Promoting Sustainable Web Agents: Benchmarking and Estimating Energy Consumption Through Empirical and Theoretical Analysis

Lars Krupp[1, 2], **Daniel Geißler**[1, 2], **Vishal Banwari**[1, 2], **Paul Lukowicz**[1, 2], **Jakob Karolus**[1, 2]
[1]RPTU Kaiserslautern-Landau
[2]German Research Center for Artificial Intelligence (DFKI)
lars.krupp@dfki.de

Abstract

Web agents, like OpenAI's Operator and Google's Project Mariner, are powerful agentic systems pushing the boundaries of Large Language Models (LLM). They can autonomously interact with the internet at the user's behest, such as navigating websites, filling search masks, and comparing price lists. Though web agent research is thriving, induced sustainability issues remain largely unexplored. To highlight the urgency of this issue, we provide an initial exploration of the energy and CO_2 cost associated with web agents from both a theoretical —via estimation— and an empirical perspective —by benchmarking. Our results show how different philosophies in web agent creation can severely impact the associated expended energy, and that more energy consumed does not necessarily equate to better results. We highlight a lack of transparency regarding disclosing model parameters and processes used for some web agents as a limiting factor when estimating energy consumption. Our work contributes towards a change in thinking of how we evaluate web agents, advocating for dedicated metrics measuring energy consumption in benchmarks.

Code — https://github.com/DFKIEI/WebAgentEnergy

Introduction

Web agents powered by large language models (LLMs) represent the next significant milestone in how we interact with the internet (GoogleDeepMind 2025; OpenAI 2025). These systems are able to "browse the web" and interact with online environments in a manner similar to human users. This concept holds immense potential to revolutionize internet usage, potentially even replacing traditional web browsers as the primary means of accessing information. However, the substantial computational costs (Samsi et al. 2023a) of these systems still remain a significant challenge.

There is an ongoing discussion on the sustainability and environmental impact of developing and deploying LLMs (Bender et al. 2021), which are core components in any web agent. OpenAI's GPT-3 has 175 billion parameters and is trained on 570 GB of data (Brown 2020), using tremendous amounts of resources for training and inference. This necessitates the creation of large-scale data centers with substantial energy consumption.

Companies are addressing this challenge in strikingly different ways. Some have opted to simply expand their energy resources. Google, for instance, is investing in the construction of nuclear power plants to support its data centers (da Silva 2024) — a controversial and arguably unsustainable approach. Others are taking a more forward-thinking route by introducing reporting standards for the energy consumption across the lifecycle of LLMs, cf. Mistral (MistralAI 2025). These initiatives aim to promote transparency, encourage sustainable development, and incentivize more energy-efficient implementations.

LLM-heavy applications like web agents are among the most computationally intensive AI systems. Yet, this reality is largely invisible to end users. Tools such as OpenAI's Operator or Google's Project Mariner present themselves as simple input fields. Indistinguishable — on the surface — from a search bar or any other interface powered by a commodity LLM. There is no immediate feedback to users about the energy consumption or environmental impact of their queries. The cost remains abstract, detached from the interaction itself. In the background, however, the infrastructure required to support such systems includes high-emission data centers. As more and more users adopt web agents, their cumulative energy impact will become significant. There is a pressing need to bring this aspect to the forefront and to assess web agents not just on their performance, but also on their energy efficiency, rewarding efficient implementations. This work highlights this critical gap by quantifying the energy consumption and CO_2 emissions of six web agents through both theoretical means and an empirical evaluation.

In our empirical evaluation, we directly benchmark the energy consumption of five web agents that use open-source LLMs on eight different GPUs using the Mind2Web benchmark (Deng et al. 2024). Our procedure allows for fast, simple and efficient energy benchmarking for any web agent using open-source LLMs and directly compares the web agents by their real energy consumption compared to their performance results. Our results show that more energy consumed does not equate to better results. The most energy efficient web agent, AutoWebGLM (Lai et al. 2024), also performed best in terms of average step success rate (SSR), a popular metric for gauging web agent performance.

Our theoretical approach proposes a method to estimate the energy consumption web agents using propri-

etary LLMs. For such agents, benchmarking the energy consumption is not possible, forcing us to rely on information present in literature. We applied this estimation approach to LASER (Ma et al. 2023), a web agent using the proprietary LLM GPT-4 and MindAct (Deng et al. 2024) as a counterexample. MindAct is a web agent using open-source LLMs and smart preprocessing for computational efficiency. Additionally, we benchmarked MindAct, acquiring its exact energy costs, thus allowing us to compare it to its theoretical estimation. By estimating and comparing the amount of energy consumed by these web agents with vastly different design philosophies, we highlight the impact of a web agent's design on its energy consumption. Using a conservative estimation, LASER spends approximately 10 times more energy than MindAct.

To quantify the effectiveness of estimating web agent energy consumption we evaluated MindAct (Deng et al. 2024) both using benchmarking and estimation. Our results shine a light on the impact that uncertainty has on the estimation of this metric. Even for MindAct, a fully open-source agent, we overestimate its energy consumption by a factor of 7. We conclude that the estimation of web agent energy consumption should only be considered when benchmarking is not possible.

Our work encourages a change in thinking of how we evaluate web agent performance. We show the stark differences in energy consumption between web agents and propose metrics to report (for proprietary LLM-driven agents) and a benchmark to use (for open-source LLM-driven agents) which allow comparison of their energy efficiency. We advocate for a holistic evaluation of web agents incorporating dedicated metrics for energy consumption.

