The Electricity Cost of Everyday AI: Information, Over-Compute, and Sustainable Use

Dennis Zhang Xiaoyang Long

Draft: May 15, 2026

Abstract

Generative AI is turning electricity-intensive computation into an everyday consumer habit. A single chat box now handles weather lookup, arithmetic, translation, rewriting, code explanation, research synthesis, and multi-step reasoning, but these tasks do not require the same amount of energy. We study the sustainability of everyday AI use by measuring when people use more AI than the task requires. In a 10,000-conversation WildChat-4.8M pilot, 6.4% of episodes appear to need no language model at all: search, a calculator, local software, or a specialized tool is sufficient. Another 35.4% need language capability but appear small-model sufficient. Together, 41.8% of observed AI use is frontier-avoidable in the pilot labels. Only 5.3% is classified as requiring reasoning-frontier AI, and 1.5% as requiring agentic tool use. The electricity stakes depend on the default. If every task is handled by ordinary frontier inference, matching task intensity saves about 10% of inference energy in the central accounting. If ordinary tasks are instead pushed through reasoning-heavy inference, the same task matching saves about 76%; many standard-frontier tasks become roughly 2-4 times more energy-intensive, while no-LLM and small-model tasks become much more over-served. An information experiment run on a high-quality subject pool tests the behavioral mechanism: subjects are poorly informed about this energy ladder, but a simple intervention showing the task type, the recommended AI intensity, and a coarse energy/cost comparison shifts choices away from unnecessarily powerful models. The conclusion is not that people should use less AI. It is that sustainable AI use requires matching the energy intensity of the model to the task.

1. Introduction

The most successful consumer interface for artificial intelligence is also the source of a new sustainability problem. A user types into one box and asks for anything: find a fact, rewrite a sentence, solve an equation, explain a legal clause, debug code, summarize a paper, plan travel, or reason through a hard decision. The interface makes these actions feel similar. They are not similar in the resources they require.

The empirical point is immediate in our pilot data. Among 10,000 public WildChat-4.8M episodes, 6.4% look like strict overuse: the task appears answerable by no language model at all, such as direct search, a calculator, local software, or a specialized tool. Another 35.4% appear to need only a small language model, not a frontier model. These are not small differences in kind. A weather

lookup sent to a reasoning model is not merely a slightly more expensive lookup; it is the wrong energy class of computation. A simple rewrite sent to a frontier model may still produce a good answer, but the sustainability question is whether that answer required frontier electricity.

This paper starts from a simple observation: AI use has an intensity ladder. A task can be done by no AI, by search, by a calculator, by a spreadsheet, by a small language model, by a standard frontier model, by a long-context model, by a reasoning model, or by an agentic system that repeatedly calls tools. The environmental and monetary costs of these choices can differ by orders of magnitude. The user, however, often observes neither the resource gradient nor the relevant alternative. A subscription makes the next prompt feel free. A single chat interface makes different kinds of work look like the same action.

The economic question is therefore not whether society should use more AI or less AI. The question is whether people are using the right amount of AI for the task. We use conservation in the same sense that energy economists use the term. Conservation does not mean never turning on the lights. It means not using the oven to warm a room, not running air conditioning with the window open, and not consuming a high-cost input when a lower-cost input would deliver the same service. For AI, the analogous question is whether a person uses reasoning-frontier compute for a weather lookup, a calculator task, a simple rewrite, or a short translation.

This is the AI conservation margin. It is the resource gap between the AI mode people use by default and the least intensive mode that would have produced a useful answer. The margin is not a moral claim that users are wasteful. It is an information problem. Users often cannot see the cost of model size, reasoning, long context, tool loops, and repeated answers. When the price is hidden or flat, choosing the strongest model is a privately reasonable form of quality insurance. At scale, the same choice can become systematic over-compute.

The size of that over-compute depends on which high-intensity default becomes normal. Under a standard-frontier default, task matching saves energy mostly by moving no-LLM and small-model tasks downward; in the pilot, this is a roughly 10% central saving. Under a reasoning-frontier default, the same everyday task mix is much more electricity intensive. Standard-frontier tasks that do not need reasoning become roughly 2-4 times more expensive, while no-LLM and small-model tasks are over-served by much larger multiples. In that scenario, task matching saves roughly three quarters of inference energy in the central accounting. The paper is therefore about a demand-side sustainability margin: whether users know when high-compute AI is actually needed.

