Research Technical Report FRITS-TR-2026-04 5 July 2026 ENERGY EFFICIENCY ROUTING SUSTAINABILITY EU POLICY

The Cruise Missile and the Lasagne: How to Make AI Chatbots Greener by Right-Sizing the Model to the Question

Frits Lyneborg

FRITS AI ApS

ABSTRACT

Most questions people ask AI chatbots are easy — a lasagne recipe, an email draft, a word explained. Yet the industry default is to answer every question with the largest model available, the way one might sink a dinghy with a cruise missile. This report assembles the public evidence for a common-sense alternative: route each question to the smallest model that answers it well, and escalate only when the question demands it. The physics is simple — inference energy scales with model size and answer length, so a ten-times-smaller model spends roughly a tenth of the energy on the same answer. The quality cost is, for the easy majority of questions, not perceptible: peer-reviewed routing studies report 40% fewer large-model calls with no drop in response quality, and 95% of frontier quality retained while most traffic runs on far smaller models. Using only published, independently converging energy figures (≈0.3 Wh per frontier chatbot 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 its inference energy by roughly two-thirds — and that siting the same workload on a low-carbon European grid instead of the average US grid multiplies the carbon saving to roughly fifty-fold. Europe is also the only place where the claim "our AI is greener" is becoming legally verifiable, through mandatory data-centre energy reporting and AI-model energy documentation. Greener chatbots require no breakthrough: right-size by default, escalate on demand, keep answers short, run on clean grids, and publish the numbers.

1. The lasagne question

Ask a popular AI chatbot for a lasagne recipe and, on the industry's default settings, the answer is produced by one of the largest artefacts ever engineered — a model built to reason about protein folding, contract law and compiler bugs. The recipe will be excellent. It would also have been excellent from a model a tenth or a fiftieth the size, because the world's lasagne knowledge saturated small models years ago.

Nobody sinks a dinghy with a cruise missile — not because the missile wouldn't work, but because the cost is absurd once you notice it and a cheaper means does the same job. The premise of this report is that AI chat has not yet had its notice-the-cost moment: **the default is frontier-sized compute for every question, when most questions are dinghies.**

Three things make this worth writing down now. First, chatbots have become infrastructure: the largest services handle on the order of a billion queries a day. Second, the energy footprint of data centres is the fastest-growing item in global electricity demand — roughly 415 TWh in 2024, projected by the International Energy Agency to reach about 945 TWh by 2030, with AI the main driver [12 (#ref-12)]. Third — and this is the encouraging part — the technique that fixes it is published, validated, and requires no new science. What follows is the argument, the published numbers, and the arithmetic.

2. The physics of overkill

A language model answers by performing a fixed amount of arithmetic per word it produces: as a rule of thumb, about two floating-point operations per **active model parameter** per generated token [2 (#ref-2)]. Energy consumption follows the arithmetic. Two consequences fall straight out of this:

1. **Model size is a multiplier on energy.** For the same length of answer, a model with ten times fewer active parameters performs roughly ten times fewer operations. Whatever the absolute energy of a frontier answer is — and we will get to the

estimates — the *ratio* between a large and a small model's energy is anchored in arithmetic, not in anyone's marketing.

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

The industry itself already validates the first point architecturally. Mixture-of-experts models — such as Mixtral 8x7B, which stores 47 billion parameters but activates only 13 billion per token — match or beat much larger dense models across 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: activate a smaller model entirely, whenever the question permits.

3. What one answer actually costs

Closed AI 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, however, independent figures have converged remarkably 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_2e, 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

Source (year)	Figure	Basis
Mistral AI life-cycle analysis (2025) [7 (#ref-7)]	1.14 gCO_2e and 45 mL water per ~400-token frontier-model answer	First peer-reviewed LLM LCA (with ADEME)

Three points from this table carry the rest of the report:

- **A frontier chatbot answer costs on the order of 0.3 Wh** — three independent 2025 sources, one of them a first-party production measurement, agree within ±30%.
- **A small-model answer of the same length costs on the order of 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** per answer.
- **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 of this report is not about one query; it is about *defaults multiplied by billions* — and about which direction the multiplication runs as usage grows tenfold.

