

# The *Thermodynamic* Efficiency Inversion

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

Charles Duprat

ORCID: 0000-0002-2734-4108

WORKING PAPER

MAR 2026

JEL: Q40, Q55, L86, O33

charles@dupr.at

DOI: 10.2139/ssrn.6287918

**4–9<sup>×</sup>**

Less total energy  
LLM vs. web search

COMPLEX TASKS ON MOBILE  
FULL PARAMETER RANGE

**5.4<sup>×</sup>**

Central scenario efficiency  
advantage

3-PAGE MOBILE SESSION  
AD-SUPPORTED

**1.6<sup>×</sup>**

Lower bound standard-LLM  
advantage

9 FREE PARAMETERS  
10K DRAWS

The dominant environmental narrative around generative AI treats each large language model (LLM) prompt as an energy-intensive event and compares it to a traditional web search query at the server level. This paper argues that such comparisons use the wrong functional unit. The relevant unit is not query-level computation, but *task-level session energy*: the total energy required to satisfy one user information need across the full retrieval stack. We therefore conduct a full-stack Comparative Energy Lifecycle Assessment (CELCA) of two modalities for complex information-seeking tasks: direct LLM synthesis and ad-supported web search.

The analysis includes server-side computation, mobile or fixed-network transmission, client-device energy during active use, webpage rendering, and the client-side burden of programmatic advertising — costs largely avoided by direct LLM responses. Using Google's peer-reviewed production benchmark for Gemini inference ([arXiv:2508.15734](https://arxiv.org/abs/2508.15734)), 0.24 Wh per median text prompt), HTTP Archive 2025 mobile page-weight data, Nokia mobile-network intensity estimates, CHI '25 experimental task-completion evidence ([Spatharioti et al. \(2025\)](#)), and a 10,000-draw Monte Carlo sensitivity analysis, we model matched user sessions rather than isolated queries.

The central finding is conditional but substantial: for complex synthesis tasks performed on mobile connections, a standard-LLM session consumes approximately **4–9× less energy** than an equivalent ad-supported web-search session, with a central estimate of 5.4× and a worst-case floor of 1.6× across all modelled parameter combinations. This advantage narrows to parity for simple zero-click queries on fixed Wi-Fi, and reverses for reasoning models and agentic workflows utilising test-time compute. The paper's contribution is therefore methodological as well as empirical: when energy accounting shifts from queries to tasks, the conventional LLM-versus-search narrative is materially altered.

## Keywords

Lifecycle assessment

LLM inference energy

Programmatic advertising

Mobile network energy

Sustainable AI

Information retrieval

For version history, see [dupr.at/thermodynamic-efficiency-inversion](https://dupr.at/thermodynamic-efficiency-inversion).

**TABLE OF CONTENTS**

#	Section
1.	Introduction and motivation
2.	Related work and analytical gap
3.	The energy physics of LLM inference in 2025
4.	Anatomy of the modern search session
5.	The programmatic advertising energy overhead
6.	Comparative energy lifecycle assessment
7.	Sensitivity analysis
8.	Behavioural dynamics and the time-on-task multiplier
9.	Counter-arguments: a rigorous interrogation
10.	Policy implications and research agenda
11.	Conclusions
–	References
A	Appendix A – Session energy summary
B	Appendix B – Mathematical formulation

§ 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](https://arxiv.org/abs/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.

## 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)) proposed holistic lifecycle evaluation of language model creation, while a recent systematic review (Oliveira et al. (2026)) confirms that narrow, server-only operational boundaries systematically underestimate the true environmental impact of deployed AI. 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, and (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. A recent PRISMA-compliant systematic review (Oliveira et al. (2026)) corroborates this gap, finding that cross-modal lifecycle comparisons remain absent from the peer-reviewed literature and that heterogeneous functional units preclude cross-study synthesis; a limitation this paper directly addresses.

§ 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.

### 3.2 Corroborating independent evidence

Epoch AI (February 2025) estimated approximately **0.30 Wh** per ChatGPT query, followed by OpenAI CEO Sam Altman disclosing **0.34 Wh** for a standard text query in June 2025. This floor has now been robustly validated by peer-reviewed literature. A comprehensive bottom-up simulation of frontier-scale deployments by Microsoft researchers [Oviedo et al. \(2026, Joule\)](#) establishes a median node-level energy of **0.31 Wh** per standard query, concluding that prior estimates based on non-production assumptions systematically overstate energy use by 4-20×. This consensus confirms that our central estimate (0.30 Wh) accurately captures typical, highly-optimised production inference as of mid-2026.†

**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.

### 3.3 The reasoning model tier (out of scope)

SOTA reasoning models — including leading frontier offerings from OpenAI, Google, and Anthropic as of Q1 2026 † — 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.