Related Work

With the rapid improvements of LLMs in recent years and their ever improving capabilities in tool-use (Dubey et al. 2024) a new frontier of research has become possible. With the goal of building agents that can interact with the internet much like a human would, web agents recently are gaining traction (Deng et al. 2024; Yao et al. 2022). Approaches in web agent construction show a high diversity, varying in input modalities, processing steps, and employed LLM. Some only use HTML for their input (Ma et al. 2023; Deng et al. 2024), others use the accessibility tree instead (Chezelles et al. 2024), supplement it with screenshots (Zheng et al. 2024) or use screenshots exclusively, like Pix2Act (Lù, Kasner, and Reddy 2024). Further characteristics include preprocessing (Deng et al. 2024; Gur et al. 2023), and the integration of memory modules (Ma et al. 2023). While some agents use open-source models (Deng et al. 2024; Gur et al. 2023), most use proprietary models (Ma et al. 2023; Zheng et al. 2024; Yang et al. 2024; Zhang et al. 2024).

Likewise, many different benchmarks to evaluate their performances are being proposed in rapid succession (Yao et al. 2022; Deng et al. 2024; He et al. 2024). While efforts to unify these benchmarks exist (Chezelles et al. 2024), comparing the performance of different web agents is challenging. Additionally, no benchmark yet takes the energy consumption of web agents into consideration and penalizes inefficient agents.

Since modern web agents are driven by LLMs, contextualizing their growing emissions is vital. In the earlier stages of LLM development, research estimated the CO_2 emissions of, at the time state-of-the-art, transformer models such as BERT and GPT-3. For BERT (Devlin 2018), trained in the US, Devlin et al. calculated a potential environmental impact of 0.754 metric tons of CO_2 for a single training of 79h on 64 Tesla V100 GPUs with an average utilization of 62.7%. For GPT-3, the predecessor of ChatGPT, it is estimated that around 550 metric tons of CO_2 emissions were produced to complete the full training, based on the US American energy mix, tremendously exceeding the previous estimations from BERT due to increased complexity and dataset size (Shi et al. 2023). On top, a significant amount of energy is commonly wasted on ineffective versions of the LLM and for tuning the hyperparameter spaces (Verdecchia, Sallou, and Cruz 2023).

After training, the deployment of LLMs to the public introduces an additional and significant layer of environmental impact. According to Samsi et al. (Samsi et al. 2023a), the energy demand during inference is influenced by several unpredictable factors, including user load and deployment duration. They propose *energy per token* as a useful metric for evaluating the trade-off between performance and sustainability, particularly for quantized LLMs. Additionally, the environmental impact of LLMs extends beyond energy consumption to include resource use, such as the water needed for cooling data centers, and the e-waste generated by the disposal of outdated hardware, highlighting the complexity and difficulty of calculating and comparing the LLM carbon footprint throughout the whole life cycle (Patterson et al. 2021).

To support the development of more sustainable web agents, our work focuses specifically on the energy costs incurred during inference, addressing a critical component of their broader impact on the environment. We benchmark the energy consumption and carbon footprint of web agents and demonstrate that, due to lack of transparency, estimating the energy usage of proprietary LLM-driven agents is unreliable.

Energy Consumptions of Web Agents

Sustainability is becoming an increasingly pressing concern in the development and deployment of generative AI systems, particularly as web agents gain broader adoption. In this work, we aim to provide insights into the energy efficiency of several prominent web agents introduced in recent research. We solely focus on the energy efficiency during inference as it will outweigh fine-tuning cost with continued use and widespread use of web agents. Our approach is twofold. First, we conduct a practical evaluation of five web agents, all of which rely on open-source LLMs and are thus available for benchmarking, reflecting realistic deployment scenarios. Second, we present a theoretical estimation of energy consumption for one of the benchmarked agents and an additional agent that uses a proprietary model, where benchmarking is infeasible due to the lack of open-source access.

By including a benchmarked agent in the estimation process, we are able to assess the accuracy of our theoretical estimation and demonstrate its applicability in case empirical measurements are not possible. Table 1 lists all web agents that we evaluated in this paper. We note that completely closed-source agents such as OpenAI's Operator (OpenAI 2025) and Google's Project Mariner (GoogleDeepMind 2025) are impossible to estimate as no details of their implementation are available.

Web Agent	LLM	BM	Est.
Fully open-source web agents			
AutoWebGLM (Lai et al. 2024)	ChatGLM3-6B	✓	
MindAct (Deng et al. 2024)	DeBERTa-86M, flan-T5$_{XL}$ -2.85B	✓	✓
MultiUI (Liu et al. 2024)	UIX-Qwen2-8.03B	✓	
Synapse (Zheng et al. 2023)	CodeLlama-Instruct-7B	✓	
Synatra (Ou et al. 2024)	CodeLlama-7B	✓	
Open-source web agent with proprietary LLM			
LASER (Ma et al. 2023)	GPT-4		✓

Table 1: Web agents evaluated in this paper. BM abbreviates benchmark. Est. abbreviates Estimate.

