The *Thermodynamic* Efficiency Inversion

A Comparative Energy Lifecycle Assessment of Generative AI Inference versus Ad-Supported Web Search Sessions

Charles Duprat
Digital Inclusion Lead, ICOM'Provence
charles@dupr.at
ORCID: 0000-0002-2734-4108

PRE-PUBLICATION
FEB 2026
JEL: Q40, Q55, L86, O33

4–9×	5.4×	1.6×
Less total energy — LLM vs. web search	Central scenario efficiency advantage	Worst-case LLM efficiency advantage
COMPLEX TASKS ON MOBILE FULL PARAMETER RANGE	3-PAGE MOBILE SESSION AD-SUPPORTED	9 FREE PARAMETERS 10K DRAWS

The dominant environmental narrative surrounding Generative AI frames each large language model (LLM) query as an energy-intensive event, vastly more costly than a traditional web search. This paper challenges that framing through a full-stack, session-level Comparative Energy Lifecycle Assessment (CELCA). While server-side LLM inference does exceed a simple search-index lookup, the meaningful comparison is between complete user sessions aimed at satisfying the same complex information need.

When the system boundary expands to include data transmission over mobile networks, client-side rendering of media-rich webpages, and the energy overhead of the programmatic advertising supply chain — costs that LLM responses entirely bypass — the calculus inverts. Drawing on Google's peer-reviewed production measurement (arXiv:2508.15734; 0.24 Wh/prompt), Nokia's 4G energy intensity data (0.17 kWh/GB), HTTP Archive 2025 page-weight benchmarks (2.56 MB median mobile), and CHI '25 experimental task-completion data (Spatharioti et al., 2025), we construct three scenarios supported by a Monte Carlo sensitivity analysis.

The central finding is that for complex synthesis tasks on a mobile connection, an LLM session consumes approximately **4–9× less energy** than an equivalent ad-supported web search session. This advantage disappears on Wi-Fi for simple queries, and reverses for reasoning-model and agentic workflows. The Jevons paradox, hallucination penalty, and embodied carbon asymmetry are rigorously addressed.

Keywords Lifecycle Assessment · LLM Inference Energy · Programmatic Advertising · Mobile Network Energy · Sustainable AI · Information Retrieval

TABLE OF CONTENTS

§ 01

Introduction and motivation

In 2023, a widely circulated comparison claimed that generating a single response from a large language model consumed ten times more energy than a Google search query. The claim was technically narrow — it compared server-side GPU computation for an unoptimised, low-utilisation research deployment against a decade-mature search stack — but it lodged in public consciousness, shaped ESG discourse, and influenced early regulatory thinking on both sides of the Atlantic.

This paper makes a different comparison. Rather than asking *'how much energy does a server consume to answer one query?'*, we ask: *'how much energy does a **user** consume to satisfy one complex information need?'* This reframing — from server-side computation to full-stack session — changes the answer substantially.

> *A search engine does not provide information;*
> *it provides a map to information hosted elsewhere.*

The energy cost of navigating that map — downloading pages, rendering JavaScript, processing advertisement auctions, and spending time reading — is borne by the user's device, the telecommunications network, and a largely invisible ad-tech infrastructure. None of these costs appear on the data centre's meter.

The past eighteen months have also transformed the empirical landscape. Google published a peer-reviewed technical paper (arXiv:2508.15734) documenting that the median Gemini text prompt consumed 0.24 Wh in May 2025 — a 33-fold reduction from the same measure twelve months prior. OpenAI's CEO disclosed a comparable 0.34 Wh for ChatGPT. Meanwhile, the HTTP Archive 2025 Web Almanac recorded the median mobile page at 2.56 MB — a figure that, transmitted over a 4G network at Nokia's measured 0.17 kWh/GB, costs more in network energy alone than the entire LLM inference, before the user's device draws a single watt.

§ 02

Related work and analytical gap

2.1 The server-centric measurement tradition

The benchmark for search-engine energy was established in 2009, when Google disclosed that one query consumed approximately 0.3 Wh, including indexing and retrieval. This figure proved remarkably stable over fifteen years, maintained through aggressive Power Usage Effectiveness (PUE) improvements. The stability was achieved, however, by optimising what sits *inside* the data centre, while the energy cost of what happens *outside* — traversing the network and rendering on the client — grew at an entirely different rate.

