# Quantifying the Carbon and Water Footprint of Artificial Intelligence: A Comprehensive Research Report for Calculator Development

## I. Executive Summary

**Purpose:** This research report provides a comprehensive foundation for the development of a user-friendly tool designed to calculate the estimated carbon and water footprint of various Artificial Intelligence (AI) activities. It aims to gather and synthesize the necessary data, methodologies, and best practices to accurately estimate these environmental impacts and suggest corresponding offset credits.

**Key Findings Overview:** The proliferation of AI technologies has brought to the forefront significant concerns regarding their energy and water consumption, alongside the associated carbon emissions. Current assessments of AI's environmental impact reveal substantial variability and a notable opacity in available data, particularly from commercial AI providers. This necessitates the use of robust estimation methodologies and a clear articulation of underlying assumptions and uncertainties within any footprint calculation tool.

The environmental footprint of AI is influenced by several core components: the specific AI models used (both for training and inference), the nature and complexity of AI tasks performed, the hardware and data center infrastructure powering these activities, and the geographical location of computation, which dictates grid carbon intensity and water stress levels. Recent academic studies have begun to quantify these impacts for popular models like the GPT series, Llama, Claude, and image generation models such as Stable Diffusion and DALL-E, highlighting, for instance, that training GPT-3 consumed approximately 1,287 MWh of energy and 700,000 liters of water for cooling alone , while a single short query to GPT-4o might consume around 0.43 Wh. Image generation tasks are shown to be orders of magnitude more energy-intensive than analytical tasks.

The development of a comprehensive AI footprint calculator faces challenges due to this data opacity, the rapid evolution of AI models and hardware, and the lack of standardized reporting. This report addresses these challenges by reviewing existing calculators, analyzing AI provider sustainability reports, detailing the consumption profiles of AI components, outlining data requirements and robust calculation methodologies (including strategies for intelligent estimation where direct data is scarce), and discussing the mechanisms for carbon and water credits.

**Report Structure:** This report is structured to provide a foundational understanding of the current AI environmental footprint landscape, delve into the specifics of quantifying AI usage across its various components, outline data requirements and calculation methodologies, discuss carbon and water credit mechanisms, and finally, consider tool design, challenges, and future-proofing strategies.

# II. The Current AI Environmental Footprint Landscape

## A. Critical Review of Existing AI Footprint Calculators and Methodologies

The increasing ubiquity and computational demand of AI have spurred the development of various tools and methodologies aimed at quantifying its environmental impact. A critical review of these existing solutions reveals a diverse landscape, ranging from educational estimators to more specialized research and enterprise tools, each with its own set of user inputs, outputs, underlying assumptions, and limitations.

The **WAB Learns AI Environmental Footprint Estimator**, for example, is designed for educational illustration, allowing users to input the number of text queries, AI searches, image/video generations, and coding tasks, along with region/provider, to receive an estimate of energy (kWh), CO_2 emissions (kg), and water usage (L). It relies on publicly available estimates, such as 0.0029 kWh/query for ChatGPT and 0.20 kWh/image for image generation, and grid factors for China (0.582 kgCO_2/kWh) and the USA (0.367 kgCO_2/kWh). However, its creators explicitly state its limitations for "serious technical use" due to the high variability and rapid evolution of AI models and data. This highlights a core challenge for any calculator: maintaining accuracy in a fast-moving field.

The **Optim.ai GreenIMPACT Calculator©** positions itself as a more comprehensive, AI-powered tool for quantifying the environmental net impact of AI systems across their lifecycle, including energy, water, emissions, e-waste, and even positive contributions. It adheres to international standards like ISO and the GHG Protocol. Users are required to undertake a significant data collection effort (2-4 weeks) before inputting data into the application (2-4 hours), which then generates an impact report and recommendations. While its lifecycle approach is thorough, the extensive upfront data requirement may be a barrier for users seeking quick estimations, and the "AI-powered" calculation mechanism lacks full transparency regarding its specific algorithms.

**Deloitte's AI Carbon Footprint Calculator** takes a different approach, using multiple-choice questions about the project's location, use case, infrastructure, and AI model to generate an overall carbon footprint score (Low, Medium, or High) and a personalized report with recommendations. It is based on published research and metrics, with outputs weighted across the input categories. A video demonstration suggested a query for script generation consumed 2 Wh and two bottles of water. This tool appears more qualitative and score-based, potentially lacking the precise quantitative energy and water figures desired for detailed footprinting. The specifics of its underlying research and weighting are not fully transparent in the available materials.

Focusing specifically on inference energy efficiency, the **Hugging Face/Salesforce/Cohere/Carnegie Mellon University AI Energy Score** initiative provides a standardized framework for evaluating AI models. It benchmarks models across common tasks on standardized NVIDIA H100 GPUs, publishing results on a public leaderboard with a 1-5 star rating system. GPU energy consumption in Wh per 1000 inferences is a key output metric. While valuable for comparing inference energy, its scope does not extend to the full lifecycle, water footprint, or training impacts. The relative nature of its scoring means a model's rating can change based on the pool of evaluated models, not just its own efficiency. A notable point arose from a discussion where the AI Energy Score for stabilityai/stable-diffusion-xl-base-1.0 was 1.64 kWh/1000 inferences (on H100), contrasting with 11.41 kWh/1000 inferences (on A100) from

another paper, a difference attributed to hardware efficiency and image dimension constraints. This underscores the critical importance of methodological consistency and transparency in hardware and task parameters.

Cloud providers like **Google Cloud, Microsoft Azure, and Amazon Web Services (AWS)** offer tools for customers to track the carbon emissions associated with their cloud usage.

- **Google Cloud Carbon Footprint** provides location- and market-based Scope 1, 2, and 3 emissions, with a published methodology reviewed against the GHG Protocol, allocating emissions via machine-level monitoring. It covers AI services like Vertex AI.
- **Microsoft Azure** offers tools like the Emissions Impact Dashboard and Azure Carbon Optimization, tracking emissions at resource and subscription levels.
- **AWS Customer Carbon Footprint Tool (CCFT)** tracks Scope 1 and 2 emissions by region, with an independently validated methodology that allocates unused capacity and overhead. While these tools are improving in granularity , they often face criticism for lacking transparency in how AI-specific impacts are disaggregated within shared infrastructure and for not fully covering the hardware lifecycle. The ease of use of AI services often means users are unaware of the environmental cost of their actions.

Other tools include the **ML CO2 Impact Calculator**, an online tool for estimating CO2 from ML model *training* based on hardware, runtime, and cloud provider inputs. However, it relies on assumptions and is considered a rough guide. **CodeCarbon** is a Python library for more granular measurement of training and inference emissions by tracking CPU/GPU energy but requires technical expertise. **Arbor.eco** outlines a 5-step process for estimating AI carbon emissions, differentiating internal (Scope 1/2) and external (Scope 3) AI model usage, and uses figures like 4.32g CO_2e per ChatGPT query from MLCO2. The **HCI GenAI CO2ST Calculator** is tailored for Human-Computer Interaction research, translating technical factors into HCI-relevant inputs and erring on the conservative side for its estimates, explicitly noting the opacity from large model providers as a challenge.

This review of existing calculators reveals a fundamental challenge: balancing usability, accuracy, and transparency. Tools that aim for high accuracy and comprehensive lifecycle assessment, like Optim.ai or detailed academic approaches, often demand extensive data input and complex methodologies, which can be prohibitive for general users. Conversely, more user-friendly tools, such as the WAB Learns estimator, achieve simplicity by relying on broader estimations and averages, thereby sacrificing granular accuracy and often being suitable only for educational purposes. Transparency is another critical dimension; "AI-powered" calculators may obscure their underlying methodologies , whereas tools based on open research offer clarity but are constrained by the public availability and granularity of that research. The development of a new, effective calculator must navigate this "Transparency-Usability-Accuracy Trilemma" carefully.

Furthermore, many existing tools are specialized, focusing on particular phases of the AI lifecycle (e.g., training for ML CO2 Impact , inference energy for the AI Energy Score ) or specific user communities (e.g., HCI researchers for the HCI GenAI CO2ST Calculator ). A general-purpose calculator, as envisioned by the user query, needs to address a wider spectrum of AI activities—including LLMs, image generation, and complex agentic platforms—and cater to a broader audience with varying levels of technical understanding. This "scope mismatch" suggests that the proposed tool will need to synthesize functionalities or offer a more holistic approach than many current options.

Finally, the reliability of all calculators is intrinsically tied to the quality and currency of their underlying data, covering model energy consumption, hardware efficiency, and emission factors. The AI field is characterized by rapid advancements in models and hardware, alongside

fluctuating grid carbon intensities. This "recency challenge," explicitly acknowledged by the WAB Learns tool , means that any calculator's database can quickly become outdated. Consequently, a critical design consideration for the new tool is the implementation of a robust and agile mechanism for continuous data updates and recalibration.

## B. Analysis of AI Provider Sustainability Reports and Public Commitments

Major AI providers, including cloud service companies and prominent model developers, are increasingly acknowledging their environmental responsibilities through sustainability reports and public commitments. An analysis of these disclosures offers insights into their transparency levels, the types of metrics reported, and the methodologies used to account for AI-specific environmental impacts.

**Google** publishes annual sustainability reports, with its 2024 report notably utilizing AI in its production. The company provides the Google Cloud Carbon Footprint tool, allowing customers to track Scope 1, 2, and 3 emissions from their cloud usage, with data exportable to BigQuery and a methodology verified against the GHG Protocol. Google reports metrics like Power Usage Effectiveness (PUE), with a fleet-wide trailing twelve-month (TTM) PUE of 1.09 in Q1 2025 , and is committed to operating on 24/7 carbon-free energy by 2030. AI is estimated to contribute 10-15% of Google's total electricity consumption , and the company attributed a rise in its overall GHG emissions in 2023 to AI. While the Carbon Footprint tool covers AI services like Vertex AI , the specific disaggregation of AI's footprint from shared infrastructure is complex, relying on bottom-up machine-level monitoring. Some critics note a lack of detailed breakdown for individual AI products' environmental impact. Google also highlights hardware efficiency, such as its Trillium TPUs being 67% more energy-efficient than the TPU v5e.