Empirical Evaluation through Benchmarks

The most precise way of measuring the sustainability of an open-source LLM-driven web agent is to benchmark its energy consumption. While benchmarking methods for the energy consumption of LLMs have been established (Samsi et al. 2023b), those methods are not sufficient in our case as they do not take into account how effectively the web agent navigates its environment. We chose the Mind2Web benchmark (Deng et al. 2024) since it is easy to set up (no external server setup that has to be accessed), guarantees comparability (tasks are not changing, in contrast to live benchmarks), and is one of the most popular benchmarks when evaluating web agents (Krupp et al. 2025). In contrast to other web agent benchmarks (Yao et al. 2022; Liu et al. 2018), Mind2Web consists of real-world websites. This allows for a realistic evaluation with respect to the average number of tokens per website. Mind2Web consists of 2350 tasks on 137 websites in 31 domains, with an average number of actions needed for task completion of 7.3 and an average of 1135 HTML elements per website. The tasks are distributed over three splits: cross-domain (across task and environment generalization), cross-task (across task, same environment generalization), and cross-website (across website, same domain generalization).

We selected five popular web agents (see Table 1) that are fully open-source (including the LLM), reproducible and already utilize Mind2Web in their evaluation.

Setup We conducted the benchmarking on our cluster equipped with a variety of commonly employed GPUs for AI and Machine Learning, as detailed in Table 2. Additionally, we ran each web agent five times on each GPU to ensure stable results. To retrieve the energy consumption of the GPUs, we modified the original web agent code to include the carbontracker library (Anthony, Kanding, and Selvan 2020). By setting a start and end flag using this library, we can acquire the actual energy consumed by the executed code on each connected GPU.

GPU Model	Architecture	VRAM	FP32 (TFLOPS)
A100-SXM4	Ampere	40 GB	19.5
A100-PCIe	Ampere	40 GB	19.5
RTX A6000	Ampere	48 GB	38.7
RTX 3090	Ampere	24 GB	35.6
H100-SXM5	Hopper	80 GB	67
H100-NVL	Hopper	94 GB	60
H200-SXM5	Hopper	141 GB	67
L40S	Ada Lovelace	48 GB	91.61

Table 2: NVIDIA GPUs used in our setup, cf. (TechPowerup 2025).

Results Our results show large differences in the energy consumption of different agents and GPUs. Clear trends become visible when analyzing Figure 1, allowing for an ordering of web agents by energy consumption from most efficient to least efficient.

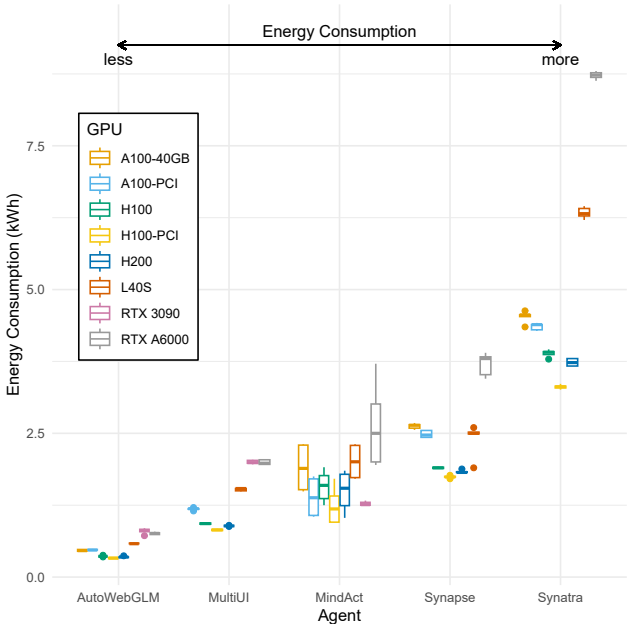

Figure 1: Energy consumption per web agent and GPU.

On average, the most energy-efficient GPU in our benchmarking tests was the Nvidia H100-NVL, a popular GPU for AI-related computing tasks. Hence, we choose to conduct further agent comparisons on this GPU only.

We found large differences in energy consumption between the most efficient web agent and the most inefficient agent (see Table 4). Synatra (Ou et al. 2024) consumes ten times more energy compared to AutoWebGLM (Lai et al. 2024). However, the increased consumption does not lead to better results on the Mind2Web benchmark as reported by the agents' average step success rate (SSR). As such, AutoWebGLM is not only the most energy-efficient web agent, but it also performs best on the Mind2Web benchmark.

The SSR is the de facto standard metric of the Mind2Web benchmark (Krupp et al. 2025), representing the ratio of successful steps towards the correct solution divided by the total steps. Since the Mind2Web benchmark is divided into three splits, each reporting a separate SSR, an average SSR (Deng et al. 2024) is calculated at the end.

Within these splits, cross-domain contains the most tasks, followed by cross-task and cross-website. To account for the different sizes, we calculated the average energy spent per token for each web agent and for each split on the Nvidia H100-NVL GPU (see Table 3). Our results show small differences between the splits for the same web agent. Additionally, it shows the importance of reducing the amount of ingested tokens into the LLM for the overall energy consumption. While the energy per token was consistently highest for AutoWebGLM, the overall energy consumed was lowest due to AutoWebGLM's preprocessing, which allowed the agent to significantly reduce the total amount of processed tokens. The energy per token metric is mostly influenced by LLM size, different tokenizers and different internal workings of the agents.