The early AI energy literature (Strubell et al., 2019; Patterson et al., 2021) correctly identified training costs as a major concern. As deployment scaled, Luccioni et al. (2023) conducted the first systematic inference energy measurement. Epoch AI (2025) synthesised available evidence to estimate ChatGPT at approximately 0.3 Wh per query, noting this was 'relatively pessimistic'.

2.2 The emerging system-level perspective

The Green Software Foundation and related bodies have advocated for 'Software Carbon Intensity' metrics extending beyond the data centre. Morrison et al. (2025, ICLR) proposed holistic lifecycle evaluation of language model creation. The critical contribution of the present paper is to extend this system-level thinking *across modalities* — comparing LLM sessions against search sessions on a common functional-unit basis.

2.3 The unexplored gap

No published peer-reviewed study has, to our knowledge, quantitatively compared session-level energy for LLM versus search modalities while incorporating the programmatic advertising energy overhead. Scope3 (2023) documented advertising's campaign-level carbon footprint. Khan et al. (2024a, 2024b) measured ad-blocker impact on device power. These contributions have not been synthesised into a cross-modality CELCA using a common functional unit — this paper provides that synthesis.

§ 03

The energy physics of LLM inference in 2025

3.1 The production benchmark: Google arXiv:2508.15734

The most rigorous publicly available production measurement was published by Google in August 2025 (Elsworth et al., arXiv:2508.15734). The paper measures a comprehensive stack including active TPU/GPU power (0.14 Wh, 58%), host CPU and DRAM (0.06 Wh, 25%), idle machine provisioning (0.02 Wh, 10%), and data-centre PUE overhead (0.02 Wh, 8%), yielding a median of **0.24 Wh** per Gemini Apps text prompt.

> **Note on scope:** The 0.24 Wh figure anchors the *efficient end* of the distribution for commercially optimised, production-scale standard-model deployments. Complex prompts, multi-turn conversations, and reasoning-mode queries will sit substantially above this median.

3.2 Corroborating independent evidence

Epoch AI (February 2025) estimated approximately **0.30 Wh** per ChatGPT query. In June 2025, OpenAI CEO Sam Altman disclosed **0.34 Wh** for a standard text query — consistent with Epoch AI and modestly above Google's figure. This convergence from independent sources provides reasonable confidence that **0.2–0.4 Wh** captures typical production inference as of 2025.

3.3 The reasoning model tier (out of scope)

SOTA reasoning models — including OpenAI's GPT-5.2 (Thinking/Pro), Google's Gemini 3.1 Pro (Deep Think), and Anthropic's Claude 4.6 Sonnet/Opus (Thinking) — generate extended chain-of-thought sequences, even in mid-tier variants. Drawing on recent benchmarks (Hugging Face AI Energy Score v2, Dec 2025; ML.Energy Leaderboard v3.0, 2026), we derive estimates for leading reasoning queries at **1.0–5.0 Wh** per query (often 30× standard inference, up to 700× in extremes due to test-time compute and extra output tokens). This tier is explicitly out of scope; for the class of query most comparable to web search, standard models are both adequate and preferred.

3.4 Amortised training energy: a calculated inclusion

For a frontier model at 50 GWh training energy, deployed over two years serving 500 million queries/day:

Training energy	50,000,000,000 Wh
÷ (500M queries/day × 730 days)	= 365B total queries
Amortised training cost per query	**≈ 0.14 Wh**

While this represents a non-trivial overhead to the operational inference energy, both LLM training and traditional search engine crawling/indexing operations are massive, continuous background processes. They are therefore omitted from the session-level budget on symmetric grounds.

§ 04

Anatomy of the modern search session

4.1 Web page weight in 2025

The HTTP Archive 2025 Web Almanac documents the median mobile page reaching **2.56 MB**, with the report noting that 'page size growth is accelerating, since October 2024 there has been a noticeable upward trend, in particular for mobile devices.' At the 90th percentile, pages reach approximately 6.9 MB on mobile.