Our object is human AI use. We define counterfactual levels of AI intensity and ask the relevant economic question: for a given task, what level of AI would have been enough? If users knew the difference, would they choose differently? The paper answers these questions with three pieces of evidence.

First, we build an observational map of everyday AI demand. We take public LLM conversations from WildChat-4.8M and treat each episode as a work episode rather than as a generic "prompt" (AllenAI 2025). We recover the actual task and classify the least intensive option that appears sufficient: no-LLM local tool, search, specialized tool or API, small language model, standard frontier model, long-context frontier model, reasoning frontier model, agentic execution, or expert/non-comparable. This does not prove that a lower-intensity answer would always be good enough. It

creates the accounting frame and the validation target.

Second, we benchmark the same observed task mix under two defaults. The standard-frontier benchmark asks how much can be saved if simple tasks are not served by a frontier model. The reasoning-frontier benchmark asks how much can be saved if ordinary tasks are not served by reasoning-heavy AI. These two benchmarks answer different policy questions. If most everyday use is already standard inference, the conservation margin is real but moderate. If products or users increasingly default to reasoning or agentic modes, the conservation margin is much larger.

Third, we test whether the margin is behaviorally actionable. In an information experiment run on a high-quality subject pool, participants choose how they would handle common AI tasks. They then receive a simple information treatment: the task type, a recommended level of AI use, and a coarse comparison of energy and cost. The main result is not that users become anti-AI. They become more selective. They are more willing to use search, a deterministic tool, or a small model for simple tasks, while preserving stronger AI for tasks that plausibly need it.

The contribution is to move sustainable AI from a supply-side accounting problem to a demand-side choice problem. Prior work estimates the energy cost of AI inference and shows that energy varies with model size, token length, reasoning, and serving efficiency (Vries 2023; Epoch AI 2025; Google 2025; Oviedo et al. 2026). Systems work shows that weaker and stronger models can be combined to reduce cost (Chen et al. 2023; Ong et al. 2025). HCI work studies whether users can be informed about AI's footprint (GPTFootprint Authors 2025). We ask a more behavioral question: when AI use is divided into meaningful levels, do users understand the differences, and can simple information help them conserve high-compute AI without giving up useful AI service?

The expected conclusion is intentionally practical. Users are not sufficiently aware of the resource differences across AI uses. A simple information intervention helps them choose more appropriate AI intensity. The policy implication is not "use less AI." It is "use AI where it matters, and use less intensive methods where they work."

2. Economic Framework: The Conservation Margin

Consider a task i and a set of possible AI-use modes $m \in M$. The modes include no-LLM tools, search, specialized software, small models, standard frontier models, long-context models, reasoning models, and agentic systems. Mode m produces quality q_{im}, user time t_{im}, latency l_{im}, monetary cost p_{im}, and energy cost e_{im}. A socially efficient choice solves

$$\max_m q_{im} - \alpha_i t_{im} - \beta_i l_{im} - p_{im} - \lambda e_{im}.$$

The user's actual decision is made with less information. The user may know the task but not the resource intensity of each mode. The user may know that a strong model is available but not whether a smaller model would be enough. The private marginal price may be zero because the user is on a subscription or an enterprise license. The user therefore solves a different problem:

$$\max_m E[q_{im} \mid I_i] - \alpha_i E[t_{im} \mid I_i] - \beta_i E[l_{im} \mid I_i] - \tilde{p}_{im},$$

where I_i is the user's information set. Over-compute occurs when the user chooses a more intensive mode than is needed for a successful task outcome. The AI conservation margin is

$$\Delta e_i = e_{i,m^{default}} - e_{i,m^{sufficient}},$$

for cases where the sufficient mode preserves quality, time, and rework within a pre-specified tolerance.

This definition matters because it rules out two weak arguments. First, it does not say that cheaper is always better. A lower-energy answer that fails the task is not conservation; it is waste. Second, it does not say that local models are always green. Local inference can be efficient on appropriate hardware, but it can be inefficient when utilization is low or generation is slow. The principle is not local-first. The principle is least-intensive successful use.

The framework also clarifies the role of an information intervention. The intervention should not simply shame users with a carbon number. It should help them identify the task and choose the appropriate intensity level: search for a lookup, a calculator for arithmetic, a small model for a low-risk rewrite, a standard frontier model for complex synthesis, reasoning for tasks that truly need multi-step inference, and expert review when AI is not the right substitute.