Theoretical Estimation

Many web agents rely on proprietary LLMs, preventing direct energy measurements through benchmarking. In this section, we propose an approach using theoretical estimation of their energy consumption based on available literature. As such, it is important to have detailed knowledge about the internal workings of each individual web agent to expose computationally inefficient implementations, making even this approach not feasible for completely closed-source agents such as OpenAI's Operator and Google's Project Mariner. However, if only the LLM is closed-source (proprietary), a theoretical estimation is possible if we can access information on the agent's internal procedure, e.g., through publications.

With MindAct (Deng et al. 2024) and LASER (Ma et al. 2023), we chose two web agents differing in many aspects to provide an example of the diversity in approaches and mentalities present in the field. While MindAct uses comparatively small open-source models and does extensive preprocessing to get the best possible performance out of the available resources, LASER uses a proprietary model at its core with minimal preprocessing being done. To estimate the energy consumption for each model, we analyzed and dissected available resources such as the accompanying publications (Deng et al. 2024; Ma et al. 2023) and the publicly available source code[1].

MindAct MindAct (Deng et al. 2024) divides the process of finding the next correct *action* to execute on the web into two stages, as depicted in Figure 2. The first stage, candidate generation, is treated as a ranking task. Here, their finetuned DeBERTa 86M (He, Gao, and Chen 2021) model is given the following *input*: the user's query, previous actions and a cleaned representation of each element in the Document Object Model (DOM) of the HTML webpage. Each cleaned element consists of its tag, its textual content, its salient attribute values and a textual representation of its respective parent and child nodes. From this, DeBERTa estimates a matching value MS_i between 0 and 1 — indicating how well an element matches the given user query. This process is repeated for all elements in the DOM. We estimate that the total number of tokens processed by DeBERTa at the end of this process is at least equivalent to the number of tokens in the original HTML. After processing all elements, the 50 elements with the highest matching values are used in the second stage. The second stage, action prediction, is constructed as a multiple-choice question answering challenge. A finetuned flan-T5$_{XL}$ model (Chung et al. 2022) is given the user query, five of the returned elements and a none element and is tasked to decide which element is most likely to help towards answering the user query (the *action* to perform). This is done a minimum of 10 times until all 50 returned elements are processed. The process is repeated if more than one possible action (besides none) is returned, until only one action remains (or all are rejected). For our estimation, we assume the maximum input length of flan-T5$_{XL}$ (Raffel et al. 2020) (512 tokens) for one multiple-choice question and that a final result is obtained after the first pass (querying flan-T5$_{XL}$ 10 times).

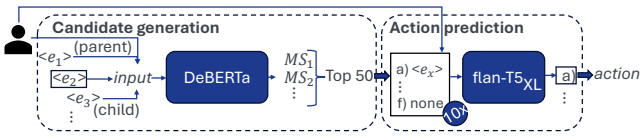

Figure 2: Pipeline depicting how an action is chosen in MindAct.

To estimate the amount of energy per action for MindAct, we need to acquire (1) the energy per token for its LLMs, DeBERTa and flan-T5$_{XL}$, and (2) an estimation of the LLMs' context sizes (number of tokens passed to the LLM). Note that we already fixed $\overline{N}_{flan-T5_{XL}} = 512$, the maximum input length of flan-T5$_{XL}$. For DeBERTa, we calculated the average number of tokens contained within an HTML page for the Mind2Web benchmark (Deng et al. 2024) to be $\overline{N}_{DeBERTa} = 118798$ using its tokenizer (He, Gao, and Chen 2021). We extracted the energy per token for both models from our benchmark results (Nvidia H100-NVL), yielding $e_{DeBERTa} = 3.77 \cdot 10^{-6}\,Wh$ and $e_{flan-T5_{XL}} = 9.08 \cdot 10^{-6}\,Wh$. Resulting in the following

[1]MindAct: https://github.com/OSU-NLP-Group/Mind2Web, LASER: https://github.com/Mayer123/LASER

Agent	Split	# 10^6 Tokens	Energy (kWh)	Energy/Token (kWh)
AutoWebGLM	cross-domain	0.25	0.22 ± 0.007	$(890 \pm 28.3) \times 10^{-9}$
	cross-task	0.09	0.06 ± 0.004	$(719 \pm 46.42) \times 10^{-9}$
	cross-website	0.06	0.05 ± 0.004	$(831 \pm 69.28) \times 10^{-9}$
MindAct	cross-domain	183.78	0.66 ± 0.152	$(3.59 \pm 0.08) \times 10^{-9}$
	cross-task	78.27	0.35 ± 0.086	$(4.47 \pm 1.10) \times 10^{-9}$
	cross-website	49.60	0.21 ± 0.066	$(4.23 \pm 1.33) \times 10^{-9}$
MultiUI	cross-domain	1.59	0.52 ± 0.008	$(326 \pm 5.04) \times 10^{-9}$
	cross-task	0.65	0.18	$(280 \pm 6.15) \times 10^{-9}$
	cross-website	0.40	0.12	$(307 \pm 10.06) \times 10^{-9}$
Synapse	cross-domain	6.88	1.07 ± 0.018	$(156 \pm 2.62) \times 10^{-9}$
	cross-task	2.97	0.42 ± 0.004	$(142 \pm 1.35) \times 10^{-9}$
	cross-website	1.93	0.25 ± 0.004	$(131 \pm 2.07) \times 10^{-9}$
Synatra	cross-domain	24.34	2.11 ± 0.027	$(86.7 \pm 1.11) \times 10^{-9}$
	cross-task	8.92	0.72 ± 0.008	$(80.9 \pm 0.9) \times 10^{-9}$
	cross-website	5.50	0.48 ± 0.004	$(86.9 \pm 0.73) \times 10^{-9}$

Table 3: Mean energy consumption per benchmark split and energy per input-token level for the H100-NVL GPU; the total number of tokens is dependent on the LLM's tokenizer.