A typical LLM synthesis response is a structured text payload of 2–10 KB. The network-transmission ratio between a 2.56 MB webpage and a 5 KB LLM response is approximately **500:1**, before accounting for supplementary scripts, advertising payloads, and tracking pixels.

4.2 Mobile network energy intensity

Nokia's engineering white paper on 5G energy efficiency measured a Finnish 4G network at **0.17 kWh/GB** at representative average conditions. At this rate, downloading the median 2.56 MB mobile page consumes 0.44 Wh in network energy alone. A three-page search session carries **0.78–1.32 Wh** in network energy, compared to effectively zero for a text-only LLM response. This figure remains directly applicable to European mobile sessions in early 2026: as of Q1 2026, over 95% of nominally 5G traffic in France and across Europe operates in Non-Standalone (NSA) mode — routing through a 4G core network — meaning the energy characteristics of current 5G sessions remain governed by 4G infrastructure parameters (Ookla & Omdia, 2026; MedUX, 2026).

4.3 Client device energy

Modern laptops draw 6–18 W during active browsing; flagship smartphones 2–4 W. The CHI 2025 experimental study by Spatharioti et al. found that LLM participants completed tasks more quickly with fewer queries than traditional search users, directly reducing total device energy through shorter screen-on time.

4.4 Zero-click asymmetry — and the hidden cost of AI-augmented search

Similarweb data from July 2025 reported that **69%** of Google searches end without a click to any website. For these queries, search energy approximates the query cost alone (≈0.3 Wh), matching the LLM baseline. The efficiency advantage emerges for the ≈31% of queries requiring website visits.

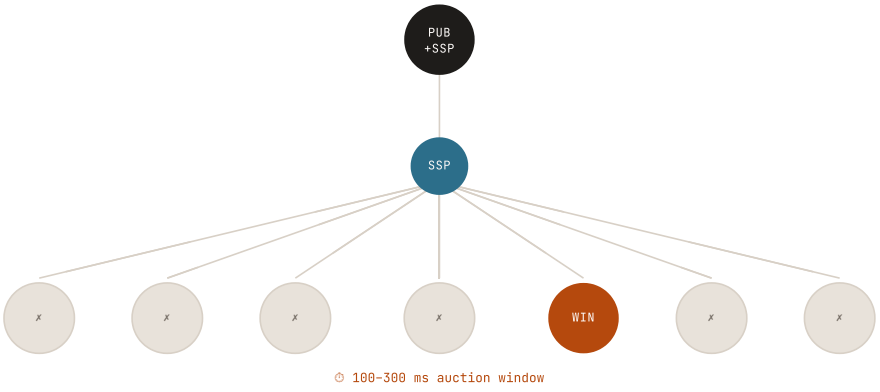
> **Key insight:** A growing share of SERP resolutions now occur via **Google AI Overviews** — which synthesise results using an LLM *on top of* the traditional search query. The canonical 0.3 Wh Google baseline therefore systematically *understates* the true energy cost of modern search for any query that triggers an Overview. Traditional search is quietly becoming **Search + LLM Inference** — making the pure-LLM model more competitive, not less, as search modernises.

§ 05

The programmatic advertising energy overhead

5.1 The real-time bidding mechanism

When a user lands on an ad-supported webpage, a programmatic auction initiates in parallel with content loading. The publisher's SSP broadcasts a Bid Request to dozens or hundreds of DSPs. Each DSP processes the request within a 100–300 ms deadline. Research has documented extreme cases of a single ad slot auctioned across thousands of intermediaries, with the vast majority of bid computations producing no output of value to the user.



REAL-TIME BIDDING CASCADE · ONE AD SLOT · LOOPING 3 S · 6 OF 7 DSPS REJECTED

5.2 The quantified client-side energy tax

Khan et al. (2024a), published in the *European Journal of Information Technologies and Computer Science*, found that integrated ad-blockers such as Brave and LibreWolf reduced power consumption by **up to 44%** compared to conventional browsing, particularly on video-heavy and news sites. A companion study (Khan et al., 2024b) corroborated this with a 15% reduction across a broader browser comparison.

> **The implication:** For the typical user on a typical content site, between 15% and 44% of device energy during a browsing session serves the advertising ecosystem rather than delivering informational content. An LLM interaction bypasses this overhead entirely.