**Amazon Web Services (AWS)** also issues sustainability reports and offers the Customer Carbon Footprint Tool (CCFT) for users to track Scope 1 and 2 emissions from their AWS usage, with breakdowns by AWS Region. The CCFT's methodology (v2.0) is independently validated and includes allocation for unutilized capacity and overhead. AWS reported a global PUE of 1.15 in 2023 and aims for 100% renewable energy usage by 2025 and to be water positive by 2030. The CCFT implicitly covers AI services like SageMaker and Bedrock as it accounts for all AWS services ; however, specific details on footprint calculation per AI API call are not extensively publicized. AWS itself uses AI, such as its FlowMS tool, to improve its own utility efficiency.

**Microsoft Azure** provides a suite of tools including the Microsoft Sustainability Manager, Emissions Impact Dashboard (a Power BI application), and the Azure Carbon Optimization tool, which tracks emissions at the resource and subscription level for all Azure services. Microsoft has committed to being carbon negative by 2030, using 100% renewable energy by 2025, being water positive by 2030, and achieving zero waste by 2030. The tools cover AI services like Azure Cognitive Services and Azure OpenAI. Microsoft, like Google, also attributed an increase in its GHG emissions partly to AI development.

**Meta** publishes an annual Sustainability Report and has maintained net-zero emissions in its global operations since 2020 by matching 100% of its electricity use with renewable energy. However, its absolute emissions rose in 2023, with AI's increased energy demand cited as a contributing factor. Meta aims to cut Scope 1 and 2 emissions by 42% by 2031 (from a 2021 baseline), but its carbon footprint has reportedly increased by approximately 38%. Meta uses AI for its own ESG data gathering and to identify emission reduction opportunities, such as

optimizing concrete mixtures for data centers and improving cooling fan efficiency. Detailed disaggregation of AI's operational footprint for public reporting is not extensively provided.
**NVIDIA**, a key hardware provider, focuses on the energy *efficiency* of its products in its Corporate Sustainability Report. For example, it states its Blackwell GPUs are 20 times more energy efficient than traditional CPUs for AI and HPC workloads, and the GB200 Grace Blackwell Superchip offers 25 times better energy efficiency than the prior Hopper generation for massive LLMs. NVIDIA argues that accelerated computing saves energy overall compared to CPU-based approaches for AI and that while training AI consumes energy, applying AI can save energy in other sectors. The company reports overall corporate GHG emissions but does not typically provide granular lifecycle assessment data (manufacturing footprint per chip) for its AI hardware beyond these efficiency claims.
Examining these provider reports reveals a pattern: while there's a strong emphasis on improving the energy *efficiency* of new hardware (e.g., NVIDIA's Blackwell ) and data center operations (e.g., Google's PUE of 1.09 , AWS's PUE of 1.15 ), the *absolute* energy consumption and greenhouse gas emissions attributed to AI are paradoxically increasing for major cloud providers like Google and Microsoft. This suggests that efficiency improvements are currently being outpaced by the sheer growth in demand for AI computation, a phenomenon akin to the Jevons Paradox, where increased efficiency leads to increased overall consumption. This is a critical consideration for any footprint calculator: efficiency metrics for components are important, but the volume of use is a dominant factor in the total environmental burden.
Another significant observation is the disparity in transparency regarding the embodied environmental costs of AI hardware. Cloud providers are becoming more transparent about the Scope 1 and Scope 2 operational emissions of their services. However, the embodied carbon and water in the manufacturing of the specialized hardware (GPUs, TPUs) that underpins AI—a substantial component of Scope 3 emissions for AI users and developers—remains largely opaque. NVIDIA, the leading GPU supplier, primarily highlights the operational efficiency and energy savings of its chips during use, rather than providing detailed lifecycle assessment (LCA) data for the manufacturing footprint per unit. This "Scope 3 Chasm" for AI hardware creates a significant data gap that a comprehensive footprint calculator must address, likely by relying on academic LCAs and estimations.
Furthermore, while carbon footprint tracking tools from cloud providers are maturing, detailed and granular reporting for AI's *water footprint* is less developed. Providers often state broad corporate goals like becoming "water positive by 2030" but do not typically offer customer-facing tools that attribute water consumption to specific cloud services or AI workloads with the same granularity as carbon. Academic research is increasingly highlighting the substantial water demands of AI for data center cooling and electricity generation , but this is not yet consistently reflected in provider-supplied data for specific AI services. This indicates an area where the proposed calculator will need to lean heavily on scientific literature and robust estimation techniques.
The following table summarizes key sustainability disclosures from major AI providers:
**Table 1: AI Provider Sustainability Disclosure Summary**

| Provider | Key Sustainability Reports/Tools | Reported Corporate-Level Metrics (PUE, WUE, Renewable %, Carbon Intensity) | Specific AI-Related Disclosures/Metrics | Stated AI Sustainability Goals/Strategies | Methodologies for AI Footprinting (if disclosed) |
|---|---|---|---|---|---|
| **Google** | Annual Sustainability Report ; Google Cloud Carbon Footprint | PUE: 1.09 (TTM Q1 2025) ; Water: 5.6B gallons (2021) , Replenish 120% by 2030; Renewable %: 24/7 CFE by 2030 | AI is 10-15% of total electricity use ; GHG increase blamed on AI ; Trillium TPU 67% more efficient | Net-zero by 2030 ; Reduce 1 GtCO2e annually via AI solutions | Bottom-up machine-level monitoring for cloud footprint ; Covers Vertex AI |
| **AWS** | Annual Sustainability Report ; Customer Carbon Footprint Tool (CCFT) | PUE: 1.15 (Global 2023) ; Water: Water positive by 2030 ; Renewable %: 100% by 2025 | CCFT covers all services including AI ; Graviton4/Inferentia2 chip efficiency | Path to 100% renewables; Water positive; Lower-carbon materials in data centers | CCFT v2.0 methodology allocates based on server rack utilization, independently validated |
| **Microsoft Azure** | Sustainability Manager ; Emissions Impact Dashboard ; Azure Carbon Optimization | PUE: Not specified in snippets; Water: Water positive by 2030 ; Renewable %: 100% by 2025 ; Carbon Negative by 2030 | GHG increase blamed on AI ; Tools cover Azure OpenAI ; Liquid immersion cooling research | Carbon negative, water positive, zero waste by 2030 | Emissions Impact Dashboard & Azure Carbon Optimization track at resource/subscription level |
| **Meta** | Annual Sustainability Report | Renewable %: 100% operations match since 2020 ; Absolute emissions up ~38% (vs 2021) | AI energy demand acknowledged as challenge ; Uses AI for own sustainability (concrete, cooling) | Net zero operations; Cut Scope 1&2 by 42% by 2031 | AI-specific operational footprint disaggregation not detailed. |
| **NVIDIA** | Corporate Sustainability Report | Focus on product energy efficiency gains (e.g., Blackwell | GB200 25x more efficient than Hopper for LLMs ; AI for | Drive industrial productivity and energy savings via AI | Focus on operational energy savings of their |

| Provider | Key Sustainability Reports/Tools | Reported Corporate-Level Metrics (PUE, WUE, Renewable %, Carbon Intensity) | Specific AI-Related Disclosures/Metrics | Stated AI Sustainability Goals/Strategies | Methodologies for AI Footprinting (if disclosed) |
|---|---|---|---|---|---|
| | | 20x > CPU) | sustainability applications (drug discovery, grid) | | hardware in use, not manufacturing LCA per chip. |

# III. Quantifying AI Usage: Environmental Impact Factors & Data Availability

## A. AI Models & Platforms: Energy, Water, and Carbon Footprints

The environmental toll of AI systems is significantly influenced by the specific models and platforms employed. This section examines the energy, water, and carbon consumption associated with various AI models during their training and inference phases, drawing primarily from academic research due to the prevalent opacity from commercial developers.

### 1. Large Language Models (LLMs)

LLMs, due to their immense size and computational requirements, are a major focus of environmental impact assessments.

- **GPT Series (OpenAI):**
    - **Training:** The training of GPT-3 (175 billion parameters) is a widely cited benchmark, estimated to have consumed around 1,287 MWh of electricity and generated approximately 552 tonnes of CO_2 equivalent (tCO_2e). The water footprint for cooling during GPT-3 training alone was estimated at 700,000 liters (0.7 million L) , with another study suggesting a total water consumption (including electricity generation) of 5.4 million liters. More recent models like GPT-4 are estimated to have an even larger training footprint, with one source citing 5,184 tCO_2e for GPT-4 training.
    - **Inference:** Estimates for GPT-3 inference vary, from 0.0003 kWh/query to 0.004 kWh per medium request (~300 words output). The WAB Learns estimator uses 0.0029 kWh/query for ChatGPT. One study suggests a ChatGPT query consumes about five times more electricity than a standard Google search. Daily energy use for ChatGPT (serving 13 million users with 65 million queries) was estimated at 77,160 kWh, producing about 30 tCO_2e per day, or roughly 0.5g CO_2 per query. Another, more alarming estimate, suggests ChatGPT uses 39.98 million kWh per day and 39.16 million gallons (148.28 million L) of water daily for cooling and electricity generation.
    - The recent academic paper by Jegham et al. (2025) provides crucial inference benchmarks for several GPT variants. For GPT-4.1 nano, a short prompt (100 input/300 output tokens) consumed 0.103 Wh, less than 2ml of water, and less than

