

Right-Sizing the Model to the Question: Cutting AI Chatbot Energy Without Losing Quality

Frits Lyneborg

FRITS AI ApS

ABSTRACT

AI chatbots answer every question with the largest model available — the same frontier system whether the user asks about contract law or for a recipe. Inference energy scales with model size and answer length, so a model roughly ten times smaller spends about a tenth of the energy, and on the easy majority of questions its answer is indistinguishable. Using only published, independently converging figures (≈ 0.3 Wh per frontier answer; 0.02–0.05 Wh for a small-model answer), we calculate that a 10,000-person service defaulting to right-sized models cuts inference energy by roughly two-thirds with no perceptible quality loss. Siting the same workload on a low-carbon European grid rather than the average US one multiplies the carbon saving to about fifty-fold. Europe is also the only place where "our AI is greener" is becoming legally verifiable, through mandatory data-centre energy reporting and AI-model energy documentation. None of it needs a breakthrough: right-size by default, escalate on demand, keep answers short, run on clean grids, publish the numbers.

1. The problem, and the finding that fixes it

Ask a chatbot for a recipe and, on the industry's default settings, the answer comes from one of the largest artefacts ever engineered — a model built to reason about

protein folding, contract law and compiler bugs. It runs the same frontier system whether the question is hard or trivial.

Most questions are trivial. So most of that compute is wasted, and it is wasted at scale: the largest services handle on the order of a billion queries a day, and data-centre energy is the fastest-growing item in global electricity demand — roughly 415 TWh in 2024, projected by the IEA to reach about 945 TWh by 2030, with AI the main driver [12 (#ref-12)].

The finding that changes the picture: **inference energy scales with model size and answer length, so a model roughly ten times smaller spends about a tenth of the energy on the same answer — and for the easy majority of questions, the smaller model's answer is indistinguishable from the frontier model's.** That recipe would have been just as good from a model a fiftieth the size, because the knowledge saturated small models years ago.

The fix follows directly. Route each question to the smallest model that answers it well, and escalate only when the question genuinely needs it. The routing technique is published and peer-reviewed; it requires no new science.

2. Why energy scales with model size and answer length

A language model performs a fixed amount of arithmetic per word it produces — roughly two floating-point operations per active parameter per generated token [2 (#ref-2)]. Energy follows the arithmetic, and two multipliers fall out.

Model size. For the same answer length, a model with ten times fewer active parameters performs roughly ten times fewer operations. Whatever the absolute energy of a frontier answer turns out to be, the *ratio* between large and small is anchored in arithmetic rather than in anyone's marketing.

Answer length. A twice-as-long answer costs twice the energy from the same model. Verbosity is an energy policy.

The industry already validates the first point architecturally. Mixture-of-experts models — Mixtral 8x7B stores 47 billion parameters but activates only 13 billion per token — match or beat much larger dense models on standard benchmarks [10 (#ref-10)]. Activating fewer parameters per token *is* the efficiency mechanism, applied inside one model. Per-question routing applies the same idea one level up.

3. What one answer actually costs

Closed systems publish little, so per-query energy was for years a guessing game. The widely quoted “3 watt-hours per ChatGPT query” traces to a 2023 third-party estimate built on stacked assumptions [13 (#ref-13)] and is now considered several times too high. Since 2025 the independent figures have converged well.

Source (year)	Figure	Basis
OpenAI, company statement (2025) [4 (#ref-4)]	≈ 0.34 Wh per average ChatGPT query	Self-reported, no methodology published
Epoch AI, independent analysis (2025) [2 (#ref-2)]	≈ 0.3 Wh per typical GPT-4o query (sensitivity: 0.1–4 Wh)	First-principles: FLOPs, H100 utilisation, ~500-token answer
Google, technical report (2025) [5 (#ref-5)]	0.24 Wh median Gemini text prompt (0.03 gCO ₂ e, 0.26 mL water)	First-party production measurement
Luccioni et al., FAccT (2024) [3 (#ref-3)]	≈ 0.047 Wh mean per text generation, models ≤ ~11 B	Measured, open models, A100
ML.ENERGY benchmark (2025–26) [6 (#ref-6)]	≈ 0.12 J per output token for an 8 B model (≈ 0.017 Wh per 500-token answer, GPU only)	Measured, H100, production-style serving
Mistral AI life-cycle analysis (2025) [7 (#ref-7)]	1.14 gCO ₂ e and 45 mL water per ~400-token frontier answer	First peer-reviewed LLM LCA (with ADEME)

Three points carry the rest of the report.