Agent	Ø SSR	Energy (kWh)	Time (min)
AutoWebGLM	**53.53**	$\mathbf{0.33 \pm 0.01}$	$\mathbf{57.0 \pm 0.8}$
MindAct	43.50	1.22 ± 0.29	296.0 ± 90.2
MultiUI	34.70	0.82 ± 0.01	130.0 ± 1.2
Synapse	21.67	1.74 ± 0.02	356.0 ± 2.8
Synatra	15.85	3.31 ± 0.04	426.0 ± 1.4

Table 4: Energy consumption, computation time and reported average step success rate (SSR) per web agent on the Nvidia H100-NVL GPU.

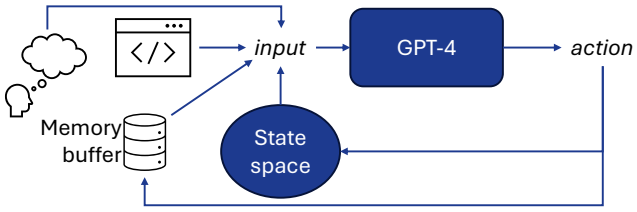

Figure 3: Pipeline depicting how an action is chosen in LASER.

energy per action for MindAct:

$$
\begin{aligned}
E_{action} &= E_{candidate\ generation} + E_{action\ prediction} \\
E_{action} &= (\overline{N}_{DeBERTa} \cdot e_{DeBERTa}) \\
&\quad + 10 \cdot (\overline{N}_{flan-T5_{XL}} \cdot e_{flan-T5_{XL}}) \\
E_{action} &= \underline{\underline{0.49}}\, Wh
\end{aligned} \tag{1}
$$

Since the Mind2Web benchmark consists of 2350 tasks with an average of 7.3 actions per task, the total energy consumption E_{total} equates to:

$$E_{total} = E_{action} \cdot 7.3 \cdot 2350 = \underline{\underline{8.5}}\, kWh \tag{2}$$

LASER In contrast to MindAct, LASER (Ma et al. 2023) makes use of a proprietary language model, specifically GPT-4 (et al. 2024). LASER introduces states and state transitions for web agents (see Figure 3), allowing for better recovery from mistakes and restricting the possible *actions* depending on the state of the agent. LASER uses one-shot prompting and makes the model think step-by-step to improve the model's capabilities when dealing with complex user queries. Additionally, LASER has access to a memory buffer to store and access intermediary results (previous *actions*). Finally, LASER is forced to produce a result after a maximum of 15 actions were generated. However, the authors do not specify explicitly what the *input* of their web agent is. We inferred that they use the raw, unmodified HTML by analyzing their results and LASER's benchmark: WebShop (Yao et al. 2022).

Analogously to MindAct's energy estimation, we need to acquire the energy per token for LASER's LLM, GPT-4, and an estimate for its context size. We calculated the average number of tokens within an HTML page for the Mind2Web benchmark using the GPT-4 (et al. 2024) tokenizer at $\overline{N}_{GPT-4} = 93778$ tokens. Since GPT-4 is not open-source, we cannot execute the LLM and benchmark its energy consumption per token.

Consequently, estimating the energy consumption of GPT-4 on a per-token basis is inherently challenging due to a lack of publicly available technical specifications. The following estimation is based on leaked data and experts heuristics about OpenAI's GPT-4 model. Its exact size is unknown, though it is estimated to have 1.8 trillion parameters. More-

over, GPT-4 is reportedly using a mixture-of-experts (MoE) architecture comprised of 16 experts, each with around 111 billion parameters. However, in each forward pass only two of these are active at a time (Schreiner 2023), leading to about 222 billion active parameters. The size of a model (the active parameters N) directly influences the number of floating-point operations (FLOP) required for inference, and thus its energy footprint. Using a conservative estimation for large parameter sizes (applicable for GPT-4) (Kaplan et al. 2020), results in a compute cost for one forward-pass of $C_{forward} = 2N$. Hence, roughly 444 billion FLOP are needed to compute one token using GPT-4. Given a Nvidia H100 SXM GPU[2] with a theoretical maximum performance for FP8 Tensor Core of $2 \cdot 10^{15}$ FLOP per second (NVIDIA 2024) in a dense configuration, yields a computation time for a single token on a H100 of about $2.22 \cdot 10^{-4} s$.

However, model size is only one component of the overall energy calculation. Other critical factors include GPU power draw (NVIDIA 2023), actual utilization and power draw during inference (Patel et al. 2024), and data center overhead (Butler 2024) (e.g., cooling, power distribution inefficiencies). In this estimation, we assume a best case scenario for GPU performance to provide a lower bound on energy costs. We note that the real energy costs are likely higher.