0.3g CO_2e; a medium prompt (1k/1k tokens) took 0.271 Wh, and a long prompt (10k/1.5k tokens) took 0.454 Wh. For GPT-4o (Mar '25 version), a short prompt was 0.421 Wh (another source states 0.43 Wh ), a medium prompt 1.214 Wh, and a long prompt 1.788 Wh. The same study estimated that OpenAI's reasoning model, o3, was the most energy-intensive, consuming 39.223 Wh for a long prompt, while GPT-4.5 consumed 30.495 Wh for a similar task. The projected annual water consumption for GPT-4o in 2025, based on estimated usage, is between 1,335 and 1,580 kiloliters, with carbon emissions between 138 and 163 kilotonnes CO_2e.
    - **OpenAI Sora (Text-to-Video):** Specific energy or water footprint data for Sora was not found in the provided materials. However, general concerns about the high computational and environmental costs of text-to-video models, which are typically transformer-based, are noted. One source mentions OpenAI's general server water use for ChatGPT at approximately 2 liters per 50 queries , while another projects that future AI (not Sora-specific) could demand 720 billion gallons of water annually in the U.S. for server cooling.
- **Claude Series (Anthropic):**
    - **Inference (Jegham et al., 2025):** Claude 3.7 Sonnet was ranked highest in eco-efficiency in this study. For inference, it consumed 0.836 Wh (short prompt), 2.781 Wh (medium), and 5.518 Wh (long). The more specialized Claude 3.7 Sonnet ET (Extended Thinking) consumed 17.045 Wh for a long prompt.
- **Llama Series (Meta):**
    - **Training:** Morrison et al. (2025) report that training Llama 2 7B required 81 MWh and resulted in 31 tCO_2e, while Llama 2 13B required 162 MWh and 62 tCO_2e. The training of Llama 3.1 405B is estimated to have produced 8,930 tCO_2e.
    - **Inference (Jegham et al., 2025):** LLaMA-3.2 1B consumed 0.070 Wh (short), 0.218 Wh (medium), and 0.342 Wh (long), with water use under 2ml and carbon under 0.3g CO_2e per query. The larger LLaMA-3.1 405B consumed 20.757 Wh for a long prompt.
    - **Inference (Rubei et al.):** This research focused on Llama 3 energy for code completion, using CodeCarbon for GPU energy measurement. A default zero-shot prompt consumed approximately 0.000016 kWh, with custom tags offering reductions.
    - Another estimate for Llama-3-70B inference is around 0.010 kWh/request, considering enterprise Service Level Objectives (SLOs).
- **Gemini Series (Google):**
    - Gemini is available in Ultra, Pro, and Nano sizes. Gemini 1.5 Pro is noted to require significantly less compute to train than the earlier Gemini 1.0 Ultra. Google has emphasized efficiency improvements, such as the Trillium TPU being 67% more energy-efficient than its predecessor, TPU v5e. However, specific public data on energy, water, or carbon footprints for Gemini model training or inference (e.g., Wh/query, Liters/query) were not found in the provided research snippets beyond these general statements about efficiency or cost per million tokens (which is a financial metric, not an energy one ). The comprehensive study by Jegham et al. (2025), which benchmarked 30 LLMs, did not explicitly include Gemini models in its detailed energy/water/carbon tables.
- **Mistral Models (Mistral AI):**
    - Mistral-7B inference energy was benchmarked by Maliakel et al. (2025), showing around 2J for BoolQ tasks and 40J for SQuADv2 tasks. The model utilizes

Grouped-Query Attention for efficiency.
    - Specific public data on training energy, water, or carbon footprints for Mistral models (including Mixtral or Mistral Large) were not found. Mistral AI's transparency report for Mistral 7B indicated "Not disclosed" for compute and data size. The Jegham et al. (2025) study did not include Mistral models in its detailed table of 30 LLMs.
- **Cohere Models:**
    - Cohere is a partner in the AI Energy Score initiative, which focuses on inference energy.
    - A policy primer from Cohere discusses the energy consumption of models like GPT-3 (1287 MWh for training, 502 tCO_2) but does not provide specific figures for Cohere's own models (e.g., Command R+). The primer emphasizes various efficiency techniques Cohere employs, such as model design, data pruning, distillation, Parameter Efficient Finetuning (PEFT), and quantization.
    - One study mentions Command R+ in a survey context but provides no environmental data.
- **DeepSeek Models (DeepSeek AI):**
    - **Inference (Jegham et al., 2025):** DeepSeek-R1 was one of the most energy-intensive models benchmarked, consuming 23.815 Wh (short), 29.000 Wh (medium), and 33.634 Wh (long prompt). Its water use was over 150ml/query, and carbon emissions exceeded 14g CO_2e/query. DeepSeek-V3 consumed 3.514 Wh (short), 9.129 Wh (medium), and 13.838 Wh (long). The high footprint of DeepSeek models was partly attributed to potential inefficiencies in their data centers.
- **Other Foundational Models (BERT, T5):**
    - **BERT Training (Strubell et al., 2019):** Training BERTbase (110M parameters) on 64 Tesla V100 GPUs consumed 1507 kWh (PUE-adjusted) and emitted 1438 lbs CO_2e (approx. 652 kgCO_2e).
    - **T5 Training (Patterson et al., 2021):** An 11-billion parameter T5 model consumed 85.7 MWh for training, resulting in 46.7 tCO_2e.
    - **BERT/T5 Inference Energy (Maliakel et al., 2025):** For T5-3B on the SQuADv2 task, inference consumed 40J/query. The study noted that larger, more complex models like T5-3B can exhibit inefficiencies, particularly in simpler tasks.

The following table consolidates key inference footprint data from Jegham et al. (2025) , which is a pivotal source for recent model comparisons.

**Table 2: LLM Inference - Energy, Water, and Carbon Footprints (per query)** *(Data primarily from Jegham et al., 2025. Water and Carbon are ranges/examples based on text)*

| Model Name | Provider/ Developer | Assumed Hardware (Inferred) | Prompt Type (Tokens In/Out) | Energy (Wh/query) | Water (ml/query) | Carbon (gCO2e/query) |
|---|---|---|---|---|---|---|
| GPT-4.1 nano | OpenAI | A100 (Small class) | Short (100/300) | 0.103 ± 0.037 | < 2 | < 0.3 |
| | | | Medium (1k/1k) | 0.271 ± 0.087 | < 2 | < 0.3 |
| | | | Long (10k/1.5k) | 0.454 ± 0.208 | < 2 | < 0.3 |
| GPT-4o (Mar '25) | OpenAI | H100/H200 (Large class) | Short (100/300) | 0.421 ± 0.127 | Low (Est.) | Low (Est.) |
| | | | Medium | 1.214 ± | | |

| Model Name | Provider/ Developer | Assumed Hardware (Inferred) | Prompt Type (Tokens In/Out) | Energy (Wh/query) | Water (ml/query) | Carbon (gCO2e/query) |
|---|---|---|---|---|---|---|
| | | | (1k/1k) | 0.391 | | |
| | | | Long (10k/1.5k) | 1.788 ± 0.363 | | |
| Claude 3.7 Sonnet | Anthropic | H100/H200 (Large class) | Short (100/300) | 0.836 ± 0.102 | Low-Med (Est.) | Low-Med (Est.) |
| | | | Medium (1k/1k) | 2.781 ± 0.277 | | |
| | | | Long (10k/1.5k) | 5.518 ± 0.751 | | |
| LLaMA-3.2 1B | Meta | H100 (Nano class) | Short (100/300) | 0.070 ± 0.011 | < 2 | < 0.3 |
| | | | Medium (1k/1k) | 0.218 ± 0.035 | < 2 | < 0.3 |
| | | | Long (10k/1.5k) | 0.342 ± 0.056 | < 2 | < 0.3 |
| DeepSeek-R1 | DeepSeek AI | H800 (Large class) | Short (100/300) | 23.815 ± 2.160 | > 150 | > 14 |
| | | | Medium (1k/1k) | 29.000 ± 3.069 | > 150 | > 14 |
| | | | Long (10k/1.5k) | 33.634 ± 3.798 | > 150 | > 14 |
| o3 | OpenAI | H100/H200 (Large class) | Short (100/300) | 7.026 ± 3.663 | High (Est. >150) | High (Est. >14) |
| | | | Medium (1k/1k) | 21.414 ± 14.273 | | |
| | | | Long (10k/1.5k) | 39.223 ± 20.317 | | |
| | | | | | | |

## 2. Image/Video/Audio Generation Models

- **Stable Diffusion (Stability AI):**
  - **SDXL Inference:** The AI Energy Score, using an NVIDIA H100 GPU and standardized image dimensions, reported 1.64 kWh per 1000 images. In contrast, a study by Luccioni et al., using an NVIDIA A100 GPU and default (likely larger) image settings, reported 11.41 kWh per 1000 images for the same model. This nearly seven-fold difference underscores the significant impact of hardware choice and output parameters.
  - General text-to-image models average around 2.9 kWh per 1000 inferences. One highly carbon-intensive image model produced 1.594 kg CO_2 (from 11.49 kWh) per 1000 images.
  - A high-level aggregate estimate suggested that 10 million Stable Diffusion users, each using the system for 1.5 hours daily on an RTX 3090, could lead to an annual energy consumption of approximately 1.92 TWh. This figure, however, relies on

broad assumptions about usage patterns and hardware.

- **DALL-E (OpenAI):**
  - **DALL-E 2 Inference (powered by GPT-3):** Estimated to produce around 2.2 kg CO_2e per 1000 queries.
  - General statements acknowledge DALL-E's substantial energy demands.
- **Midjourney:**
  - Generating a single image is estimated to take 5-50 petaoperations, potentially up to 40 seconds on an NVIDIA A100 GPU. This translates to approximately 4.5 kWh per 1000 images and around 1.9 kg CO_2e per 1000 images.
  - While many sources discuss Midjourney , specific, comparable energy and water data per image are scarce beyond the estimate in.
- **Sora (OpenAI Text-to-Video):**
  - No specific energy or water consumption figures for Sora were found in the provided research snippets. However, text-to-video models, typically based on computationally intensive transformer architectures, are generally acknowledged to have high environmental costs.
- **Text-to-Audio Models:**
  - A study analyzed seven state-of-the-art text-to-audio diffusion models, noting their high computational demands and the variability in energy consumption based on generation parameters. Specific figures per model were not available in the snippet.