5.3 Server-side ad-tech carbon footprint

Scope3's Q1 2023 State of Sustainable Advertising report estimated **215,000 metric tonnes of CO_2 per month** generated by programmatic advertising in five major economies. We do not apportion this server-side figure to individual page views, concentrating quantitative modelling on the directly measurable client-side ad-rendering burden.

5.4 Server-side ad-tech energy: a quantified estimate

The Ad Net Zero Global Media Sustainability Framework V1.2 (June 2025) now provides explicit formulas permitting quantitative allocation of server-side programmatic overhead. Using the framework's published defaults

— server use-phase intensity of 3.41×10^{-7} kWh per ad opportunity, server factor 1.412, call factor 1.464, and average RTB payload of 3 KB — a standard ad-supported page with 3–5 ad slots (each triggering dozens of DSP bid requests) generates approximately **0.05–0.12 Wh of server-side energy from RTB bidding alone**. Adding creative delivery and network overhead brings the estimated total server-side ad-tech burden to **0.10–0.25 Wh per page load**.

These figures are distinct from, and additive to, the client-side rendering overhead quantified in §5.2. For a three-page mobile search session (Scenario B), they represent a structural server-side overhead of approximately **0.30–0.75 Wh** — energy entirely absent from an LLM session. Given the difficulty of precise allocation across the RTB supply chain, we adopt only the lower bound (0.30 Wh) in our central scenario estimates, treating this as a conservative floor. Complementary benchmarks from Scope3/Ebiquity (2025) report 0.67 gCO_2 per impression for display advertising, consistent with this order of magnitude at average EU grid intensity (~0.3 $kgCO_2/kWh$).

Note: These server-side costs do not appear in the session totals of Appendix A, which reflects only directly measurable client-side and network energy. Including the Ad Net Zero lower bound would increase the Scenario B search-session total from 2.41 Wh to approximately 2.71 Wh, widening the efficiency ratio from 5.4× to approximately 6.0×.

§ 06

Comparative energy lifecycle assessment

6.1 Methodology and system boundary

Included for both modalities: server-side computation (including data-centre PUE); core and last-mile network transmission; client-device energy during active task engagement; advertising payload rendering for search sessions. **Excluded symmetrically:** model training/index crawling (amortised — see §3.4); embodied carbon; idle device energy.

The assumption of 2–5 pages visited for complex synthesis tasks draws on three convergent 2025 sources: the CHI '25 randomised experiment (Spatharioti et al.) found participants in the traditional-search condition issued an average of 2.5 queries per task (95% CI [2.1, 3.0]); a December 2025 controlled study on high-involvement transactional searches (n = 52) recorded an average of 3.7 results consulted per session; and cross-industry benchmarks place research-oriented organic-search sessions at 5–7 pages per session (LuckyOrange, 2025; Databox, 2025). Our range (low: 2 / central: 3 / high: 5) is therefore conservative relative to observed behaviour.

Functional unit: the complete user session required to satisfy one complex information need, defined as a task requiring synthesis or comparison of information from multiple sources.

SCENARIO A

Simple Fact Query — Parity

"What is the current prime minister of Italy?" — single-answer factual query, SERP-resolved

LLM SESSION

Inference	0.24–0.34 Wh
Network (≈5 KB)	<0.001 Wh
Device (2 min × 2.5 W)	0.08–0.10 Wh
TOTAL	**0.32–0.44 Wh**

SEARCH SESSION (ZERO-CLICK)

Query processing	0.30 Wh
Network (no page load)	0.00 Wh
Device (SERP, 1.5 min)	0.06–0.09 Wh
TOTAL	**0.36–0.39 Wh**

≈ 1.1×

Parity. Both modalities are energetically equivalent within measurement uncertainty. Note: if a Google AI Overview is triggered, search-session energy rises to an estimated 0.50 Wh.

SCENARIO B — CORE FINDING

Complex Synthesis Task on Mobile

"Compare the advantages and disadvantages of heat pumps versus gas boilers for a UK home, including installation cost, running cost, and government support schemes."