**Specifically, at their upper bound:** OpenAI GPT-5.4 (xhigh/Pro), Google Gemini 3.1 (Pro/Deep Think), and Anthropic Claude 4.6 Sonnet/Opus (adaptive reasoning, max effort), as of March 2026. Model naming conventions and capability tiers evolve rapidly.

### 3.4 Amortised training energy

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
-----	
<b>AMORTISED TRAINING COST PER QUERY</b>	<b>≈ 0.14 Wh</b>

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

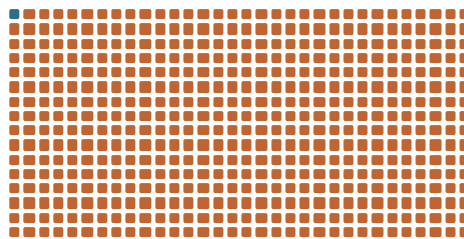
## Anatomy of the modern search session

### 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.

LLM Payload (5 KB) vs Webpage Payload (2,560 KB)



NETWORK TRANSMISSION RATIO – APPROXIMATELY 500:1 BEFORE SUPPLEMENTARY SCRIPTS AND TRACKING

### 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 ( $\approx 0.3$  Wh), matching the LLM baseline. The efficiency advantage emerges for the  $\approx 31%$  of queries requiring website visits.<sup>†</sup>

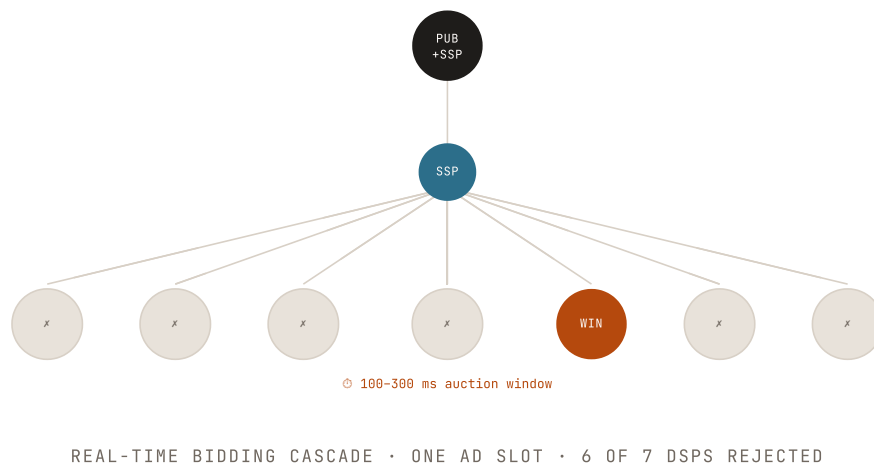
**Key insight:** A growing share of SERP resolutions now occur via **Google AI Overviews**, which synthesise results using an LLM *on top* of the traditional search query. The canonical 0.3 Wh Google baseline therefore systematically *understates* the true energy cost of modern search for any query that 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.



### 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.<sup>†</sup>

**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<sub>2</sub> 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  $3.41 \times 10^{-7}$  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. This server-side ad-tech estimate is deliberately excluded from the quantitative session totals in §6 and Appendix A, which reflect only the primary search inference and modelled network and client-side energy. Including this lower bound would increase the Scenario B search-session total from 2.41 Wh to approximately 2.71 Wh, widening the central ratio from 5.4× to approximately 6.0×. We therefore treat this overhead as a qualitative margin of safety rather than a modelled input.

§ 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 two convergent 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]); 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.