The following table summarizes available inference data for image generation models:

**Table 3: Image Model Inference - Energy and Carbon Footprints**

| Model Name | Metric Unit | Energy (kWh) | Carbon (kgCO2e) | Hardware (if specified) | Source(s) |
|---|---|---|---|---|---|
| Stable Diffusion XL | per 1000 images | 1.64 | N/A | NVIDIA H100 | |
| Stable Diffusion XL | per 1000 images | 11.41 | N/A | NVIDIA A100 | |
| Generic Text-to-Image | per 1000 inferences | 2.9 | 1.594 (from 11.49 kWh) | N/A | |
| DALL-E 2 | per 1000 queries | N/A | 2.2 | GPT-3 powered | |
| Midjourney | per 1000 images | 4.5 | 1.9 | NVIDIA A100 (implied) | |
| *(N/A: Not Available in provided snippets for that specific metric/hardware combination)* | | | | | |

## 3. Other AI Model Types (Classification, Translation, etc.)

Simpler, task-specific AI models generally have a much lower environmental footprint compared to large generative models.

- **General Tasks (from Greenly.earth, based on Luccioni et al. ):**
  - **Text classification:** 0.002 kWh per 1000 queries (approximately 0.3g CO_2e).

    - **Image classification:** 0.007 kWh per 1000 queries (approximately 1.1g CO_2e).
    - **Object detection:** 0.038 kWh per 1000 queries (approximately 6.1g CO_2e). These figures highlight that discriminative or analytical tasks are orders of magnitude less energy-intensive per query than generative tasks like image generation (2.9 kWh per 1000 queries ).

### 4. Agentic "Operator" Platforms

Agentic AI platforms (e.g., ChatGPT Operator, Manus, ServiceNow AI Agent Orchestrator ) represent a new layer of complexity for footprint estimation. These platforms typically use an LLM as an orchestrator to plan tasks, select appropriate tools (which can be other AI models or APIs), execute them, and synthesize results.

- **Methodologies to estimate orchestrator overhead:** The orchestrator itself consumes resources for its internal processing, which involves LLM calls for planning, reasoning about tool selection, and aggregating outputs. This "orchestrator overhead" is distinct from the footprint of the tools it invokes. The **CarbonCall framework** is designed to optimize and reduce the carbon footprint of function-calling LLMs on edge devices. It integrates dynamic tool selection (using embeddings, FAISS, and Cross-Encoders to pick relevant tools), carbon-aware execution (adjusting device power based on real-time carbon intensity forecasts), and quantized LLM adaptation (switching to lower-precision model versions to maintain throughput under power constraints). While CarbonCall aims to minimize the overall operational footprint of an agentic task on an edge device, it does not explicitly detail a methodology for isolating and quantifying the orchestrator's specific computational overhead versus the footprint of the executed tools, particularly in a cloud-based multi-tool workflow scenario. A general approach to estimate orchestrator overhead would involve:
    1. Identifying the LLM used by the orchestrator.
    2. Logging all LLM calls made by the orchestrator for its internal decision-making processes (planning, tool selection, result synthesis).
    3. Estimating the footprint of these internal LLM calls using per-query/per-token data for that specific orchestrator LLM. The primary challenge lies in the dynamic and variable nature of these internal calls, which depend on the complexity of the user's request and the agent's reasoning path.
- **How to aggregate the footprint of underlying tools/models:** The total footprint of an agentic session is the sum of the orchestrator's overhead and the individual footprints of all underlying tools and models invoked. To achieve this aggregation, the following data points would be necessary:
    - **From the user:** A clear description of the complex task or goal.
    - **From the agentic platform (requiring detailed tracing):**
        - The specific LLM acting as the orchestrator.
        - The number, type (model used), input/output token counts, and context for each LLM call made by the orchestrator for its internal processing.
        - A list of all external tools or sub-models invoked during the session.
        - For each tool/model invocation:
            - The specific tool/model used (e.g., GPT-4 for text summary, Stable Diffusion for image generation).
            - Relevant usage metrics (e.g., tokens processed for an LLM tool, number/complexity of images for an image tool, duration/parameters

for a data analysis tool).
    - The geographical location/provider where each tool was executed (if different from the orchestrator and known). The study by Deng et al. (2024) and Ma et al. (2023) on web agents (MindAct and LASER) estimated energy per action and then multiplied by the average number of actions per task. For instance, LASER, using GPT-4 via API, was estimated to emit approximately 9.6 kg CO_2 per task on the Mind2Web benchmark. This action-based summation approach could be adapted for more general agentic platforms, where each "action" corresponds to an orchestrator LLM call or an external tool usage.

The inherent complexity of agentic systems, involving multiple, often unpredictable sequences of LLM calls and tool interactions, presents a significant challenge. This "multiplier of unknowns" means that estimating their footprint will heavily rely on detailed logging from the platform and robust models for each component's impact. Without such detailed tracing, estimations will carry high uncertainty.

## Influence of Model Size & Architecture on Consumption

- **Model Size (Parameters):** Generally, models with more parameters require more computational resources and thus consume more energy for both training and inference. For instance, Maliakel et al. (2025) demonstrated that larger models like GPT-J-6B (6 billion parameters) are considerably more energy-intensive for inference tasks than smaller models like GPT-2 (1.5 billion parameters). Jegham et al. (2025) stratified models into hardware classes based on parameter count (Nano <7B, Micro 7-20B, etc.) for their analysis. The training carbon emissions for Llama 3.1 405B (8,930 tCO_2e) are substantially higher than those reported for GPT-3 175B (552 tCO_2e). However, the relationship is not always linear or simple. The specific hardware used for deployment can significantly alter this dynamic. For example, the smaller GPT-4o mini, when inferred on less efficient A100 GPUs, consumed slightly more energy per query than the larger GPT-4o inferred on more modern H100/H200 GPUs. This illustrates that hardware efficiency can sometimes outweigh the benefits of a smaller model size in determining real-world energy use.
- **Model Architecture:** The architecture of an AI model plays a crucial role in its environmental footprint. Transformer models, which underpin most modern LLMs and generative AI, are known for their computational intensity. Encoder-decoder architectures, like that of T5-3B, can exhibit higher computational overhead and thus higher energy consumption for simpler tasks compared to decoder-only models. Models designed for complex reasoning, often employing chain-of-thought prompting (e.g., OpenAI's o3, DeepSeek-R1), tend to be very energy-intensive during inference. However, architectural optimizations can mitigate this; for example, Claude 3.7 Sonnet ET, also a reasoning model, was found to be more energy-efficient than o3. Innovations like Mixture of Experts (MoE) architectures aim to improve efficiency by activating only a subset of the model's weights for any given input. However, their total parameter counts can be extremely large, and standardized methods for accounting for their footprint (e.g., whether to count all parameters or only active ones for training FLOPs) are still debated. Techniques like grouped-query attention, used in Mistral 7B, can enhance inference speed and reduce memory requirements, thereby improving energy efficiency.

The interplay between model design choices made during training and the subsequent

inference demands is critical. While inference often constitutes the larger portion of a model's lifecycle energy consumption , the architecture and size determined during training establish a fundamental baseline for the potential inference footprint per operation. Optimizations at the inference stage, such as quantization or efficient serving infrastructure, can reduce this, but they cannot fundamentally alter the inherent computational load dictated by a model's design if it is oversized or ill-suited for the task at hand.
A significant challenge in accurately assessing these influences is the opacity surrounding proprietary "frontier" models (like GPT-4/4o, Claude 3, Gemini, Sora, Midjourney). Specific, verifiable data on their training energy/water consumption and granular inference characteristics are often scarce. Researchers frequently resort to estimations based on API performance, inferring underlying hardware, or relying on high-level statements from providers. In contrast, open models (e.g., Llama, OLMo, some versions of Stable Diffusion) tend to have more research available detailing their training processes and inference characteristics. This data disparity is a major hurdle for creating a universally accurate footprint calculator.

### Typical Carbon Footprint (gCO_2eq per unit)

The carbon footprint per unit of AI use is highly variable, depending on the model, task complexity, hardware efficiency, and, crucially, the carbon intensity of the electricity grid powering the computation.

- **LLM Inference:** Reported figures range from less than 0.3 gCO_2e per query for highly efficient models like GPT-4.1 nano to over 14 gCO_2e per query for energy-intensive models like DeepSeek-R1. A general estimate for a ChatGPT query is around 0.5g CO_2e.
- **Image Generation:** Estimates include approximately 1.6g to 2.2g CO_2e per image for models like Midjourney and DALL-E 2.
- **Text Classification:** Significantly lower, around 0.0003 gCO_2e per query (0.3g per 1000 queries).

## B. Task Types & Complexity

The type and complexity of tasks performed by AI systems significantly influence their environmental footprint.