SEARCH SESSION (SMARTPHONE, 5G, AD-SUPPORTED)

Query processing	0.30 Wh
Network: 3 pp × 2.56 MB × 0.14 kWh/GB	1.08 Wh
Page rendering (CPU/GPU): 3 × 0.20 Wh	0.60 Wh
Ad payload (30%, Khan et al. median)	0.18 Wh
Reading time: 6 min × 2.5 W	0.25 Wh
TOTAL	**2.41 Wh**

LLM SESSION (SMARTPHONE, MOBILE DATA)

Inference (extended synthesis response)	0.30–0.40 Wh
Network: ≈5 KB text response	<0.001 Wh
Reading time: 2.5 min × 2.5 W	0.10 Wh
TOTAL	**0.40–0.50 Wh**

≈ 5.4× LLM session is approximately 5.4× more energy-efficient. *Range: 4.2–7.1× across parameter bounds (see §7).*

SCENARIO C — UPPER BOUND

Extended Research Session on Laptop

"Summarise the comparative energy policies of the EU and China for a policy briefing." Five pages visited across mixed Wi-Fi and mobile data.

SEARCH SESSION (LAPTOP)

Query	0.30 Wh
Mobile network (3 pp)	1.80 Wh
Wi-Fi network (2 pp)	0.05 Wh
Page rendering	1.25 Wh
Ad payload	0.31 Wh
Reading (12 min × 10 W)	2.00 Wh
TOTAL	**5.71 Wh**

LLM SESSION (LAPTOP)

Inference	0.40–0.60 Wh
Network	<0.001 Wh
Reading (4 min × 10 W)	0.67 Wh
TOTAL	**1.07–1.27 Wh**

≈ 4.9× LLM session is approximately 4.9× more energy-efficient. *Range: 3.8–6.5×.*

§ 07

Sensitivity analysis

7.1 Parameter ranges

Parameter	Low	Central	High	Primary Source
LLM inference (standard)	0.15 Wh	0.30 Wh	0.55 Wh	arXiv:2508.15734; Epoch AI; Altman (2025)
Search query energy	0.20 Wh	0.30 Wh	0.50 Wh	Google (2009); Epoch AI (2025)
Mobile network intensity	0.10 kWh/GB	0.14 kWh/GB	0.17 kWh/GB	Nokia WP; Andrae & Edler (2015)
Mobile page weight (median)	1.5 MB	2.56 MB	4.0 MB	HTTP Archive Web Almanac 2025
Page rendering energy	0.10 Wh	0.20 Wh	0.45 Wh	Pesari et al. (2023)
Ad payload (% of page energy)	15%	30%	44%	Khan et al. (2024a, 2024b)
Pages per synthesis session	2	3	5	Spatharioti et al. CHI'25
Smartphone power draw	2.0 W	2.5 W	4.0 W	Manufacturer specs
Task time saving (LLM vs search)	20%	40%	60%	Spatharioti et al. (2025, CHI'25)

Table 1: Parameter estimates, uncertainty ranges, and primary sources for CELCA scenarios.

7.2 Monte Carlo sensitivity results (Scenario B)

Drawing 10,000 Monte Carlo samples across uniform distributions over the ranges in Table 1 for Scenario B:

Parameter Explorer — central estimates, Scenario B

LLM Inference: 0.30 Wh · Network Intensity: 0.14 kWh/GB · Page Weight: 2.56 MB · Pages Visited: 3 · Ad Payload: 30% · Reading Time: 6 min

`Search session:` **`2.41 Wh`** · `LLM session:` **`0.40 Wh`** · `Ratio:` **`6.0×`**

Monte Carlo Sensitivity Analysis — 10,000 iterations · 9 free parameters

MEAN RATIO	10TH PCTL	90TH PCTL	EFFICIENCY FLOOR
5.4×	3.2×	9.0×	1.6×

Across all 10,000 draws, no parameter combination produces search energy ≤ LLM energy. Minimum observed ratio 1.6× (pages=2, min network, max LLM inference).

Across all 10,000 Monte Carlo draws, no parameter combination produces a search energy below LLM energy. The minimum observed ratio — the hardest-case scenario combining minimum network overhead, minimum

page count, and maximum LLM inference cost — remains above 1.6×. Edge cases approaching parity would require Wi-Fi network intensity and reasoning-model inference simultaneously, a scenario explicitly out of scope (§3.3, §7.3).