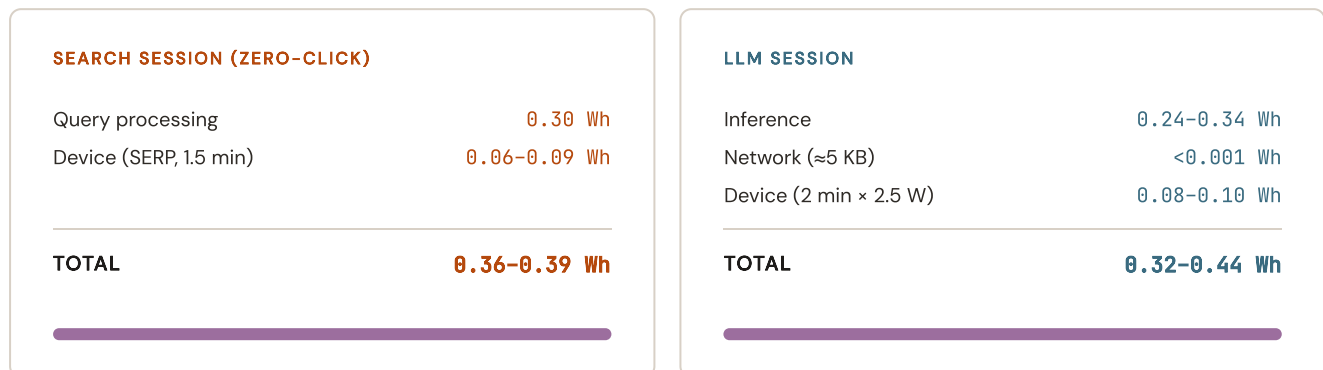
*Functional unit: the complete user session required to satisfy one complex information need, defined as a task requiring synthesis or comparison of information from multiple sources.*

SCENARIO A · ZERO-CLICK · MOBILE / WI-FI

## Simple fact query

*Who is the current prime minister of Italy?*

— A single-answer factual lookup resolved instantly on the SERP without clicking through.



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

≈ 1.1×

SCENARIO B · CORE FINDING · MOBILE / 5G

## Complex synthesis task

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

— A multi-page research session navigating ad-heavy content on a mobile cellular connection.

### 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. 2024a 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**

LLM session is approximately 5.4× more energy-efficient.

Range: 3.7–7.1× across primary sensitivity bounds (see §7, §9.5)

≈ 5.4×

SCENARIO C · UPPER BOUND · LAPTOP / WI-FI

## Extended research session

*Summarise the comparative energy policies of the EU and China for a policy briefing.*

— A deep-dive synthesis session spanning five pages across mixed Wi-Fi and mobile data.

### SEARCH SESSION (LAPTOP)

Query	0.30 Wh
Mobile network (3 pp)	1.08 Wh
Wi-Fi network (2 pp)	0.03 Wh
Page rendering	1.25 Wh
Ad payload	0.31 Wh
Reading (12 min × 10 W)	2.00 Wh

**TOTAL** **4.97 Wh**

### LLM SESSION (LAPTOP)

Inference	0.46 Wh
Reading (5 min × 10 W)	0.83 Wh

**TOTAL** **1.29 Wh**

LLM is ≈ 3.9× more efficient. The higher baseline power of the laptop narrows the advantage compared to mobile, but faster task completion preserves a strong efficiency gap.

≈ 3.9×

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

TABLE 1: PARAMETER ESTIMATES, UNCERTAINTY RANGES, AND PRIMARY SOURCES FOR CELCA SCENARIOS.<sup>†‡</sup>

**LLM inference:** The central value (0.30 Wh) is deliberately set above the empirical point estimate of arXiv:2508.15734 (0.24 Wh) to account for multi-provider variance across frontier model classes and to ensure scenarios remain conservative. The directional finding is robust to this choice. The upper bound (0.55 Wh) is set as a conservative ceiling above the highest disclosed production figure, to stress-test the model under adverse inference assumptions.

**Search query server energy:** The central value (0.30 Wh) is deliberately retained as a conservative ceiling. More recent independent estimates, including [Vanderbauwhede \(2025\)](#), which recalibrates the 2009 Google disclosure against documented PUE improvements and hardware efficiency gains, situate the figure closer to 0.03–0.04 Wh. Substituting this lower bound shifts the web session total from 2.41 Wh to approximately 2.15 Wh, and the central ratio from 5.4x to approximately 5.3x. The directional finding is unchanged: server-side query energy represents under 2% of the corrected total, which remains dominated by mobile network transmission (45%) and client-side rendering with ad overhead (33%).