Given a DGX server of eight H100 GPUs with a maximum power rating of 10.2 kW (NVIDIA 2023) and data center overheads of about 10-20% (Butler 2024) results in max 1.5 kW per GPU. However, literature suggests GPU power consumption during inference is only about 70% (Patel et al. 2024), leaving us with roughly 1 kW power consumption per GPU during inference. Theoretically, this results in an estimated energy consumption per token of approximately $e_{GPT-4} = 0.22\,Ws = 6.17 \cdot 10^{-}5\,Wh$. We note that GPT-4 cannot be run on a single H100 GPU, but rather is run on a full cluster. This, however, does not change our calculation, since the computation time inversely scales with every additional GPU. As such, a DGX server with eight H100 will use eight times the power, but only requires one eighth of the computation time.

$$E_{action} = \overline{N}_{GPT-4} \cdot e_{GPT-4} = \underline{\underline{5.78}} Wh \tag{3}$$

Using Mind2Web task numbers as above, yields E_{total}:

$$E_{total} = E_{action} \cdot 7.3 \cdot 2350 = \underline{\underline{99.21}}\, kWh \tag{4}$$

Comparing MindAct and LASER Due to the lack of relevant information about the energy consumption of LLMs and the challenges involved in gathering reliable information as a third party, our results can only provide a conservative estimation. Using MindAct as an example — estimated at $8.5\,kWh$ and benchmarked at $1.22\,kWh$ — we showcase that a theoretical estimation approach works as a very coarse estimation in terms of orders of magnitude, highlighting the need for greater transparency and standardization in reporting the energy usage of LLMs. The best option still remains relying on actual benchmark results if available. However, our theoretical estimation allowed us valuable insights into the importance of adequate preprocessing and efficient implementations.

When comparing the energy consumption for LASER and MindAct, it becomes clear that the preprocessing done in MindAct is vital. Additionally, the smaller models, DeBERTa and flan-T5$_{XL}$, are much more energy efficient than the large GPT-4 model used in LASER.

Carbon Dioxide Emissions

To calculate the CO_2 emissions of the different web agents on the Mind2Web benchmark, we multiply their energy consumption by the CO_2 emissions per Wh. We provide results for multiple energy mixes in Table 5. CO_2 emissions serve as a good proxy for energy consumption and provide end-users with a more tangible understanding of an agent's energy efficiency.

To further aid in understanding the differences between the agents, we additionally converted the CO_2 emissions into distance traveled by the average car, assuming 248.55 g of CO_2 emissions per kilometer driven[3]. The emissions generated by running AutoWebGLM on the Mind2Web benchmark equate to 0.6 km traveled, assuming US emissions. For Synatra, the least efficient open-source LLM-driven agent, the range of the car equates to 6 km. For running LASER once on the Mind2Web benchmark, one can drive up to 181 km.

Discussion

While web agents are not widely used yet, it is increasingly evident that they will play a central role in how users interact with the internet. Examples like OpenAI's Operator and Google's Project Mariner already signal this paradigm shift. As these systems become more widespread and integrated into everyday web interactions, the importance of scrutinizing their environmental impact grows.

A critical problem is that the energy consumption of web agents is not readily apparent to the end user. Interfaces are often as simple as a search bar or text box, making it impossible for users to distinguish between "good" and "bad" agents, as their energy and environmental costs are hidden (Mazzucato 2024). There's further value in communicating environmental costs directly to the user (Li et al. 2023). Displaying estimated CO_2 emissions per task could help users become aware of the environmental impact of their interactions and guide them toward more sustainable agents. This not only encourages responsible development practices but also empowers users to make informed choices. However, vastly different energy mixes across countries (see Table 5) make comparison difficult and should be considered.

Our results highlight that the differences in energy consumption between web agents are quite extensive. Some

[2]Assuming H100 GPUs clusters (NVIDIA 2023) in OpenAI data centers.

[3]https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle#typical-passenger

Method	Agent	Energy (kWh)	Energy Mix		
			Norway 20g CO2e kWh	US 453g CO2e kWh	Australia 800g CO2e kWh
Benchmarking	AutoWebGLM	0.33	6g	149g	264g
	MindAct	1.22	24g	552g	976g
	MultiUI	0.82	16g	371g	656g
	Synapse	1.74	34g	783g	1392g
	Synatra	3.31	66g	1499g	2648g
Estimation	MindAct	9.01	180g	4081g	7208g
	LASER	99.21	1984g	44942g	79368g

Table 5: The CO_2 emissions of each web agent on the Mind2Web benchmark (Deng et al. 2024) using the energy mixes of Norway, Australia (Lannelongue, Grealey, and Inouye 2021) and the US (U.S. Energy Information Administration 2022).

agents leverage smart implementations and clever use of computational resources, achieving strong performance while consuming significantly less energy. Importantly, our findings show that energy efficiency does not compromise performance. In fact, some of the most efficient agents also perform well on benchmarks. Consequently, the choice of the internal model(s) is a crucial design decision that affects both energy efficiency and step success rate. Choosing a suitable LLM for one's web agent that integrates well for its given task is an inherent trait of each web agent. This highlights a critical design choice in web agent development: the use of resource-intensive, brute-force approaches versus more efficient, fine-tuned implementations. Choosing the latter can result in significant environmental and computational advantages, as highlighted in our work.

However, incentives for developers are missing, as benchmarks do not penalize energy consumption—yet. To ensure sustainable use of web agents at scale, it is essential to evaluate them not only on performance, but also on energy efficiency. We advocate for augmenting existing benchmarks with standardized energy consumption metrics as we have done for Mind2Web[4]. We propose to use energy per benchmark as a core evaluation metric, enabling transparent comparison across systems.