- **Differential Impact of Common AI Tasks:** There is a clear and substantial difference in the environmental cost associated with generative AI tasks versus analytical or discriminative tasks.
    - **Image and Video Generation:** These are among the most energy-intensive AI tasks. Generating 1000 images with AI can produce CO_2 emissions equivalent to driving 4.1 miles (6.6 km) in a gasoline car. AI image generation is reported to be approximately 60 times more energy-intensive than text generation. The WAB Learns estimator uses a figure of 360 kWh per minute for video generation, although this is an estimate.
    - **Text Generation (Long-form vs. Simple Q&A):** While less intensive than image generation, text generation's footprint scales with length and complexity. Simple Q&A might involve fewer tokens and less computational effort than generating long-form content like reports or articles. Multi-turn conversations also accumulate footprint with each turn.
    - **Code Generation:** The energy impact can be influenced by factors like prompt

engineering. Rubei et al. found that for Llama 3 inference in code completion, a zero-shot prompt consumed about 0.000016 kWh per query (GPU energy via CodeCarbon), and specific prompt tags could reduce this.
    - **Summarization, Analysis, Classification:** These tasks generally have lower footprints. For instance, text classification is estimated at 0.002 kWh per 1000 queries, and summarization at 0.049 kWh per 1000 queries. This stark contrast, particularly between generative tasks like image/video creation and analytical tasks like classification, implies that user choices regarding *which type* of AI task to employ can have a more profound effect on the overall environmental footprint than micro-optimizations of a single, less intensive task. The calculator should aim to make these relative impacts clear to the user.
- **Metrics for Task Units:** The choice of metric for task units is crucial for consistent and comparable footprint estimation. Common units include:
    - **Per query:** Widely used for LLM interactions.
    - **Per 1000 tokens (input and/or output):** Often used in AI service pricing and can be a useful proxy for computational work in LLMs. Energy per token can be derived if total tokens and total energy per query are known. Response length in tokens is a major driver of LLM inference energy.
    - **Per image generated:** Standard for image generation models.
    - **Per minute of video/audio generated:** Used for continuous media generation.
    - **Per API call:** A general metric, but the footprint per call is highly dependent on the specific function and parameters of that call.
    - **Per hour of processing:** Can be relevant for sustained workloads or training sessions.
- **Impact of Input/Output Length, Context Window Size, and Iterative Refinement:** These factors significantly modulate the footprint of a given AI task.
    - **Input/Output Length:** Directly correlates with energy consumption, particularly for LLMs. Longer responses require more token generation steps, each incurring computational cost and thus consuming more energy. Maliakel et al. (2025) demonstrated that energy and latency increase with input sequence length for various models and tasks.
    - **Context Window Size:** Larger context windows enable models to process and "remember" more information from the input. While this can improve performance on complex tasks, processing larger contexts can increase the initial computational load (latency to first token) and thus the energy for that phase of inference. Google's Gemini, for example, supports a 2 million token context window, allowing it to process entire books. The energy implications of processing such vast contexts need careful consideration.
    - **Iterative Refinement and Multi-Step Workflows:** The process of refining prompts, engaging in multi-turn conversations, or using AI in an iterative design loop (as often happens in creative tasks or complex problem-solving) leads to a cumulative footprint. Each iteration or conversational turn is effectively an additional query or processing step, adding to the total energy and water consumption. This "silent" or "meta-work" footprint, associated with the human interaction patterns necessary to achieve a desired AI output (as alluded to in the "Long-Form AI Prompting Guide" mentioned in the user query), is often not captured in single-query estimates but is a real component of AI usage. For agentic platforms, this is even more pronounced as the agent itself may perform multiple internal AI calls or tool uses for a single

user request.

## C. Hardware & Infrastructure

The physical hardware and data center infrastructure are fundamental determinants of AI's environmental footprint, encompassing both operational energy/water use and the embodied resources in manufacturing.

### 1. Energy Consumption of Different Hardware (CPUs, GPUs, TPUs)

- **GPUs (Graphics Processing Units):** GPUs are the workhorses of modern AI due to their parallel processing capabilities.
    - **NVIDIA H100:** This GPU is a current industry standard, used for benchmarking by the AI Energy Score. It is noted to be significantly more energy-efficient than its predecessor, the A100 (nearly twice the FLOPs/Watt). Jegham et al. (2025) assume H100/H200/H800 for most proprietary model inferences, with a DGX system critical power of 10.20 kW.
    - **NVIDIA A100:** Widely deployed, with an estimated average system power per accelerator of around 330W during GPT-3 training. Jegham et al. (2025) infer its use for models like GPT-4o mini, GPT-4, and GPT-4 Turbo, with a DGX A100 system critical power of 6.50 kW.
    - **NVIDIA Blackwell Architecture (e.g., B200):** Touted as significantly more efficient than previous generations. NVIDIA claims Blackwell GPUs are 20 times more energy-efficient than CPUs for comparable AI/HPC workloads, and the GB200 Grace Blackwell Superchip offers 25 times better energy efficiency than the Hopper generation (H100) for large LLM inference.
    - General GPU power draw varies with utilization and model. GPUs can consume 10-15 times more energy than CPUs for the same workload if not optimized, but can also complete tasks much faster, potentially leading to overall energy savings for AI.
- **TPUs (Tensor Processing Units):** Google's custom ASICs designed for machine learning.
    - **Google TPU v5e:** Used by Google for model optimization efforts.
    - **Google Trillium (TPU v6):** Stated to be 67% more energy-efficient than the TPU v5e.
    - **Older TPUs (Patterson et al., 2021 ):**
        - TPU v3: Measured system average power per accelerator around 245-310W for various large models during training.
        - TPU v2: Measured system average power per accelerator around 208-296W.
- **CPUs (Central Processing Units, e.g., Intel Xeon, AMD EPYC):** While essential for general compute and managing AI workloads, CPUs are generally less efficient for the highly parallel computations central to deep learning compared to GPUs or TPUs. Specific power consumption figures for CPUs during AI inference tasks were not detailed in the provided snippets, but their role is often as host processors or for tasks not accelerated by specialized hardware. NVIDIA claims its DPUs can offload data center networking and infrastructure functions from CPUs, reducing CPU power consumption by 25%.
- **AMD MI300X:** Mentioned as a competitor to NVIDIA's H100, but specific power

benchmarks were not available in the snippets.

The following table summarizes available energy profiles for common AI hardware.

**Table 4: Energy Profiles of Common AI Hardware**

| Hardware | Type | Typical Power Draw (Watts under AI load) / System Power | Energy Efficiency Metric (Example) | Source(s) |
|---|---|---|---|---|
| NVIDIA H100 (in DGX system) | GPU | System: 10.20 kW | Higher FLOPs/Watt than A100 | |
| NVIDIA A100 (in DGX system) | GPU | System: 6.50 kW; Accelerator: ~330W (GPT-3 training) | Baseline for many comparisons | |
| NVIDIA Blackwell (GB200) | GPU | Not specified | 25x more energy efficient than Hopper for LLMs | |
| Google TPU v3 | TPU | Accelerator: 245-310W (training) | Used for T5, Meena, GShard training | |
| Google Trillium (TPU v6) | TPU | Not specified | 67% more energy-efficient than TPU v5e | |
| Generic CPU | CPU | Lower per chip, but many needed for same AI task | Significantly lower TFLOPs/Watt for AI than GPUs | |

## 2. Embodied Carbon and Water in Hardware Manufacturing, Transport, and Disposal

The environmental footprint of AI hardware extends beyond its operational phase to include the significant resources consumed and emissions generated during its manufacturing, transportation, and eventual disposal. This "embodied" footprint is often overlooked but is a critical component of a full lifecycle assessment (LCA).

- **General Challenges:** Quantifying embodied impacts is difficult due to a lack of transparency from hardware manufacturers. Hardware production can account for over 40% of total data center emissions.
- **Semiconductor Manufacturing:** This is an extremely energy- and water-intensive process. The fabrication of complex chips like GPUs involves numerous steps, significant power, and large quantities of ultrapure water.
- **Raw Material Extraction:** Obtaining the raw materials for semiconductors and other hardware components (e.g., rare earth elements) can involve environmentally damaging mining practices and the use of toxic chemicals.
- **GPU Manufacturing Estimates (from Morrison et al., 2025 ):** This study provides some of the few available detailed estimates for modern GPUs:
    - **Embodied Carbon:** Approximately 463 kg CO_2eq per GPU (derived from an estimate of 3700 kg CO_2eq per 8x GPU server node).
    - **Embodied Water:** Approximately 100.4 liters per H100 GPU (based on TSMC

water usage estimates of 12.33 liters per square centimeter of hardware) plus an additional 2.2 liters for rare earth metal mining, totaling around 102.6 liters per GPU.
    - **Amortized Impact:** Assuming a 4-year GPU lifespan, the amortized embodied impact is estimated at 0.013 kg CO_2eq per GPU-hour and 0.003 liters of water per GPU-hour.
- **Lifecycle Assessment Databases:** Standard LCA databases like Ecoinvent and GaBi exist, but the provided snippets do not confirm their direct application or the availability of specific datasets for the latest AI hardware components within them.

The rapid innovation cycle in AI hardware, with new, more powerful GPU generations appearing every 1-2 years (e.g., NVIDIA's progression from Pascal to Volta, Turing, Ampere (A100), Hopper (H100), and now Blackwell (B200) ), creates a "hardware treadmill." Even if newer chips are more operationally efficient per computation, the frequent replacement of entire server fleets means that the substantial embodied energy and water costs of manufacturing are incurred repeatedly. This recurring embodied footprint can become a dominant factor in the overall lifecycle impact of AI infrastructure, a critical aspect that a calculator focusing solely on operational energy would miss.

**Table 5: Estimated Embodied Environmental Footprint of Key AI Hardware Components**
*(Data primarily from Morrison et al., 2025 and general principles)*

| Hardware Component | Embodied Carbon (kgCO2e/unit) | Embodied Water (Liters/unit) | Assumed Lifespan (years) | Amortized Carbon (kgCO2e/hr of use) | Amortized Water (L/hr of use) | Source(s) |
|---|---|---|---|---|---|---|
| High-Performance GPU (e.g., NVIDIA H100 type) | ~463 | ~102.6 | 3-4 (typical refresh cycle) | ~0.013 (based on 4yr, full use) | ~0.003 (based on 4yr, full use) | |
| Server CPU | Lower than GPU, but varies | Lower than GPU, but varies | 3-5 | Highly variable | Highly variable | General LCA principles, data scarce |
| Server Rack (populated) | Several thousand (e.g., 3700 for 8xGPU node) | Hundreds to thousands | 4-8 | Highly variable | Highly variable | (for GPU node), data scarce |

## 3. Data Center Specifics

Data centers are the physical backbone of most AI computation, and their efficiency characteristics are paramount.