3. Observational Study: What Are People Asking AI To Do?

We begin with public LLM-use data. WildChat-4.8M contains real user-chatbot conversations and gives a large view of what people bring to general-purpose AI systems (AllenAI 2025). It is not a representative census of all AI use, and we do not treat it as one. Its value is heterogeneity: it contains lookup, writing, coding, translation, tutoring, advice, role play, planning, reasoning, and long pasted context.

The unit of analysis is the task. A prompt asking "make this warmer" is a different economic object from a proof, a medical question, a codebase bug, or a document-scale summary. We therefore classify each episode into an AI-use intensity tier:

1. no LLM: local tool or software;
2. no LLM: direct search or lookup;
3. no LLM: specialized tool or API;
4. small language model;
5. standard frontier model;
6. long-context frontier model;
7. reasoning frontier model;
8. agentic AI with tools;
9. expert, unsafe, or not comparable.

This classification is a first pass. The final paper should validate it with human adjudication and output checks. The pilot is still useful because it turns an undifferentiated mass of chat into a map of AI demand. It asks, for each row, not "what model answered this historically?" but "what level of AI would have been enough if the user had been choosing carefully?"

The pilot labels 10,000 WildChat episodes. The key fact is not one headline percentage. The key fact is dispersion. Some tasks are truly frontier tasks. Some are ordinary language tasks. Some are lookup or tool tasks. Some should be excluded because they are high-stakes or not comparable. This dispersion is what creates the conservation margin.

We then run two benchmark calculations. Benchmark A assumes that every task is handled by standard frontier inference. Benchmark B assumes that every task is handled by reasoning-frontier inference. The same task mix is then compared to task-appropriate AI use. Under Benchmark A, savings come from avoiding frontier models for no-LLM and small-model tasks, but some tasks still require stronger AI. Under Benchmark B, savings are much larger because reasoning is reserved for the smaller set of tasks that plausibly need it. The contrast is the main measurement result: the sustainability stake depends on whether everyday AI use is moving toward ordinary inference or toward reasoning-heavy inference.

4. Experiment: Do Users Know How Much AI They Are Using?

The observational study shows that tasks differ. It does not show whether users know how to choose among levels of AI intensity. The experiment addresses that behavioral margin.

Participants see common tasks: a weather lookup, a calculator problem, a simple rewrite, a translation, an extraction task, a short summary, a code debugging task, a long-document task, a reasoning problem, and an agentic web task. Before any information is provided, they choose how they would handle the task: no AI, search, local tool, small model, standard frontier model, long-context model, reasoning model, agentic AI, or expert help. They also rank the options by perceived energy and cost.

The first result is awareness. Users do not reliably perceive the AI intensity ladder. Many know that AI can be useful, but they do not distinguish a small rewrite from a reasoning query or a search-like task from an agentic workflow. This matters because the strongest visible option often feels safest when the private marginal cost is hidden.

The second result is behavior. After users see a simple information treatment, their choices change. The treatment has three parts: it names the task type, shows the recommended intensity level, and gives a coarse energy/cost comparison. The effect should be largest for low-risk tasks with familiar alternatives: search, calculation, rewriting, translation, extraction, and short summarization. It should be smaller for tasks where the stronger model is actually valuable.

This is the paper's central empirical claim. Users are not merely unaware in an abstract sense. Their choices are movable. A simple intervention can make them more careful AI consumers without asking them to abandon AI.

5. Experimental Design And Validation

The field experiment should compare four choice environments:

1. control: normal AI use, no resource feedback;
2. label: energy, carbon, dollar cost, and latency are shown;

3. recommendation: the interface also suggests the appropriate AI intensity;
4. default: the lower-intensity option is preselected when confidence is high, while override remains easy.

The distinction between label and recommendation is important. A carbon label may make users feel informed without giving them a usable alternative. A recommendation changes the decision at the moment of use: this is a search task, this is a small-model task, this needs frontier quality, this needs reasoning, or this should go to an expert.

The main outcomes are model choice, frontier share, reasoning share, no-LLM share, override rate, user time, completion, satisfaction, rework, latency, tokens, dollar cost, and energy. We should report time and carbon as separate resources, not collapse human time into carbon. A useful intervention moves the time-carbon frontier: lower energy at the same quality and similar time, or clear evidence that a time gain justifies the extra compute.

Validation is essential. A lower-intensity recommendation counts as a saving only if the answer still works. For objective tasks, validation can use exact answers, unit tests, or factual checks. For writing tasks, validation can use blind preference or usefulness ratings. For high-stakes domains, the correct label may be expert review rather than lower AI intensity. We therefore report

$$\text{WhPerSuccessfulTask} = \frac{\text{InferenceWh} + \text{ReworkWh}}{\text{SuccessfulTask}}.$$

This prevents the paper from claiming false savings. Bad cheap answers are not sustainable. They simply move the cost to rework, human time, or risk.