Benchmarking Open-Source Web Agents Our findings demonstrate that energy benchmarking does not require substantial overhead and can be conducted efficiently alongside standard performance evaluation. This enables a holistic assessment of web agents that captures both their task performance as well as energy efficiency. By incorporating energy consumption as a measurable and comparable metric, developers can identify implementations that achieve optimal task performance while minimizing energy consumption. This proposal supports informed design decisions and promotes the development of web agents that are not only powerful but also sustainable. As such, energy benchmarking should become a standard practice in the evaluation of open-source web agents, complementing accuracy-focused metrics such as step success rate with system-level efficiency.

[4]Source code available in our GitHub repo.

Estimating Energy Use of Proprietary Web Agents For web agents powered by proprietary LLMs, direct energy benchmarking is infeasible due to the lack of access to low-level system information, such as GPU usage. As a result, energy consumption must be estimated, which introduces significant uncertainty. Our comparison using MindAct, for which both estimation and direct benchmarking are possible, revealed discrepancies of up to a factor of 7. This gap mainly stems from conservative, upper-bound token assumptions we made in MindAct's candidate-generation stage. In practice, early termination, token truncation, and token reuse reduce the effective load, explaining the wide but unpredictable range between worst- and best-case energy usage and highlights how unreliable such estimations can be, even for a fully open-source agent.

Estimating energy use becomes even more problematic for proprietary LLMs, where energy per token cannot be measured and model parameters are undisclosed. If there is no other option, we recommend that developers report at the very least two key metrics: the energy consumption per token and, crucially, the number of tokens consumed. As we have shown in Table 3, it is not enough for web agents to only report the energy consumption on a per token level, as recommended in previous literature (Samsi et al. 2023a). Due to the structural complexity of web agents, often involving multiple LLM calls, preprocessing, and action steps, reporting energy per token alone is insufficient. For comparability, we propose to report the number of consumed tokens for established benchmarks, such as Mind2Web.

Conclusion

In this work, we compare the energy consumption of multiple web agents using both an empirical approach through benchmarking and a theoretical approach through estimation. We show that web agent design and used language models significantly influence the energy consumption and propose the introduction of web agent sustainability benchmarking to penalize inefficient energy consumption of web agents.

Acknowledgments

This work is supported by the European Union's Horizon Europe research and innovation program (HORIZON-CL4-2021-HUMAN-01) through the "SustainML" project (grant agreement No 101070408) and supported by the Carl Zeiss Foundation through the project "Sustainable Embedded AI".

References

Anthony, L. F. W.; Kanding, B.; and Selvan, R. 2020. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems. ArXiv:2007.03051.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 610–623.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Butler, G. 2024. AWS Global Data Centers Achieved PUE of 1.15 in 2023. https://www.datacenterdynamics.com/en/news/aws-global-data-centers-achieved-pue-of-115-in-2023/.

Chezelles, T. L. S. D.; Gasse, M.; Drouin, A.; Caccia, M.; Boisvert, L.; Thakkar, M.; Marty, T.; Assouel, R.; Shayegan, S. O.; Jang, L. K.; Lù, X. H.; Yoran, O.; Kong, D.; Xu, F. F.; Reddy, S.; Cappart, Q.; Neubig, G.; Salakhutdinov, R.; Chapados, N.; and Lacoste, A. 2024. The BrowserGym Ecosystem for Web Agent Research. arXiv:2412.05467.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. arXiv:2210.11416.

da Silva, J. 2024. Google turns to nuclear to power AI data centres. https://www.bbc.com/news/articles/c748gn94k95o.

Deng, X.; Gu, Y.; Zheng, B.; Chen, S.; Stevens, S.; Wang, B.; Sun, H.; and Su, Y. 2024. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

et al., O. 2024. GPT-4 Technical Report. arXiv:2303.08774.

GoogleDeepMind. 2025. Project Mariner. https://deepmind.google/technologies/project-mariner/. Accessed: 2025-01-17.

Gur, I.; Furuta, H.; Huang, A.; Safdari, M.; Matsuo, Y.; Eck, D.; and Faust, A. 2023. A real-world webagent with planning, long context understanding, and program synthesis. *arXiv preprint arXiv:2307.12856*.

He, H.; Yao, W.; Ma, K.; Yu, W.; Dai, Y.; Zhang, H.; Lan, Z.; and Yu, D. 2024. WebVoyager: Building an End-to-End Web Agent with Large Multimodal Models. arXiv:2401.13919.

He, P.; Gao, J.; and Chen, W. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. arXiv:2001.08361.

Krupp, L.; Geißler, D.; WOŹNIAK, P. W.; Lukowicz, P.; and Karolus, J. 2025. Quantifying Web Agents-A Survey on Web Agent Performance and Efficiency.

Lai, H.; Liu, X.; Iong, I. L.; Yao, S.; Chen, Y.; Shen, P.; Yu, H.; Zhang, H.; Zhang, X.; Dong, Y.; et al. 2024. Autowebglm: A large language model-based web navigating agent. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5295–5306.

Lannelongue, L.; Grealey, J.; and Inouye, M. 2021. Green algorithms: quantifying the carbon footprint of computation. *Advanced science*, 8(12): 2100707.