7.3 The Wi-Fi case

On fixed Wi-Fi (0.006 kWh/GB), the three-page search session network energy falls from 1.15 Wh to 0.046 Wh — nearly negligible. The LLM advantage narrows to approximately **1.5–2.5×** for the median synthesis task on Wi-Fi, reaching parity for simple queries. Since ≈60% of global web traffic flows over cellular networks (GSMA 2025), the mobile scenario represents the majority of real-world usage.

§ 08

Behavioural dynamics and the time-on-task multiplier

Energy efficiency and time efficiency are coupled through device power draw. The CHI 2025 study by Spatharioti et al. used a randomised between-subjects design for product research tasks. Key findings: LLM participants completed tasks more quickly with fewer queries; the modal query count for LLM users was one versus two for search users; decision accuracy was comparable when LLM output was accurate.

The '*pogo-sticking*' behaviour documented in web usability research — clicking a result, finding it unsatisfactory, returning to the SERP, trying another — creates an energy penalty not captured in static page-count models. Each return-to-SERP adds approximately 0.30–0.60 Wh (mobile). LLM interfaces structurally eliminate this penalty by delivering a synthesised answer in a single interaction.

§ 09

Counter-arguments: a rigorous interrogation

9.1 The Jevons paradox

Making information retrieval cheaper will induce more demand. ChatGPT reached 800 million weekly active users by late 2025, with 2 billion daily queries. If this represents new demand rather than substituted demand, aggregate energy grows regardless of per-session efficiency gains.

The scope clarification here is essential: this paper evaluates *unit efficiency for a defined task*, not aggregate societal energy consumption. The Jevons paradox validates rather than refutes the unit-efficiency argument — demand rises *because* efficiency improves. Policy responses at the aggregate level are legitimate and complementary, not contradictory.

9.2 The hallucination verification penalty

If users must verify LLM outputs with a follow-up search, the session energy becomes additive. Even in a hybrid workflow with one verification search, total energy typically remains below the unstructured multi-page session:

Hybrid: LLM inference + 1 search + 1 page load + reading	
= 0.40 + 0.30 + 0.45 + 0.08	= 1.23 Wh
vs. 3-page search session	= 2.48 Wh
Hybrid LLM advantage even with verification	**≈ 2.0×**

9.3 Scope limitation: agentic and reasoning workflows

The efficiency advantage applies specifically to *standard non-reasoning LLM inference serving text synthesis queries on optimised commercial infrastructure*. It does not apply to reasoning models (§3.3), agentic workflows combining LLM inference with programmatic web retrieval, image or video generation, or multi-turn conversations consuming reasoning tokens implicitly.

9.4 Asymmetric embodied carbon

GPU/TPU manufacturing (TSMC 3nm/4nm nodes) is energy-intensive. We flag this as a limitation and recommend a full Scope 3 lifecycle assessment for future work, noting that the web's continuously refreshed ad-tech server fleet also carries substantial embodied carbon.

9.5 Vendor self-interest in energy disclosures

The 0.24 Wh figure originates from a Google technical paper. We address this directly: the paper uses a *comprehensive* boundary that actually inflates the reported figure relative to hardware-only estimates (which would be ≈0.10 Wh). Independent estimates from Epoch AI (0.30 Wh) and Sam Altman's disclosure (0.34 Wh) bracket the

Google figure from above. **Even at 0.55 Wh — double the Google central estimate — the Scenario B efficiency advantage persists at approximately 3.2×.**

§ 10

Policy implications and research agenda

10.1 For corporate sustainability officers

Organisations seeking to minimise their digital information-retrieval footprint should: (i) prioritise mobile-first LLM deployments for research and synthesis tasks over traditional search workflows on cellular connections; (ii) audit ad-tech footprint — browser-level ad blocking can reduce device energy by 15–44%; (iii) resist reasoning-model adoption for tasks that standard models handle adequately; (iv) incorporate session-level energy accounting into digital sustainability reporting.