### 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 · SCENARIO B CENTRAL ESTIMATES

LLM INFERENCE <b>0.30 Wh</b>	NETWORK INTENSITY <b>0.14 kWh/GB</b>	PAGE WEIGHT <b>2.56 MB</b>
PAGES VISITED <b>3</b>	AD PAYLOAD <b>30%</b>	READING TIME <b>6 min</b>

---

SEARCH SESSION	LLM SESSION	EFFICIENCY RATIO
<b>2.41 Wh</b>	<b>0.40 Wh</b>	<b>6.0×</b>

#### MONTE CARLO SENSITIVITY ANALYSIS 10,000 ITERATIONS · 9 FREE PARAMETERS

MEAN RATIO	10TH PCTL	90TH PCTL	MIN OBSERVED
<b>5.4×</b>	<b>3.2×</b>	<b>9.0×</b>	<b>1.6×</b>

*Across all 10,000 parameter combinations, no scenario produces search energy ≤ LLM energy. The efficiency floor of 1.6× occurs with minimal configuration: 2 pages visited, lowest network intensity, maximum LLM inference cost.*

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 boundary constraint*

The efficiency inversion described in this assessment is fundamentally a **mobile phenomenon**. On fixed Wi-Fi (0.006 kWh/GB), network transfer energy collapses by over 95%, reducing the LLM efficiency advantage to approximately 1.5–2.5× for complex tasks and reaching parity for simple queries. The macro-argument therefore scales with global cellular usage share – approximately 60% of all web sessions as of 2025 ([GSMA 2025](#)). The inversion holds at population scale, but the headline efficiency ratio should not be generalised to desktop or Wi-Fi-primary environments without adjustment.

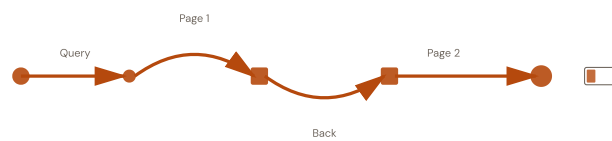
§ 08

## 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.

### SEARCH · COMPOUNDING



### LLM · SYNTHESIS



INFORMATION RETRIEVAL WORKFLOWS: SEARCH (COMPOUNDING) VS. LLM (SYNTHESIS)

§ 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	= 2.41 Wh
<b>HYBRID LLM ADVANTAGE EVEN WITH VERIFICATION</b>	<b>≈ 2.0×</b>

### 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 programmatic web retrieval with LLM inference, **test-time compute (TTC)** architectures that extend inference through chain-of-thought generation, or multi-turn conversations consuming reasoning tokens implicitly.

This risk mirrors a fundamental thermodynamic constraint of deep learning documented by [Yang et al. \(2024\)](#), where a tenfold (10×) increase in energy yields only a marginal **~3% increase in model accuracy** at the frontier. This "accuracy-at-all-costs" regime is rapidly arriving in LLM inference. Harvard's audit of TTC protocols [Jin et al. \(2025\)](#) recorded an average **4.4× token explosion**. The thermodynamic cost of this expansion is severe: [Oviedo et al. \(2026\)](#) estimate that long reasoning queries (~5,000 tokens) consume a median of **3.91 Wh**, a 13× increase over standard queries.

We can invert the analysis to locate the thermodynamic breaking point: at what reasoning-token expansion factor does the LLM efficiency advantage disappear entirely? Depending on the base inference assumptions applied from Table 1, the model identifies a crossover threshold between roughly **4×** and **8×** against the mobile search baseline. [Jin et al. \(2025\)](#) independently report a mean production reasoning expansion of **4.4×**, with more demanding cases reaching **10×** and, in the extreme, **113×**. The convergence here is structural, not

statistical: the mechanistically derived parity threshold sits within the observed operating envelope of current reasoning models. The efficiency advantage of standard LLM inference over mobile web search cannot be assumed to extend to reasoning workflows.

## THE CLIFF, NOT THE SLOPE

---

It is tempting to model the reasoning energy penalty as a proportional, gradual cost. The thermodynamic reality is a phase transition. Below approximately 3× token expansion, the LLM retains a meaningful efficiency advantage across the vast majority of modelled scenarios. Between 3× and 10×, that advantage collapses into parity or net loss. Above 10×, a threshold documented in extreme production cases, the LLM becomes substantially more energy-intensive than the web session it replaces. This is not a slope. It is a regime change. The efficiency narrative surrounding standard LLM inference cannot be straightforwardly extended to reasoning or agentic workflows; they constitute a categorically distinct, highly emissive operating regime. This reinforces the absolute necessity of **difficulty-aware model routing** as a first-order sustainability intervention.

