	1×10^7 J
Total LLM workflow	4×10^8 J	3×10^7 J
Unassisted human, food only	8×10^8 J	3×10^7 J
Unassisted human, per capita	1.6×10^{10} J	6×10^8 J
Ratio (food basis)	2×	1×
Ratio (per-capita basis)	40×	20×

From the energy supply point of view, it clearly does, and it sits within a broader trend: the world needs more and more energy. Population growth is a factor but lifestyle improvement is arguably a stronger one, and lifestyle improvements are themselves powered by energy-consuming machines.

Energy efficiency, however, is not the only criterion that matters when choosing a tool. The most important thing the case studies suggest is not just that LLM-assisted workflows are competitive on energy per task — it is that they make some tasks feasible that were previously ruled out by the time and human effort required. Comparing energy per task assumes the task gets done either way. When the alternative is that it does not get done, the comparison changes shape.

A transport analogy helps frame the trade-off. A single person travelling 100 km on foot consumes approximately 20 MJ [12]. Cycling the same distance uses roughly half as much energy [13, 14]. A common petrol car consumes $\approx$ 290 MJ [14] — an order of magnitude more than walking. An electric car uses perhaps a quarter of that; a fully occupied electric train is approximately as efficient as a bicycle (Figure 2).

It would be wonderful if all transport were by bicycle and electric train. But society's needs are more varied. A heavy steel beam still has to reach a construction site that may be nowhere near a railway, and

walking with it is neither practical nor efficient. The right choice of transport depends on the task, not solely on energy per kilometre.

The same holds for LLM use. There are problems where LLM workflows deliver enormous productivity gains regardless of their energy footprint, and there are cognitive tasks better done at walking pace, e.g. where the overhead of using an LLM exceeds the cost of just doing the work yourself. Recognising which is which is itself a useful skill.

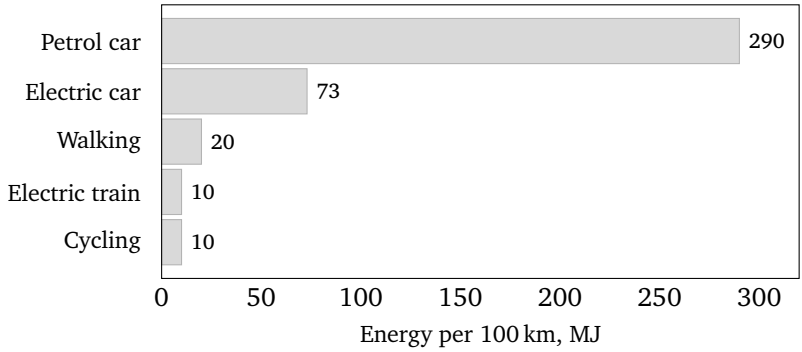

FIGURE 2: Energy use per 100 km for a single traveller across transport modes. Sources: walking [12], cycling [13, 14], vehicles [14].

7 Conclusions

For substantial knowledge work tasks, LLM-assisted workflows consume energy in the same order of magnitude as the human work they partially replace, and often less once the broader energy cost of sustaining a knowledge worker is taken into account.

For an individual practitioner, the more useful question is rarely whether an LLM is the more energy-efficient choice in some abstract sense, but whether a given task is well-suited to LLM assistance at all.

References

[1] Toby Ord. *Are the Costs of AI Agents Also Rising Exponentially?* 2025. URL: https://www.tobyord.com/writing/hourly-costs-for-ai-agents (visited on 05/04/2026).

[2] OpenAI. *Tokenizer*. OpenAI Platform. URL: https://platform.openai.com/tokenizer (visited on 05/04/2026).

[3] Stephen Wolfram. *What Is ChatGPT Doing … and Why Does It Work?* 2023. URL: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work (visited on 05/04/2026).

[4] Sam Altman. *The Gentle Singularity*. 2024. URL: https://blog.samaltman.com/the-gentle-singularity (visited on 05/04/2026).

[5] JarvisLabs Team. *NVIDIA H100 Price Guide 2026: GPU Costs, Cloud Pricing & Buy vs Rent*. 2026. URL: https://jarvislabs.ai/blog/h100-price#real-world-cost-benchmarks (visited on 05/04/2026).

[6] James O'Donnell and Casey Crownhart. *We Did the Math on AI's Energy Footprint. Here's the Story You Haven't Heard.* 2025. URL: https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech (visited on 05/04/2026).

[7] Erik Johannes Husom et al. *The Price of Prompting: Profiling Energy Use in Large Language Models Inference*. 2024. arXiv: 2407.16893 [cs.CY]. URL: https://arxiv.org/abs/2407.16893.

[8] Commonwealth of Australia. *Nutrient Reference Values for Australia and New Zealand Including Recommended Dietary Intakes*. Commonwealth of Australia, 2006. URL: https://www.eatforhealth.gov.au/sites/default/files/2022-04/n35-dietaryenergy_0.pdf.

[9] Hannah Ritchie. *Global Comparison: How Much Energy Do People Consume?* Our World in Data. URL: https://archive.ourworldindata.org/20260417-112857/per-capita-energy.html (visited on 04/17/2026).

[10] Energy Institute. *Statistical Review of World Energy*. The Energy Institute, 2025. URL: https://www.energyinst.org/statistical-review.

[11] Matthew Schwartz. *Vibe Physics: The AI Grad Student*. Anthropic Research. 2026. URL: https://www.anthropic.com/research/vibe-physics (visited on 05/04/2026).

[12] C. Hall et al. "Energy Expenditure of Walking and Running: Comparison with Prediction Equations". In: *Medicine and Sci-*

ence in Sports and Exercise 36.12 (Dec. 2004), pp. 2128–2134. URL: https://pubmed.ncbi.nlm.nih.gov/15570150.

[13] Anja Mizdrak et al. "Fuelling Walking and Cycling: Human Powered Locomotion Is Associated with Non-Negligible Greenhouse Gas Emissions". In: *Scientific Reports* 10.1 (2020), p. 9196. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC7280492.

[14] David J. C. MacKay. *Sustainable Energy — Without the Hot Air*. Cambridge, England: UIT, 2009. URL: http://www.withouthotair.com.