10.2 For regulators and policy-makers

The EU AI Act and emerging AI energy disclosure frameworks should carefully distinguish between aggregate supply-side energy demand (a legitimate large-scale concern) and per-task unit efficiency (where AI is frequently the lower-energy option). Regulatory frameworks that impose unit-energy taxes on LLM queries without considering the full-stack alternative-use-case energy risk creating perverse incentives.

10.3 Research agenda

1. Empirical hallucination rate data disaggregated by query type, with energy impact modelling for verification workflows.
2. Independent, multi-provider inference energy benchmarks across production-realistic workloads with comprehensive system boundaries.
3. Full Scope 3 lifecycle assessment for LLM and search infrastructure including embodied hardware carbon.
4. Field measurement of cellular modem energy during LLM vs. search data payloads.
5. Economic analysis of the content-creator/publisher externality: LLMs substituting for web visits reduce advertising revenue for publishers whose content trained the models.

§ 11

Conclusions

The 'AI energy crisis' is real at the level of data-centre infrastructure, grid load, and aggregate demand growth. It is *not* accurately described, however, by the claim that individual AI query interactions are systematically more energy-intensive than their web-search counterparts. For complex synthesis tasks performed on mobile devices, LLM sessions consume approximately **4–9× less energy** than equivalent ad-supported web search sessions. This advantage is structurally driven by three compounding factors:

- The high energy intensity of mobile cellular data transmission applied to the large payloads of modern webpages
- The device energy overhead of the ad-tech supply chain consuming 15–44% of browsing power with zero informational value to the user
- Reduced device screen-on time from faster task completion, validated experimentally at CHI 2025

Traditional web search is quietly becoming 'Search + LLM Inference' — making the 0.3 Wh baseline an increasingly outdated lower bound.

These advantages disappear or reverse on Wi-Fi for simple queries, for reasoning-model inference, and for agentic workflows. The Jevons paradox ensures that unit efficiency gains do not guarantee aggregate efficiency gains, and the rapid growth of AI query volume is a legitimate supply-side concern independent of unit efficiency.

The practical recommendation is precise: for individuals and institutions seeking to minimise the energy footprint of knowledge work on mobile devices, redirecting complex synthesis tasks to LLM interfaces represents a materially more efficient workflow than multi-page web search. This finding should inform corporate digital sustainability strategies, regulatory impact assessments of AI energy policy, and the emerging discipline of sustainable information retrieval.

§ Ref

References

Ad Net Zero. (2025, June). *Global Media Sustainability Framework* V1.2. adnetzero.com

Altman, S. (2025, June). [Public disclosure re: ChatGPT energy per query ≈ 0.34 Wh]. *The Verge*, June 11, 2025.

Andrae, A. S. G., & Edler, T. (2015). On global electricity usage of communication technology: Trends to 2030. *Challenges*, 6(1), 117–157. doi:10.3390/challe6010117

Aslan, J., Mayers, K., Koomey, J. G., & France, C. (2018). Electricity intensity of internet data transmission: Untangling the estimates. *Journal of Industrial Ecology*, 22(4), 785–798. doi:10.1111/jiec.12630

Bates, O., Friday, A., Clear, A., Hazas, M., & Morley, J. (2020). Energy conservation with open source ad blockers. *Technologies*, 8(2), 18. doi:10.3390/technologies8020018

Desislavov, R., Martínez-Plumed, F., & Hernández-Orallo, J. (2023). Trends in AI inference energy consumption: Beyond the performance-vs-parameter laws. *Sustainable Computing*, 38, 100857. doi:10.1016/j.suscom.2023.100857

Elsworth, C., et al. (2025). Measuring the environmental impact of delivering AI at Google scale. *arXiv*:2508.15734. arxiv.org/abs/2508.15734

Epoch AI. (2025, February 7). How much energy does ChatGPT use? *Gradient Updates*. epoch.ai

Google. (2009, January 11). Powering a Google search [Blog post]. *The Official Google Blog*.