Li, P.; Yang, J.; Islam, M. A.; and Ren, S. 2023. Making ai less" thirsty": Uncovering and addressing the secret water footprint of ai models. *arXiv preprint arXiv:2304.03271*.

Liu, E. Z.; Guu, K.; Pasupat, P.; Shi, T.; and Liang, P. 2018. Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration. In *International Conference on Learning Representations (ICLR)*.

Liu, J.; Ou, T.; Song, Y.; Qu, Y.; Lam, W.; Xiong, C.; Chen, W.; Neubig, G.; and Yue, X. 2024. Harnessing webpage uis for text-rich visual understanding. *arXiv preprint arXiv:2410.13824*.

Lù, X. H.; Kasner, Z.; and Reddy, S. 2024. Weblinx: Real-world website navigation with multi-turn dialogue. *arXiv preprint arXiv:2402.05930*.

Ma, K.; Zhang, H.; Wang, H.; Pan, X.; and Yu, D. 2023. Laser: Llm agent with state-space exploration for web navigation. *arXiv preprint arXiv:2309.08172*.

Mazzucato, M. 2024. The Ugly Truth behind ChatGPT: AI Is Guzzling Resources at Planet-Eating Rates. https://www.theguardian.com/commentisfree/article/2024/may/30/ugly-truth-ai-chatgpt-guzzling-resources-environment.

MistralAI. 2025. Our contribution to a global environmental standard for AI. https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai.

NVIDIA. 2023. NVIDIA DGX H100 Datasheet. https://resources.nvidia.com/en-us-dgx-systems/ai-enterprise-dgx.

NVIDIA. 2024. NVIDIA H100 GPU Datasheet. https://nvdam.widen.net/s/fdllbtmmbv/h100-datasheet-2430615.

OpenAI. 2025. Introducing Operator. https://openai.com/index/introducing-operator/.

Ou, T.; Xu, F. F.; Madaan, A.; Liu, J.; Lo, R.; Sridhar, A.; Sengupta, S.; Roth, D.; Neubig, G.; and Zhou, S. 2024. Synatra: Turning indirect knowledge into direct demonstrations for digital agents at scale. *Advances in Neural Information Processing Systems*, 37: 91618–91652.

Patel, P.; Choukse, E.; Zhang, C.; Goiri, I.; Warrier, B.; Mahalingam, N.; and Bianchini, R. 2024. Characterizing Power Management Opportunities for LLMs in the Cloud. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 3*, volume 3 of *ASPLOS '24*, 207–222. Association for Computing Machinery. ISBN 979-8-4007-0386-7.

Patterson, D.; Gonzalez, J.; Le, Q.; Liang, C.; Munguia, L.-M.; Rothchild, D.; So, D.; Texier, M.; and Dean, J. 2021. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Samsi, S.; Zhao, D.; McDonald, J.; Li, B.; Michaleas, A.; Jones, M.; Bergeron, W.; Kepner, J.; Tiwari, D.; and Gadepally, V. 2023a. From Words to Watts: Benchmarking the Energy Costs of Large Language Model Inference. *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–9.

Samsi, S.; Zhao, D.; McDonald, J.; Li, B.; Michaleas, A.; Jones, M.; Bergeron, W.; Kepner, J.; Tiwari, D.; and Gadepally, V. 2023b. From words to watts: Benchmarking the energy costs of large language model inference. In *2023 IEEE High Performance Extreme Computing Conference (HPEC)*, 1–9. IEEE.

Schreiner, M. 2023. GPT-4 architecture, datasets, costs and more leaked. https://the-decoder.com/gpt-4-architecture-datasets-costs-and-more-leaked/.

Shi, M.; Currier, K.; Liu, Z.; Janowicz, K.; Wiedemann, N.; Verstegen, J.; McKenzie, G.; Graser, A.; Zhu, R.; and Mai, G. 2023. Thinking geographically about AI sustainability. *AGILE: GIScience Series*, 4: 42.

TechPowerup. 2025. GPU Specs Database. https://www.techpowerup.com/gpu-specs/.

U.S. Energy Information Administration. 2022. How much carbon dioxide is produced per kilowatthour of U.S. electricity generation? https://www.eia.gov/tools/faqs/faq.php?id=74&t=11.

Verdecchia, R.; Sallou, J.; and Cruz, L. 2023. A systematic review of Green AI. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 13.

Yang, K.; Liu, Y.; Chaudhary, S.; Fakoor, R.; Chaudhari, P.; Karypis, G.; and Rangwala, H. 2024. AgentOccam: A Simple Yet Strong Baseline for LLM-Based Web Agents. arXiv:2410.13825.

Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35: 20744–20757.

Zhang, Y.; Ma, Z.; Ma, Y.; Han, Z.; Wu, Y.; and Tresp, V. 2024. WebPilot: A Versatile and Autonomous Multi-Agent System for Web Task Execution with Strategic Exploration. arXiv:2408.15978.

Zheng, B.; Gou, B.; Kil, J.; Sun, H.; and Su, Y. 2024. Gpt-4v (ision) is a generalist web agent, if grounded. *arXiv preprint arXiv:2401.01614*.

Zheng, L.; Wang, R.; Wang, X.; and An, B. 2023. Synapse: Trajectory-as-exemplar prompting with memory for computer control. *arXiv preprint arXiv:2306.07863*.