- **Power Usage Effectiveness (PUE):** A key metric representing the ratio of total facility energy to IT equipment energy. A lower PUE indicates higher efficiency.
    - **Google:** Reports a fleet-wide TTM PUE of 1.09 for Q1 2025. Individual campus PUEs vary, for example, Central Ohio (Columbus) at 1.05, while Changua County, Taiwan is at 1.12 for Q1 2025 TTM.
    - **AWS:** Reported a global PUE of 1.15 in 2023.
    - **Microsoft Azure:** While committed to efficiency, specific recent PUE values were not found in the provided snippets.

    - **Industry Average:** Historically around 1.58 , highlighting that major cloud providers operate significantly more efficiently. PUE is a necessary factor in calculating operational energy but is not sufficient for a complete environmental assessment. It indicates how efficiently energy is delivered to IT equipment but says nothing about the *source* of that energy (i.e., its carbon intensity) or the *water efficiency* of the facility. A data center with an excellent PUE located in a region heavily reliant on coal and experiencing high water stress could have a more detrimental overall environmental impact than a facility with a slightly higher PUE that is powered by renewables in a water-abundant area. The calculator must therefore consider PUE in conjunction with grid carbon intensity and water usage factors.
- **Water Usage Effectiveness (WUE) and Local Water Stress:**
    - **WUE:** Measures the water used per kilowatt-hour of IT energy, encompassing on-site cooling (Scope 1 water use, e.g., for cooling towers) and off-site water use for electricity generation (Scope 2 water use). Some also consider embodied water in hardware (Scope 3).
    - **Provider Commitments:** AWS and Azure aim to be "water positive" by 2030. Google aims to replenish 120% of the freshwater it consumes by 2030.
    - **Quantitative Data:** Specific, comparable WUE values are less commonly reported by providers than PUE. Jegham et al. (2025) use estimated WUEsite (on-site cooling water per IT kWh) and WUEsource (source electricity water per IT kWh) values for their calculations (e.g., for Microsoft Azure US data centers: WUEsite 0.30 L/kWh, WUEsource 3.142 L/kWh).
    - **Local Water Stress:** Data centers can impose significant stress on local water resources, especially in arid or drought-prone regions. Google's data centers, for instance, consumed 5.6 billion gallons of water globally in 2021. The WRI Aqueduct Water Risk Atlas is a tool that can map these water stress levels. To reflect the true environmental impact, a water footprint calculation should ideally weight water consumption by the local water scarcity. This means a liter of water consumed in a highly water-stressed region has a greater environmental consequence than a liter consumed in a water-abundant region.
- **Cooling Methods and Their Water Use:**
    - Data center cooling is a major consumer of both energy and water.
    - **Evaporative Cooling:** Traditional cooling towers use water evaporation and can consume vast quantities – estimates range from 2 liters to as much as 9 liters of water per kWh of energy used by the data center. Much of this water is lost to evaporation.
    - **Liquid Immersion Cooling:** An advanced technique where servers are submerged in dielectric fluid. This can significantly reduce or even eliminate water consumption for cooling and improve energy efficiency.
    - **Adiabatic Cooling:** Uses outside air for cooling when ambient temperatures allow, reducing water use. Microsoft Azure is deploying this in some new regions.
    - Data centers require clean, treated water for their cooling systems to prevent blockages and bacterial growth.

The actual environmental footprint of an AI task emerges from a complex interplay of these factors: the specific AI model's efficiency, the hardware it runs on (both its operational efficiency and embodied footprint), the data center's PUE and WUE, the cooling technologies employed, and the geographical context (grid carbon intensity and local water stress). These elements are not merely additive; they are interdependent. For instance, a highly efficient AI model run on

older, less efficient hardware in a data center powered by a carbon-intensive grid and located in a water-stressed region might have a worse overall footprint than a moderately efficient model run on the latest hardware in a data center powered by renewables in a water-abundant region. The calculator's methodology must strive to capture these interdependencies.

## D. Geographical Factors & Energy Source Factors

The geographical location where AI computation occurs and the characteristics of the energy sources used are critical determinants of the final environmental footprint, particularly for carbon emissions and water impact.

- **Grid Carbon Intensity (gCO_2eq/kWh):** The carbon intensity of the electricity grid measures the amount of greenhouse gas emissions produced per unit of electricity generated. This factor varies dramatically across different countries and even within regions of the same country, depending on the local energy mix (e.g., reliance on coal, natural gas, nuclear, renewables).
    - **Examples:** Values can range from as low as ~20 gCO_2eq/kWh in regions with high hydropower or nuclear generation (e.g., Norway, Switzerland) to over 800 gCO_2eq/kWh in areas heavily reliant on fossil fuels (e.g., parts of Australia, South Africa, and the USA).
    - The WAB Learns AI Environmental Footprint Estimator uses an average of 0.367 kgCO_2/kWh for the USA (2023, from EIA) and 0.582 kgCO_2/kWh for China (2023, from Statista).
    - Jegham et al. (2025) use specific grid factors in their model footprint calculations: 0.3528 kgCO_2e/kWh for Microsoft Azure US data centers, 0.6 kgCO_2e/kWh for DeepSeek in China, and 0.385 kgCO_2e/kWh for AWS US data centers.
    - **Data Sources:** Key sources for this data include the International Energy Agency (IEA), Ember Climate, the U.S. Environmental Protection Agency (EPA), and Our World in Data. A comprehensive calculator will need to compile and regularly update a table of these factors for relevant geographical regions.

  The temporal granularity of carbon intensity data is also an important consideration. While many calculators use average annual grid carbon intensity figures, these can mask significant hourly or daily fluctuations in the actual carbon content of electricity due to varying contributions from intermittent renewables (like solar and wind) and baseload power plants. Google, for its internal calculations, utilizes hourly greenhouse gas emission factors matched with hourly electricity load data to achieve a more precise measure of location-based emissions. This allows for optimizing workload timing to coincide with periods of lower grid carbon intensity. While accessing real-time hourly data may be challenging for a general-purpose calculator, acknowledging this variability and using the most granular available data (e.g., regional or national annual averages) is crucial.
- **Availability and Impact of Renewable Energy Sourcing:** The use of renewable energy can significantly reduce the operational carbon footprint of AI workloads. AI providers employ various strategies for renewable energy sourcing:
    - **24/7 Carbon-Free Energy (CFE) Matching:** Pioneered by companies like Google , this approach aims to match electricity consumption with carbon-free energy sources on an hourly basis within the same grid region. This is considered highly impactful because it ensures that clean energy is actually being generated and used when and where the AI workload is running, directly displacing fossil fuel generation on that local grid.

- **Renewable Energy Certificates (RECs) and Power Purchase Agreements (PPAs):** These are common mechanisms used by many cloud providers to claim renewable energy use. RECs represent the environmental attributes of renewable electricity generation, and PPAs are long-term contracts to buy electricity from renewable projects. While these contribute to the overall greening of the grid and support renewable energy development, their direct impact on the emissions of a specific AI workload at a specific time can be less certain than 24/7 CFE matching, especially if the RECs are "unbundled" (traded separately from the electricity) or sourced from distant grids.
- **Impact:** The use of 100% carbon-free energy sources can dramatically reduce, or even eliminate, the operational carbon footprint of electricity consumption. For example, the OLMo 7B model trained on the LUMI supercomputer in Finland, which runs entirely on hydroelectric power, was reported to have zero carbon emissions from its electricity use during training.

When considering renewable energy claims, the concept of "additionality" is important. A truly impactful renewable energy procurement strategy is one that leads to the development of new renewable energy capacity that would not have been built otherwise. Generic RECs have faced criticism regarding their additionality. The 24/7 CFE matching approach inherently has stronger additionality as it requires sourcing sufficient clean energy in real-time on the local grid where consumption occurs. For the AI footprint calculator, particularly when suggesting offsets or evaluating the impact of running workloads in "green" data centers, understanding the nature and credibility of these renewable energy claims will be important.

The following table should be compiled with the latest available data for key regions relevant to AI computation.

**Table 6: Grid Carbon Intensity Factors for Key Regions**

| Country/Region | Grid Carbon Intensity (gCO2eq/kWh) | Year of Data | Source (e.g., IEA, Ember, EPA) |
|---|---|---|---|
| USA (Average) | 367 | 2023 | EIA (via ) |
| China (Average) | 582 | 2023 | Statista (via ) |
| EU (Average) | *Populate with recent data* | *Year* | *Source* |
| France | *Populate with recent data* | *Year* | *Source* |
| Germany | *Populate with recent data* | *Year* | *Source* |
| UK | *Populate with recent data* | *Year* | *Source* |
| India | *Populate with recent data* | *Year* | *Source* |
| Canada (Average) | *Populate with recent data* | *Year* | *Source* |
| Brazil | *Populate with recent data* | *Year* | *Source* |
| Australia | *Populate with recent data* | *Year* | *Source* |
| Japan | *Populate with recent data* | *Year* | *Source* |

| Country/Region | Grid Carbon Intensity (gCO2eq/kWh) | Year of Data | Source (e.g., IEA, Ember, EPA) |
|---|---|---|---|
| South Korea | *Populate with recent data* | *Year* | *Source* |
| *Other key regions...* | *Populate with recent data* | *Year* | *Source* |

## E. Usage Parameters (User Inputs for the Calculator)

To calculate the footprint of a specific AI activity, the user will need to provide several key parameters. These inputs are fundamental to any calculator, as seen in existing tools , and their required granularity will depend on the chosen calculation methodologies (discussed in Section IV). Essential usage parameters include:

1. **For LLM tasks:**
    - Selected LLM (e.g., GPT-4o, Claude 3 Sonnet).
    - Number of queries or interactions.
    - Average number of input tokens per query.
    - Average number of output tokens per query.
    - Context window size used (if significantly impacting performance/energy).
2. **For Image/Video/Audio Generation tasks:**
    - Selected generation model (e.g., Stable Diffusion XL, DALL-E 3, Sora).
    - Number of images generated.
    - Complexity of images (e.g., resolution, detail level, number of generation steps if known/controllable by user).
    - Duration of video/audio generated (in minutes or seconds).
    - Complexity of video/audio (e.g., resolution, frame rate, content type).
3. **For Agentic Platform Usage:**
    - Selected agentic platform/orchestrator model.
    - Description of the overall task or project.
    - If possible, number of LLM calls made by the agent and their token counts.
    - If possible, number and type of tools/APIs invoked by the agent.
    - (This area will likely require more abstract inputs like "duration of agent session" or "complexity of agent task" if detailed tracing is unavailable, leading to higher uncertainty estimates).
4. **General Parameters:**
    - Geographical location of computation (country/region, or specific cloud provider region).
    - Type of hardware used (if known, e.g., specific GPU type; otherwise, model choice may imply typical hardware).
    - Duration of the AI task or project (if time-based rather than unit-based).
    - Total computational time (if known from logs, e.g., GPU hours).