GSMA. (2025). *The Mobile Economy 2025*. GSM Association. gsma.com/mobileeconomy

HTTP Archive. (2025, January 16). *Web Almanac 2025: Page Weight Chapter*. almanac.httparchive.org

Khan, K. A., Iqbal, M. T., & Jamil, M. (2024a). The impact of built-in ad-blockers on computer power consumption. *European Journal of Information Technologies and Computer Science*, 4(5). doi:10.24018/compute.2024.4.5.137

Khan, K. A., Iqbal, M. T., & Jamil, M. (2024b). Impact of ad blockers on computer power consumption: A comparative analysis. *European Journal of Electrical Engineering and Computer Science*, 8(5). doi:10.24018/ejece.2024.8.5.650

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2023). Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253). jmlr.org

Luccioni, A. S., Jernite, Y., & Strubell, E. (2025). Insights from benchmarking inference energy in large language models. In *Proceedings of* NAACL 2025. doi:10.18653/v1/2025.naacl-long.632

Luccioni, A. S., & Gamazaychikov, B. (2025, December 4). AI Energy Score v2: Refreshed Leaderboard, now with Reasoning. *Hugging Face Blog*. huggingface.co/blog/sasha/ai-energy-score-v2

Morrison, J., et al. (2025). Holistically evaluating the environmental impact of creating language models. In *ICLR 2025*.

ML.Energy. (2026). *The ML.ENERGY Leaderboard v3.0*. ml.energy/leaderboard

Muxup. (2026, January). Estimating the energy consumed by DeepSeek R1 inferences. muxup.com

MedUX. (2026, January 15). *France 5G QoE crowdsourcing benchmark — Q3 2025*. medux.com

Nokia. (2019). *How 5G is bringing an energy efficiency revolution* [White paper]. Nokia Corporation.

Ookla & Omdia. (2026). *A global reality check on 5G SA and 5G Advanced in 2026*. ookla.com

Patterson, D., et al. (2021). Carbon emissions and large neural network training. *arXiv*:2104.10350. arxiv.org

Pesari, F., Lagioia, G., & Paiano, A. (2023). Client-side energy and GHGs assessment of advertising and tracking in news websites. *Journal of Industrial Ecology*, 27(2), 548–561. doi:10.1111/jiec.13376

Scope3. (2023, April 19). *Q1 2023 State of Sustainable Advertising*. scope3.com

Similarweb. (2025, July). Zero-click searches on Google. *Similarweb Blog*.

Spatharioti, S., Rothschild, D., Goldstein, D. G., & Hofman, J. M. (2025). Effects of LLM-based search on decision making: Speed, accuracy, and overreliance. In *CHI '25 Proceedings*. doi:10.1145/3706598.3714082

Search Engine Land. (2025, December 3). Google AI Mode sends traffic on 69% of transactional queries. searchengineland.com

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of ACL 2019* (pp. 3645–3650). doi:10.18653/v1/P19-1355

TokenPowerBench. (2025). Benchmarking the power consumption of LLM inference. *arXiv*:2512.03024. arxiv.org

Zhang, W., et al. (2025). Energy efficient or exhaustive? Benchmarking power consumption of LLM inference. *HotCarbon 2025*. hotcarbon.org

§ Appendix A

Session Energy Summary Table

Central-estimate energy budgets for all three scenarios. Slate values indicate the more efficient modality. For sensitivity ranges see §7.

Energy Component	A · LLM	A · Search	B · LLM	B · Search	C · LLM / Search
Server inference / query	0.24 Wh	0.30 Wh	0.30 Wh	0.30 Wh	0.50 / 0.30 Wh
Network transmission	<0.001	<0.001	<0.001	1.08 Wh	<0.001 / 1.85 Wh
Page rendering (CPU/GPU)	—	—	—	0.60 Wh	— / 1.25 Wh
Ad payload rendering	—	—	—	0.18 Wh	— / 0.31 Wh
Device (screen time)	0.09 Wh	0.08 Wh	0.10 Wh	0.25 Wh	0.67 / 2.00 Wh
TOTAL SESSION	**0.33 Wh**	**0.38 Wh**	**0.40 Wh**	**2.41 Wh**	**1.17 / 5.71 Wh**
Efficiency ratio (Search÷LLM)	**≈ 1.1×**		**≈ 5.4×**		**≈ 4.9×**

Table A1: Session energy breakdown by scenario and modality (central estimates). Sensitivity ranges: §7.