A frontier chatbot answer costs on the order of **0.3 Wh** — three independent 2025 sources, one a first-party production measurement, agreeing within ±30%. A small-

model answer of the same length costs **0.02–0.05 Wh**, bracketed by direct measurement of 8–11 B models [3 (#ref-3), 6 (#ref-6)] and consistent with scaling the frontier figure by the parameter ratio [2 (#ref-2)]. Call it six to fifteen times less.

And a single query is genuinely tiny either way. 0.3 Wh is about a fiftieth of a phone charge, or two metres of driving in an electric car [14 (#ref-14), 16 (#ref-16)]. Nobody should feel guilty about asking for a recipe. The argument here is about defaults multiplied by billions, and about which direction that multiplication runs as usage grows tenfold.

4. Does the smaller model give worse answers? On the easy majority, no

The published evidence says no, provided the routing is competent.

- **Hybrid LLM** (ICLR 2024, Microsoft Research) trained a difficulty-aware router between a small and a large model and reports “up to 40% fewer calls to the large model, with no drop in response quality” [9 (#ref-9)]. That is the cleanest peer-reviewed statement of the thesis.
- **RouteLLM** (Berkeley/LMSYS, ICLR 2025) learned routers from human preference data and retained 95% of GPT-4’s benchmark quality while sending half or more of traffic to a Mixtral-class model, cutting cost by up to 85% [8 (#ref-8)].
- **FrugalGPT** (Stanford, 2023) showed cascades — try cheap, verify, escalate — matching GPT-4 accuracy at up to 98% lower cost [1 (#ref-1)].
- **Small models are no longer weak.** A 3.8-billion-parameter model (phi-3-mini) now rivals models the size of GPT-3.5 on standard benchmarks, small enough to run on a phone [11 (#ref-11)].

Two gaps are worth stating. The routing literature measures dollar cost, not energy — dollars track compute closely, but the energy claim here comes from combining those routing rates with the per-model figures of Section 3, and the studies put power meters on nothing. And no study we know of has tested human indistinguishability on recipe-grade questions directly. Neither gap changes the direction of the conclusion.

The deeper reason routing works is distributional. Real chat traffic is dominated by short, factual, low-difficulty requests. The frontier model earns its energy on the hard tail. What is wasteful is not that frontier models exist, but that they answer traffic which never needed them.

5. Five moves to a greener chatbot

None requires new research. **Right-size by default and escalate on demand:** classify each question's difficulty — trivially cheap next to answering it — and send it to the smallest model that handles its class, escalating on difficulty, user request, or a failed first attempt. **Spend fewer tokens**, since answer length is a linear multiplier and the industry habit of padding answers with restated questions and closing pleasantries is a free saving. **Serve efficiently:** batching, quantisation and caching produced Google's reported 33× per-prompt energy reduction in one year [5 (#ref-5)]. **Site the compute on a clean grid**, a sixteen-fold difference decided purely by geography [14 (#ref-14), 15 (#ref-15)]. And **measure and publish** per-query energy, grid and water, because claims without numbers are marketing.

6. Worked example: what right-sizing saves

Every input comes from Section 3. Assume a service of 10,000 people asking 10 questions a day; a frontier answer at 0.3 Wh and a small-model answer at 0.03 Wh; and 75% of questions routable with no perceptible loss. Section 4 supports 40–75%, so the conservative case is shown too.