4. Is the lasagne worse?

The objection writes itself: surely the smaller model gives worse answers, and the energy saving is quietly paid for in quality. For the easy majority of questions, 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 the traffic to a model in the Mixtral class**, cutting cost by up to 85% on the tested benchmark [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 — a model small enough to run on a phone [11 (#ref-11)].

One honesty note: the routing literature measures **dollar cost, not energy**. Dollars track compute closely, but the energy claim in this report comes from *combining* those routing rates with the per-model energy figures of Section 3 — the studies themselves did not put power meters on anything. And no study we know of has tested "human indistinguishability on recipe-grade questions" directly; the evidence is benchmark- and preference-based. We flag both gaps rather than paper over them — neither changes the direction of the conclusion.

The deeper reason routing works is distributional: real chat traffic is dominated by short, factual, low-difficulty requests — recipes, rewordings, translations, explanations. The frontier model earns its energy on the hard tail. The waste is not that frontier models exist; it is that they are the *default* for traffic that does not need them.

5. The recipe: how to make a chatbot greener

Everything above compresses into five moves, none of which requires new research:

1. **Right-size by default; escalate on demand.** Classify each incoming question's difficulty (a trivially cheap operation compared to answering) and send it to the smallest model that handles its class well. Escalate to the frontier model on signal — difficulty, user request, or a failed first attempt. This single move converts the 6–15× per-answer gap of Section 3 into fleet-level savings at the routing rates of Section 4.
2. **Spend fewer tokens.** Answer length is a linear energy multiplier. Concise defaults, sensible output caps, and resisting the industry habit of padding answers with restated questions and closing pleasantries are all free energy savings.
3. **Serve efficiently.** Batching, quantisation, caching of repeated questions — the serving-stack improvements that produced Google's reported 33× per-prompt

energy reduction in one year [5 (#ref-5)] are available to everyone running open models.

4. **Site the compute on a clean grid.** The same workload emits ~22 gCO_2 per kWh in France and ~361 g on the average US grid — a sixteen-fold difference decided purely by geography [14 (#ref-14), 15 (#ref-15)]. Within Europe, France and the Nordics are the standouts.
5. **Measure and publish.** Per-query energy, grid, water. Claims without numbers are marketing; Section 7 shows why publishing them is about to become normal in Europe anyway.

6. Worked example: what right-sizing saves

The arithmetic is deliberately simple enough to check on paper, and every input is from Section 3. Assumptions: a service of **10,000 people**, each asking **10 questions a day**; a frontier answer at **0.3 Wh**, a small-model answer at **0.03 Wh**; and **75% of questions routable** to the small model with no perceptible loss (Section 4 supports 40–75%; we show the conservative case 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
Saving vs default	—	**≈ 7.4 MWh/yr (−67%)**	≈ 5.0 MWh/yr (−45%)

Rendered in units a person can feel, the 7.4 MWh a year that right-sizing saves this one modest service 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)].

Honesty about magnitude: 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 **one billion answers a day** — the scale of today's largest — 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 lasagne: **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 here the European position is materially ahead:

- **Data centres must now report.** The recast Energy Efficiency Directive (EU 2023/1791, Article 12) requires data centres of 500 kW and above to report energy performance and water footprint annually to a public European database, first deadline September 2024 [18 (#ref-18)].
- **Efficiency floors are arriving.** Germany's Energy Efficiency Act sets a power-usage-effectiveness ceiling of 1.2 for new data centres commissioned from mid-2026, with waste-heat-reuse quotas (an easing amendment was pending in the Bundestag as

this report went to press — the direction stands, the exact ceiling may shift) [19 (#ref-19)].