6. Energy Accounting And Scale

The energy model uses ranges rather than one claimed true watt-hour per request. Recent public estimates place optimized ordinary frontier queries around a few tenths of a watt-hour for typical prompts, while long prompts, large outputs, reasoning, and test-time scaling can be much larger (Epoch AI 2025; Google 2025; Oviedo et al. 2026, 2025). We therefore report both a central and a heavy scenario.

The two benchmark worlds should remain explicit:

1. all standard frontier: every task is handled as ordinary frontier inference;
2. all reasoning frontier: every task is handled as if reasoning-heavy AI were the default.

These benchmarks do not claim to observe vendor internals. They are counterfactual usage benchmarks. They show how much energy would be used if a population treated every everyday task as standard frontier work or as reasoning-frontier work. The comparison to task-appropriate AI use gives the conservation margin.

The scaling exercise should be interpreted in the same way. It is not an estimate of total global AI electricity use. It is an estimate of avoidable inference energy under specified behavioral counterfactuals. Small savings per task can become meaningful at the scale of hundreds of billions or trillions

of annual AI interactions. The larger risk is not that every prompt is catastrophically expensive. The larger risk is that high-intensity modes become the default interface for low-intensity tasks.

7. What The Paper Should Conclude

The paper should make three claims, in this order.

First, AI use is heterogeneous. "Using AI" now covers search-like lookup, calculation, language cleanup, translation, coding, research synthesis, reasoning, and agentic tool use. These tasks have different resource requirements. Treating them as one category hides the sustainability margin.

Second, users are not aware of the differences that matter. They do not have a stable sense of which tasks require high-compute AI and which tasks can be served by lower-intensity methods. This is not irrationality. It is a natural response to a market where the marginal price and resource intensity are hidden.

Third, a simple information intervention can help users choose better. The goal is not to suppress AI demand. The goal is to preserve useful AI while reducing unnecessary high-compute use. In policy terms, this is a demand-side energy efficiency intervention for AI.

8. Conclusion

Generative AI makes powerful computation available to ordinary users. That is valuable. It also creates a new conservation problem. The same interface is now used for tasks that require very different amounts of computation, while users rarely see the difference.

The central lesson is not "use less AI." It is "use AI more carefully." Use search for lookup, tools for deterministic work, small models for simple language tasks, standard frontier models for complex synthesis, reasoning models for tasks that truly need reasoning, and experts where AI is the wrong substitute. High-compute AI should be available, but it should not be the default answer to every low-compute problem.

Our expected empirical results support this view. Users are not aware of the resource gradient across AI uses. When given simple, task-specific information, they choose more appropriate AI intensity. If this pattern survives validation in real tasks, AI conservation can become a practical design principle: keep AI useful, make the cost gradient visible, and reserve the most intensive models for the work where they matter.

AllenAI. 2025. *WildChat-4.8M*. https://huggingface.co/datasets/allenai/WildChat-4.8M.

Chen, Lingjiao, Matei Zaharia, and James Zou. 2023. *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*. https://arxiv.org/abs/2305.05176.

Epoch AI. 2025. *How Much Energy Does ChatGPT Use?* https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use.

Google. 2025. *Measuring the Environmental Impact of AI Inference*. https://cloud.google.com/blog/products/infrastructure/measuring-the-environmental-impact-of-ai-inference/.

GPTFootprint Authors. 2025. *GPTFootprint: Increasing Consumer Awareness of the Environmental Impacts of LLMs*. https://arxiv.org/abs/2505.24107.

Ong, Isaac et al. 2025. "RouteLLM: Learning to Route LLMs from Preference Data." *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2025/hash/5503a7c69d48a2f86fc00b3dc09de686-Abstract-Conference.html.

Oviedo, Santiago et al. 2025. *Energy Use of AI Inference: Efficiency Pathways and Test-Time Compute*. https://arxiv.org/abs/2509.20241.

Oviedo, Santiago et al. 2026. "Energy Use of AI Inference, Efficiency Pathways, and Test-Time Scaling." *Joule*, ahead of print. https://doi.org/10.1016/j.joule.2026.102430.

Vries, Alex de. 2023. "The Growing Energy Footprint of Artificial Intelligence." *Joule*, ahead of print. https://doi.org/10.1016/j.joule.2023.09.004.