	Frontier-by-default	Right-sized (75% routed)	Right-sized (50% routed)
Energy per average answer	0.30 Wh	0.10 Wh	0.17 Wh
Per day (100,000 answers)	30 kWh	9.8 kWh	16.5 kWh
Per year	≈ 11.0 MWh	≈ 3.6 MWh	≈ 6.0 MWh

	Frontier-by-default	Right-sized (75% routed)	Right-sized (50% routed)
Saving vs default	—	≈ 7.4 MWh/yr (−67%)	≈ 5.0 MWh/yr (−45%)

In units a person can feel, that 7.4 MWh a year is roughly 530,000 phone charges [16 (#ref-16)], or 37,000 km of electric driving — almost once around the planet [14 (#ref-14)] — or the annual electricity of two European households [17 (#ref-17)].

For one 10,000-person service the absolute numbers are modest, which is itself a finding and a reason chatbot energy guilt is misplaced at the individual level. The point is what the default does at ecosystem scale. A service handling a billion answers a day saves at the same ratio about **73 GWh a year**, the electricity of a 20,000-household town, from routing alone.

Now add geography, because the two levers multiply. Take the year of workload above and place it on different grids [14 (#ref-14), 15 (#ref-15)]:

Scenario	Energy	Grid	CO ₂ per year
Frontier-by-default, average US grid	11.0 MWh	361 g/kWh	≈ 3,970 kg
Right-sized, average US grid	3.6 MWh	361 g/kWh	≈ 1,290 kg
Right-sized, French grid	3.6 MWh	22 g/kWh	≈ 79 kg

Same users, same questions, same-quality answers: a fifty-fold difference in carbon between the lazy configuration and the deliberate one. Neither lever required inventing anything.

7. Europe’s head start: greener will be verifiable

Green AI claims are cheap. What makes them checkable is disclosure law, and the European position is materially ahead. The recast Energy Efficiency Directive requires data centres of 500 kW and above to report energy performance and water footprint annually to a public European database [18 (#ref-18)]. Germany’s Energy Efficiency Act

sets a power-usage-effectiveness ceiling for new data centres from mid-2026 [19 (#ref-19)]. And the EU AI Act requires providers of general-purpose models to document known or estimated energy consumption [20 (#ref-20)].

None of this legislates right-sizing. What it does is make energy-per-answer a reportable, comparable quantity — the condition under which “how green is your chatbot?” stops being a slogan and becomes a procurement question. Operators who can answer with published numbers will be asked. Operators who cannot will be asked harder.

8. Limitations and honest objections

The absolute numbers are estimates. For closed systems, per-query energy rests on company statements and first-principles analysis rather than audited meters, and Epoch’s own sensitivity range spans 0.1–4 Wh [2 (#ref-2)]. The ratio argument of Section 2 survives any value in that range, which is why the report leans on it. Routing is not free either: the classifier costs compute, negligible next to generation, and mis-routing wastes either energy or a round trip. The published systems of Section 4 price that in — their quality-retention numbers are net of routing errors.

Cheaper queries invite more queries. The rebound effect is real across the history of efficiency and chatbots are not exempt. But the growth is coming regardless — the IEA projection already assumes it [12 (#ref-12)] — and a query distribution growing on a right-sized default grows several times slower in energy than the same distribution on a frontier default. Efficiency does not cancel growth. It changes the slope.

Scope, and the hard tail. This addresses the energy of answering questions. Training is a separate, large, one-off cost, and the only peer-reviewed life-cycle analysis attributes the great majority of a frontier model’s footprint to training plus cumulative inference [7 (#ref-7)]. And some questions genuinely need the largest models, so a right-sized system must escalate without friction.

9. Conclusion

The energy question in AI chat is usually framed as a trade-off between capability and footprint. For the majority of what people actually ask, that trade-off does not exist.

Making chatbots greener requires no breakthrough and no sacrifice. Only the willingness to stop answering every question with the largest model that runs.