The calculator must guide the user in providing these inputs, possibly offering defaults or ranges if exact values are unknown, and clearly communicating how these inputs affect the final footprint estimate.

# IV. Data Requirements & Sources

Developing a robust AI carbon and water footprint calculator necessitates access to a wide

array of specific data points and reliable sources. Given the current landscape of data opacity and rapid technological evolution, a multi-pronged strategy for data acquisition, including the use of direct measurements, academic research, provider reports, and intelligent proxies, is essential.

**A. Collection of Specific Energy (kWh or Wh per task unit) and Water (Liters per task unit) Consumption Figures:**

This is the core data required. As detailed in Section III.A, figures vary significantly.

- **LLM Inference:** Data like Wh/query, L/query (e.g., from Jegham et al. (2025) ) for various models (GPT series, Claude, Llama, DeepSeek) across different prompt lengths. Per-token data is less common but can be derived if TPS and query energy are known.
- **Image/Video/Audio Generation:** Data like kWh/1000 images, L/image (e.g., for Stable Diffusion, DALL-E, Midjourney from sources like ). Video/audio data is scarcer (e.g., WAB's estimate of 360 kWh/minute for video ).
- **Model Training:** Total MWh and total Liters for training specific models (e.g., GPT-3: ~1287 MWh, ~0.7-5.4 million Liters water ; Llama 2 7B: 81 MWh ; OLMo 2 13B: 892 kL water ). This is more for contextual understanding than direct user calculation of training, unless the tool supports estimating training for custom models (an advanced feature).
- **Hardware Operation:** Power draw (Watts) for specific GPUs (H100, A100), TPUs, CPUs under typical AI loads.
- **Data Center Operations:** PUE values and WUE values (on-site and source electricity, e.g., from ) for major cloud providers and regions.

**B. Carbon Emission Factors (CO_2e per kWh):**

- Grid carbon intensity factors (gCO_2eq/kWh or kgCO_2eq/kWh) for various countries and regions are essential to convert electricity consumption into carbon emissions.
- Sources: National reports (e.g., EPA for the US ), international bodies (IEA), research organizations (Ember Climate, OurWorldInData) [User Query III.E].
- Examples: USA ~367 gCO_2eq/kWh, China ~582 gCO_2eq/kWh (2023 estimates ). These need to be regularly updated.

**C. Lifecycle Assessment (LCA) Databases for Hardware and Infrastructure:**

- Embodied carbon and water data for manufacturing GPUs, CPUs, servers, and data center construction.
- This data is notoriously difficult to obtain directly from manufacturers.
- Academic LCAs and estimations (e.g., Morrison et al. (2025) for H100 GPUs: ~463 kgCO_2eq/GPU, ~102.6 L/GPU ) will be primary sources.
- Standard LCA databases like Ecoinvent or GaBi might contain some relevant data for materials or generic electronic components, but specific AI hardware data is often proprietary or not yet included.

**D. Strategies for Identifying and Utilizing Proxy Data or Benchmarks When Direct Data is Unavailable:**

Given data opacity, especially for proprietary models and new hardware, proxy data and benchmarks are crucial.

- **Model Analogies:** If data for a specific model (e.g., "NewModel X") is unavailable, use data from a known model with similar architecture, parameter count, and release date as a proxy, clearly stating this assumption and applying an uncertainty range.
- **Performance Benchmarks as Proxies for Energy:**
    - **MLPerf** : Industry-standard benchmark for ML performance (training and inference). While primarily focused on speed/accuracy, it has started including power measurement. MLPerf Power results (energy per inference, inferences per joule)

can serve as a proxy for relative energy efficiency if direct energy consumption data for a specific task/model isn't available.
    - **AI Energy Score** : Provides relative energy efficiency for inference. While not absolute Wh/query, the star rating or underlying (normalized) energy scores could be used to scale estimates from a known model.
- **Extrapolation from Simpler/Older Models:** Data from older or smaller models (e.g., energy per parameter or per FLOP for training) can be cautiously extrapolated to newer, larger models, though scaling is often non-linear.
- **Hardware Datasheets (TDP):** Thermal Design Power (TDP) from datasheets is often used as a proxy for maximum power draw, but actual power consumption under AI workloads can be lower or vary significantly. Measured power is preferred (as in ).
- **Academic Research:** Continuously monitor and incorporate findings from peer-reviewed studies and pre-print archives (like arXiv) that benchmark new models or hardware (e.g.).
- **Cloud Provider Billing Data:** For users running workloads on the cloud, their billing data (if it provides compute hours for specific VMs/accelerators) can be combined with hardware power profiles and PUE to estimate operational energy, even if the cloud provider's own carbon tool is not used or lacks granularity.

**E. Publicly Available Research, Academic Studies, Reports, and Recognized Benchmarks:**

This forms the backbone of the calculator's database.

- **Academic Studies:**
    - **Key Papers for LLM Inference:** Jegham et al. (2025) "How Hungry is AI?" (energy, water, carbon for 30 LLMs). Maliakel et al. (2025) "Investigating Energy Efficiency..." (energy for 6 LLMs across tasks, DVFS impact). Rubei et al. "Prompt engineering..." (Llama 3 energy for code generation).
    - **Key Papers for Training/Lifecycle:** Strubell et al. (2019) "Energy and Policy Considerations..." (BERT, Transformer training). Patterson et al. (2021) "Carbon Emissions and Large Neural Network Training" (GPT-3, T5 training). Morrison et al. (2025) "Holistically Evaluating..." (OLMo, Llama 2 lifecycle). Li et al. (2023) "Making AI Less 'Thirsty'..." (AI water footprint methodology, GPT-3 water).
    - **Key Papers for Image/Video:** Luccioni et al. (cited in ) for image generation. Karaarslan & Aydin (2024) review text-to-video models including Sora, noting computational intensity.
- **Organizational Reports:**
    - **IEA (International Energy Agency):** Reports on data center energy consumption trends, grid intensity.
    - **Greenpeace, OECD.AI:** General reports on AI sustainability, policy considerations [User Query III.E].
    - **AI Provider Sustainability Reports:** (Google, AWS, Microsoft, Meta, NVIDIA) provide corporate goals, PUE, some high-level AI impact data (see Section II.B).
- **Recognized Benchmarks:**
    - **MLPerf Power:** For hardware energy efficiency during training/inference.
    - **AI Energy Score:** For relative inference energy efficiency of models.

The process of data collection must be ongoing, with a system for regularly scanning these sources for new data and updates.

# V. Calculation Methodologies & Estimation

# Techniques

To provide users with meaningful estimates of their AI carbon and water footprint, robust calculation methodologies are required. These methodologies must combine user inputs with the collected data on AI components, infrastructure, and geographical factors. Given the inherent uncertainties and data gaps, intelligent estimation techniques are also crucial.

**A. Formulas to Calculate Environmental Impacts:**

1. **Operational Energy Consumption (E_{op}):** The core of the footprint calculation is determining the energy consumed by the AI workload. E_{op} = (\frac{E_{model,task}}{U_{task}}) \times N_{tasks} \times F_{hardware} \times PUE
   Where:
   - E_{model,task} is the energy consumed by the specific AI model for a defined task unit (e.g., Wh/query, Wh/image, Wh/token). This data comes from sources like Jegham et al. or Luccioni et al..
   - U_{task} is the number of base units in E_{model,task} (e.g., if E_{model,task} is Wh per 1000 queries, U_{task} is 1000).
   - N_{tasks} is the number of task units performed by the user (e.g., number of queries, images generated).
   - F_{hardware} is a hardware efficiency factor. If E_{model,task} is already benchmarked on specific hardware (e.g., H100 by AI Energy Score ), and the user's (assumed) hardware is different, this factor adjusts for the difference in energy efficiency (e.g., A100 might be 1.5x less efficient than H100 for a given task). If E_{model,task} is hardware-agnostic (rare), then F_{hardware} would be the energy consumption of the user's hardware per fundamental operation, and N_{tasks} would be total operations. More practically, specific E_{model,task,hardware} values should be used.
   - PUE (Power Usage Effectiveness) of the data center where the computation occurs. This accounts for energy used by cooling, lighting, and power distribution in addition to IT equipment.
     Alternatively, if computational time is known: E_{op} = P_{hardware} \times T_{compute} \times PUE Where:
   - P_{hardware} is the average power draw of the IT hardware (e.g., GPU server) during the AI task (Watts).
   - T_{compute} is the total computational time for the user's tasks (hours).
2. **Operational Carbon Footprint (C_{op}):** C_{op} = E_{op} \times CI_{grid} Where:
   - CI_{grid} is the carbon intensity of the electricity grid in the region of computation (gCO_2eq/kWh or kgCO_2eq/kWh). This factor is adjusted if the provider uses certified renewable energy (e.g., 24/7 CFE matching could reduce CI_{grid} effectively to near zero for that portion of energy ).
3. **Operational Water Consumption (W_{op}):** Water consumption has two main components: on-site for cooling and off-site for energy generation. W_{op} = W_{cooling} + W_{energy_source} W_{cooling} = E_{IT} \times WUE_{site} = (\frac{E_{op}}{PUE}) \times WUE_{site} W_{energy_source} = E_{op} \times WUE_{source} So, W_{op} = (\frac{E_{op}}{PUE}) \times WUE_{site} + E_{op} \times WUE_{source} Where:
   - E_{IT} is the energy consumed directly by IT equipment (E_{op} / PUE).
   - WUE_{site} is the on-site Water Usage Effectiveness (e.g., Liters/kWh of IT energy for cooling). This depends on cooling technology (e.g., evaporative cooling uses