### 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 *The "conservative baseline" validation*

Early drafts of this assessment relied on Google's 0.24 Wh figure, which invited skepticism regarding vendor self-interest. However, independent peer-reviewed evidence [Oviedo et al. \(2026\)](#) now confirms that production inference costs are indeed substantially lower than widely cited historical estimates, validating our use of a low baseline. By adopting 0.30 Wh as our central estimate, our model deliberately builds in a margin of safety against multi-provider variance. **Even if inference costs rise to 0.55 Wh, the Scenario B efficiency advantage persists at approximately 3.7×.** On the search side, [Vanderbauwhede \(2025\)](#) recalibrates the 2009 Google server figure downward to approximately 0.04 Wh, accounting for PUE improvements and hardware efficiency gains. Substituting this value shifts the web session total from 2.41 Wh to 2.15 Wh and the central ratio from 5.4× to 5.3×, leaving the directional finding intact.

## 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*

This distinction is now entering institutional analysis. In their 2026 joint report, the French telecom regulator (ARCEP), working with PEReN, highlights that evaluating generative AI requires moving beyond isolated query metrics toward the total energy required to deliver the service rendered, including the advertising-auction overhead of traditional search alternatives. Imposing unit-energy taxes on LLM queries without accounting for this full-stack alternative-use-case risks 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.

## 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 narrow to parity for simple queries on Wi-Fi, and reverse for reasoning-model inference or 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: redirecting complex synthesis tasks from mobile browsers to standard LLM interfaces represents a materially more efficient workflow. At population scale, the efficiency delta is significant. As an illustrative scenario, if 500 million daily informational queries were to shift from ad-supported mobile search to standard non-reasoning LLM inference, the implied net energy savings would approach **365 GWh annually**, equivalent to eliminating the full power demand of a mid-sized data centre fleet or a small municipality, through software routing alone.

This finding should inform corporate digital sustainability strategies, regulatory impact assessments, and the emerging discipline of sustainable information retrieval. The ultimate implication of this assessment is not merely that LLM queries can be cheaper than search queries. It is that the dominant paradigm of the ad-supported, multi-page mobile web carries a structural, largely invisible energy tax – and that a plausible near-term substitution effect would represent a material, system-level efficiency gain. For the class of synthesis tasks identified here, generative AI acts as a **thermodynamic compression engine**, sparing the mobile network and client device from the accumulated burden of the modern web.

§ Ref

## References

---

- Ad Net Zero. (2025, June). *Global Media Sustainability Framework V1.2*. [adnetzero.com](https://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](https://doi.org/10.3390/challe6010117)
- ARCEP–PEReN. (2026, May). *Intelligence artificielle générative : quels défis environnementaux ?* Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. [arcep.fr](https://arcep.fr)
- 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](https://doi.org/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](https://doi.org/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](https://doi.org/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](https://arxiv.org/abs/2508.15734)
- Epoch AI. (2025, February 7). How much energy does ChatGPT use? *Gradient Updates*. [epoch.ai](https://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](https://gsma.com/mobileeconomy)
- HTTP Archive. (2025, January 16). *Web Almanac 2025: Page Weight Chapter*. [almanac.httparchive.org](https://almanac.httparchive.org)
- Jin, Y., Wei, G.-Y., & Brooks, D. (2025). The energy cost of reasoning: Analyzing energy usage in LLMs with test-time compute. *arXiv:2505.14733*. [arxiv.org/abs/2505.14733](https://arxiv.org/abs/2505.14733)
- 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](https://doi.org/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](https://doi.org/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](https://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](https://doi.org/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](https://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](https://ml.energy/leaderboard)
- Muxup. (2026, January). Estimating the energy consumed by DeepSeek R1 inferences. [muxup.com](https://muxup.com)
- MedUX. (2026, January 15). *France 5G QoE crowdsourcing benchmark – Q3 2025*. [medux.com](https://medux.com)
- Nokia. (2019). *How 5G is bringing an energy efficiency revolution* [White paper]. Nokia Corporation.
- Oliveira, A. P., Carraquico, T., & Martinez-Perez, C. (2026). Beyond efficiency: A systematic review of energy consumption and carbon footprint across the AI lifecycle. *Sustainability*, 18(3), 1359. [doi:10.3390/su18031359](https://doi.org/10.3390/su18031359)
- Ookla & Omdia. (2026). *A global reality check on 5G SA and 5G Advanced in 2026*. [ookla.com](https://ookla.com)
- Oviedo, F., et al. (2026). Energy use of AI inference, efficiency pathways, and test-time scaling. *Joule*, 10, 102430. [doi:10.1016/j.joule.2026.102430](https://doi.org/10.1016/j.joule.2026.102430)
- Patterson, D., et al. (2021). Carbon emissions and large neural network training. *arXiv:2104.10350*. [arxiv.org](https://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/jieec.13376](https://doi.org/10.1111/jieec.13376)
- Scope3. (2023, April 19). *Q1 2023 State of Sustainable Advertising*. [scope3.com](https://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](https://doi.org/10.1145/3706598.3714082)
- Search Engine Land. (2025, December 3). Google AI Mode sends traffic on 69% of transactional queries. [searchengineland.com](https://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](https://doi.org/10.18653/v1/P19-1355)
- TokenPowerBench. (2025). Benchmarking the power consumption of LLM inference. *arXiv:2512.03024*. [arxiv.org](https://arxiv.org)
- Vanderbauwhede, W. (2025, January 6). Estimating the increase in emissions caused by AI-augmented search. *arXiv:2407.16894v2*. University of Glasgow. [arxiv.org/abs/2407.16894](https://arxiv.org/abs/2407.16894)
- Yang, Z., Adámek, K., & Armour, W. (2024). Double-exponential increases in inference energy: The cost of the race for accuracy. *arXiv:2412.09731*. [arxiv.org/abs/2412.09731](https://arxiv.org/abs/2412.09731)
- Zhang, W., et al. (2025). Energy efficient or exhaustive? Benchmarking power consumption of LLM inference. *HotCarbon 2025*. [hotcarbon.org](https://hotcarbon.org)