- **AI models must document energy.** The EU AI Act requires providers of general-purpose AI models to include known or estimated energy consumption in their technical documentation (Annex XI), and establishes voluntary codes of conduct on energy-efficient AI design (Article 95) [20 (#ref-20)].
- **The grid itself.** The EU's electricity averaged 213 gCO_2/kWh in 2024 against a world average of 473 and a US average of 361 — with France at ~22 and Sweden in the same league [14 (#ref-14), 15 (#ref-15)].

None of this legislates right-sizing. What it does is make energy-per-answer a *reportable, comparable* quantity — which is precisely the condition under which "how green is your chatbot?" stops being a slogan and becomes a procurement question. Operators who can answer it 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, not audited meters; Epoch's own sensitivity range spans 0.1–4 Wh [2 (#ref-2)]. The *ratio* argument of Section 2 — small models cost proportionally less — survives any value in that range, which is why the report leans on it.

Routing is not free. The classifier itself costs compute (negligible next to generation), and mis-routing has two costs: an easy question sent large wastes energy; a hard question sent small wastes a round-trip before escalation. The published systems of Section 4 already price this 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 we will not pretend chatbots are exempt: some of any per-query saving will be consumed by growth. But the growth is coming regardless — the IEA projection already assumes it [12 (#ref-12)] — and a query distribution that grows 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: inference, not training. This report addresses the energy of answering questions. Training is a separate, large, one-off cost — 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 right-sizing helps there too, indirectly: a world that routes to small models needs fewer frontier-scale training runs to serve the same traffic, though we do not quantify that here.

The hard tail is real. Some questions genuinely need the largest models, and a right-sized system must escalate to them without friction. Greener AI is not smaller AI everywhere; it is *proportionate* AI — the cruise missile stays in the arsenal, reserved for ships.

9. Conclusion

The energy question in AI chat is usually framed as a dilemma: capability or footprint. For the majority of what people actually ask, the dilemma is false. The physics says a right-sized model spends a fraction of the energy; the peer-reviewed routing literature says users cannot taste the difference on the easy majority; the arithmetic says the combination cuts a chatbot fleet's inference energy by half to two-thirds; and European grids and European disclosure law turn the remaining footprint into something both small and provable.

Making AI chatbots greener therefore requires no breakthrough and no sacrifice — only the willingness to stop answering every question with the biggest thing that runs. Right-size by default. Escalate on demand. Keep answers short. Run on clean grids. Publish the numbers.

The lasagne is just as good.

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 ≈ 415 TWh (2024) → ≈ 945 TWh (2030).
13. de Vries, A. — *The growing energy footprint of artificial intelligence.* Joule, 2023. (Source of the superseded ≈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 gCO_2/kWh; US 361; world 473); Weiss, M., et

al. — *Energy Consumption of Electric Vehicles in Europe.* Sustainability 16(17), 2024 (real-world ≈ 0.2 kWh/km).

15. RTE — *Annual Electricity Review 2024.* France: 21.7 gCO_2/kWh direct, the lowest on record.
16. US EPA — *Greenhouse Gas Equivalencies Calculator, Calculations and References* (smartphone charge ≈ 0.014 kWh, citing US DOE).
17. Odyssee-Mure — *Households energy efficiency profile:* EU average household electricity ≈ 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). The Cruise Missile and the Lasagne: How to Make AI Chatbots Greener by Right-Sizing the Model to the Question. FRITS AI ApS, Technical Report FRITS-TR-2026-04. https://frits.ai/research/greener-ai-right-sizing/

```
@techreport{lyneborg2026cruise,
  title       = {The Cruise Missile and the Lasagne: How to Make AI Chatbots Greener by Right-S
  author      = {Lyneborg, Frits},
  institution = {FRITS AI ApS},
  number      = {FRITS-TR-2026-04},
  year        = {2026},
  month       = {jul},
  url         = {https://frits.ai/research/greener-ai-right-sizing/}
}
```