References

1. Chen, L., Zaharia, M., Zou, J. — *FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance*. Stanford University, 2023. arXiv:2305.05176.
2. You, J. — *How much energy does ChatGPT use?* Epoch AI, Gradient Updates, February 2025. epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use.
3. Luccioni, A. S., Jernite, Y., Strubell, E. — *Power Hungry Processing: Watts Driving the Cost of AI Deployment?* ACM FAccT 2024. arXiv:2311.16863.
4. Altman, S. — *The Gentle Singularity*. Personal blog, June 2025. blog.samaltman.com/the-gentle-singularity.
5. Google Cloud — *Measuring the environmental impact of AI inference*. Technical report and blog, August 2025.
6. Chung, J.-W., et al. — *LLM Inference Energy: A Longitudinal Analysis*. ML.ENERGY, February 2026; benchmark methodology in arXiv:2505.06371.
7. Mistral AI — *Our contribution to a global environmental standard for AI*. Life-cycle analysis with Carbone 4 and ADEME, peer-reviewed by Resilio and Hubblo, July 2025.
8. Ong, I., et al. — *RouteLLM: Learning to Route LLMs with Preference Data*. UC Berkeley / LMSYS, ICLR 2025. arXiv:2406.18665.
9. Ding, D., et al. — *Hybrid LLM: Cost-Efficient and Quality-Aware Query Routing*. Microsoft Research, ICLR 2024. arXiv:2404.14618.
10. Jiang, A. Q., et al. — *Mixtral of Experts*. Mistral AI, 2024. arXiv:2401.04088.
11. Abdin, M., et al. — *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. Microsoft, 2024. arXiv:2404.14219.

12. International Energy Agency — *Energy and AI*. April 2025. Data centres $\approx$ 415 TWh (2024) $\rightarrow$ $\approx$ 945 TWh (2030).
13. de Vries, A. — *The growing energy footprint of artificial intelligence*. Joule, 2023. (Source of the superseded $\approx$ 3 Wh/query estimate; retained here for provenance only.)
14. Ember — *European Electricity Review 2025 and Global Electricity Review 2025* (direct grid intensities, 2024: EU-27 213 gCO₂/kWh; US 361; world 473); Weiss, M., et al. — *Energy Consumption of Electric Vehicles in Europe*. Sustainability 16(17), 2024 (real-world $\approx$ 0.2 kWh/km).
15. RTE — *Annual Electricity Review 2024*. France: 21.7 gCO₂/kWh direct, the lowest on record.
16. US EPA — *Greenhouse Gas Equivalencies Calculator, Calculations and References* (smartphone charge $\approx$ 0.014 kWh, citing US DOE).
17. Odyssee-Mure — *Households energy efficiency profile*: EU average household electricity $\approx$ 3.6 MWh/year.
18. Directive (EU) 2023/1791 on energy efficiency (recast), Article 12 and Annex VII; Commission Delegated Regulation (EU) 2024/1364.
19. Energieeffizienzgesetz (EnEfG), Germany, data-centre provisions (§11 ff.).
20. Regulation (EU) 2024/1689 (AI Act), Article 53 and Annex XI Section 1; Article 95(2) (b).

How to cite

Frits Lyneborg (2026). Right-Sizing the Model to the Question: Cutting AI Chatbot Energy Without Losing Quality. FRITS AI ApS, Technical Report FRITS-TR-2026-04. <https://frits.ai/research/greener-ai-right-sizing/> doi:10.5281/zenodo.21274515.

```
@techreport{lyneborg2026right,  
  title      = {Right-Sizing the Model to the Question: Cutting AI Chatbot Energy Without Losi  
  author     = {Lyneborg, Frits},  
  institution = {FRITS AI ApS},  
  number     = {FRITS-TR-2026-04},  
  year       = {2026},  
  month      = {jul},
```

```
doi      = {10.5281/zenodo.21274515},  
url      = {https://frits.ai/research/greener-ai-right-sizing/}  
}
```