## § Appendix A

## Session energy summary

Central-estimate energy budgets for all three scenarios. 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 / 1.11 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	0.83 / 2.00 Wh
<b>TOTAL SESSION</b>	<b>0.33 Wh</b>	<b>0.38 Wh</b>	<b>0.40 Wh</b>	<b>2.41 Wh</b>	<b>1.29 / 4.97 Wh</b>
<b>Efficiency ratio (Search/LLM)</b>		<b>≈ 1.1×</b>		<b>≈ 5.4×</b>	<b>≈ 3.9×</b>

TABLE A1: SESSION ENERGY BREAKDOWN BY SCENARIO AND MODALITY (CENTRAL ESTIMATES). SENSITIVITY RANGES: §7.

INTERACTIVE BREAKDOWN OF ENERGY DISTRIBUTION ACROSS THE SYSTEM BOUNDARY FOR EACH SCENARIO

## § Appendix B

## Mathematical formulation

---

Let  $E$  represent the total energy required to satisfy a single complex information need.

### 1. THE LLM MODALITY

The total energy is a function of a single, heavy server inference, negligible payload transmission, and brief device display time:

$$E_{\text{LLM}} = E_{\text{server}}^{\text{LLM}} + E_{\text{net}}^{\text{LLM}} + E_{\text{device}}^{\text{LLM}}$$

### 2. THE WEB SEARCH MODALITY

Where  $n$  is the number of page interactions (queries, clicks, back-navigations) required to satisfy the user's intent. Note that device screen energy  $E_{\text{device}}^{\text{Search}}$  is modelled as a continuous session-level cost rather than per-page:

$$E_{\text{Search}} = \sum_{i=1}^n \left( E_{\text{server},i}^{\text{Search}} + E_{\text{net},i}^{\text{Search}} + E_{\text{render},i} + E_{\text{ads},i} \right) + E_{\text{device}}^{\text{Search}}$$

### 3. THE EFFICIENCY INVERSION CONDITION

The traditional narrative assumes  $E_{\text{server}}^{\text{LLM}} \gg E_{\text{server}}^{\text{Search}}$ . The thermodynamic inversion occurs when the systemic, multiplicative costs of the web search journey exceed the heavier initial cost of the LLM:

$$E_{\text{LLM}} < E_{\text{Search}}$$

---

### PHYSICAL EXPANSIONS:

$E_{\text{net}} = I_{\text{net}} \times D$  (where  $I_{\text{net}}$  is network intensity in Wh/GB and  $D$  is data payload in GB).

$E_{\text{device}} = P_{\text{dev}} \times t$ , where  $P_{\text{dev}}$  is power draw (W) and  $t$  is duration (h), yielding Wh.