more water than liquid immersion or efficient air cooling ).
    - WUE_{source} is the water consumed at the source of electricity generation (e.g., Liters/kWh of total energy consumed by the data center). This varies greatly depending on the energy mix (e.g., thermal power plants have high water consumption, while solar PV and wind have very low operational water use).
    - The water footprint should also be contextualized by local water stress levels. This might involve applying a water stress index multiplier to W_{op} to reflect the true environmental burden.
4. **Embodied Energy, Water, and Carbon from Hardware (E_{emb}, W_{emb}, C_{emb}):** This accounts for the environmental impact of manufacturing, transporting, and disposing of the hardware. C_{emb, prorated} = (\frac{C_{unit}}{L_{unit} \times H_{year}}) \times T_{usage} Similar formulas apply for E_{emb, prorated} and W_{emb, prorated}. Where:
    - C_{unit} is the total embodied carbon to produce one unit of hardware (e.g., one GPU server or one GPU).
    - L_{unit} is the average lifespan of the hardware unit (years).
    - H_{year} is the average operational hours per year for such hardware.
    - T_{usage} is the duration the user's AI task utilized that type of hardware (hours). This prorated embodied impact is then added to the operational impact for a more complete lifecycle view. Data for C_{unit} is scarce and often estimated.

**B. Methods for Aggregating Impacts for Multi-Step Projects or Agentic Platform Usage:**
For complex AI projects involving multiple steps or the use of agentic platforms that orchestrate several AI models/tools, the total footprint is the sum of the footprints of individual components and the orchestrator's overhead.

- **Step 1: Deconstruct the Workflow:** Identify each distinct AI call (LLM query, image generation, tool API call) and each significant processing step performed by the orchestrator LLM (planning, tool selection, result synthesis).
- **Step 2: Calculate Footprint for Each Component:**
    - For each AI model call: Use the formulas in V.A with the specific model, task parameters, and assumed hardware/location.
    - For each external tool API call: If the tool itself has a known footprint per API call (rarely available), use that. Otherwise, this might be a data gap requiring estimation or exclusion with clear caveats.
    - For orchestrator LLM processing: Estimate the number and token count of internal LLM calls made by the orchestrator for its decision-making. Calculate their footprint using LLM inference formulas. This is the "orchestrator overhead."
- **Step 3: Sum the Impacts:** Aggregate the energy, carbon, and water footprints from all components. Impact_{total} = \sum Impact_{orchestrator_calls} + \sum Impact_{tool_calls}The CarbonCall framework, for example, aims to optimize function calling in agentic systems on edge devices by dynamically selecting tools and adjusting power based on carbon intensity. While it reduces overall footprint, estimating the specific overhead of its orchestration logic versus the tools would require detailed internal profiling not typically exposed. For cloud-based agentic platforms, users would need detailed logs of all underlying LLM calls and tool invocations to perform this aggregation accurately.

**C. Approaches for Intelligent Estimation:**
Given the pervasive data gaps and uncertainties, particularly for new models, proprietary systems, and embodied impacts, the calculator must employ intelligent estimation techniques.

1. **Developing Plausible Ranges (Min/Max/Average) for Key Parameters:**

- Instead of single point estimates, provide outputs as a range (e.g., "Your estimated carbon footprint is 50-150 gCO_2eq").
- **Sources of Ranges:**
  - **Model Efficiency:** Use data from benchmarks like AI Energy Score or studies like Jegham et al. which show variability (e.g., GPT-4.1 nano 0.103 Wh/query vs. o3 39.223 Wh/query for similar prompt types). The min could be based on the most efficient known model of a similar class, and the max on a less efficient one or by adding uncertainty margins.
  - **Hardware Power:** Use TDP as a max power draw, and measured average power under load (if available from benchmarks like ) as an average or min.
  - **PUE/WUE:** Use provider-specific values if known for a region, or a range from best-case (e.g., 1.05 for Google ) to average industry PUE (~1.58 ) if location/provider is unknown.
  - **Grid Carbon Intensity:** Use country/regional averages, but acknowledge hourly variations by potentially providing a range based on typical peak/off-peak intensities if that data is available for a region.
  - **Embodied Impacts:** These have high uncertainty. Use ranges from different LCA studies or apply significant error margins to point estimates (e.g., Morrison et al.'s estimate for H100 embodied carbon could be a central value in a wider range).

2. **Sensitivity Analysis to Identify Most Impactful Variables:**
   - Internally, conduct sensitivity analysis to understand which input parameters (e.g., model choice, usage volume, location, hardware efficiency) have the most significant influence on the final footprint estimates.
   - This helps prioritize data collection efforts (focus on parameters with high impact and high uncertainty) and can inform user interface design (e.g., highlight to users which of their choices will most drastically alter the footprint).
   - For example, the choice of image generation vs. text classification has a massive impact. Similarly, running a workload in a region with high carbon intensity vs. low carbon intensity will yield vastly different carbon footprints for the same energy consumption.
3. **Clearly Stating Assumptions and Uncertainty Levels in the Output:**
   - **Transparency is paramount.** For every calculation, the tool must clearly state:
     - The source of the data used for E_{model,task}, P_{hardware}, CI_{grid}, PUE, WUE, embodied values.
     - Any proxy data used and the rationale.
     - The specific assumptions made (e.g., assumed hardware if not specified by user, assumed PUE/WUE if location is generic).
     - The level of uncertainty associated with the estimate (e.g., qualitative: "high confidence," "medium confidence," "low confidence/highly uncertain," or quantitative if possible via error propagation).
   - The WAB Learns estimator's disclaimer about being for "educational illustration only" due to data volatility is a good example of communicating limitations. The HCI GenAI CO2ST Calculator also emphasizes that results are estimates and err on the conservative side.

**D. Consideration of "Rebound Effects" or Jevons Paradox:**

While not directly part of the calculation, the tool could include a contextual note about rebound effects. As AI becomes more efficient and cheaper, its usage might increase, potentially

offsetting or even negating the environmental benefits of efficiency gains at a macro level. This is an important awareness point for users, encouraging mindful AI use beyond just selecting efficient options.

**E. Lifecycle Perspective - Practical Incorporation:** Incorporating a full LCA for every AI task is impractical for a user-friendly calculator. However, elements can be included:

- **Operational Phase:** This is the primary focus, calculating energy, water, and carbon from AI model execution and data center operations.
- **Hardware Manufacturing (Embodied Footprint):** Include prorated embodied carbon/water from hardware (Section V.A.4). This requires estimates for manufacturing impacts of GPUs/CPUs/servers and their lifespans. The data from Morrison et al. (2025) for OLMo models, which includes hardware manufacturing estimates, is a key reference.
- **Hardware Disposal (E-waste):** While hard to quantify per task, general information about the e-waste problem from rapid hardware obsolescence in AI can be provided contextually.
- **Data Center Construction:** Generally, the embodied energy/carbon of data center construction is amortized over its lifespan and often included in the PUE or facility overheads by providers, rather than being a direct per-task calculation for the user.

The key is to balance comprehensiveness with usability, clearly indicating which lifecycle stages are included in the estimate and which are discussed for broader context.

# VI. Carbon and Water Credits

Once the carbon and water footprints are calculated, the tool can suggest corresponding offset credits. This section outlines the translation of footprint into credits and information on sourcing and verification.

**A. Translation of Calculated Footprint into Offset Credits:**

- **Carbon Footprint:**
  - Typically measured in tonnes of CO_2 equivalent (tCO_2e).
  - Conversion: 1 carbon credit typically represents the reduction or removal of 1 tonne of CO_2e from the atmosphere.
  - So, if the calculated carbon footprint is X kgCO_2e, the number of carbon credits needed is X / 1000.
- **Water Footprint:**
  - Typically measured in cubic meters (m^3) or Liters (L) of water consumed or withdrawn.
  - Water offset credits (e.g., Water Restoration Certificates (WRCs) or similar volumetric water benefits) aim to restore or replenish a certain volume of water, often in water-stressed basins.
  - Conversion: 1 WRC might represent 1,000 gallons (approximately 3.785 m^3) or 1 m^3 of water restored or replenished, depending on the specific credit type and standard. The tool would need to use a defined conversion factor based on the type of water credit being referenced.

**B. Information on Types of Credits:**

- **Carbon Credits:**
  - **Avoidance/Reduction Credits:** Generated from projects that reduce emissions compared to a baseline (e.g., renewable energy projects displacing fossil fuels, energy efficiency improvements, avoided deforestation).

    - **Removal Credits:** Generated from projects that actively remove CO_2 from the atmosphere (e.g., afforestation/reforestation, direct air capture (DAC), bioenergy with carbon capture and storage (BECCS), soil carbon sequestration). Removal credits are often considered higher quality for offsetting residual emissions.
    - **Nature-based solutions** (e.g., forestry, mangroves) vs. **Technology-based solutions** (e.g., DAC). Xpansiv CBL marketplace, for instance, has standardized contracts like N-GEO for nature-based and C-GEO for technology-based credits.
- **Water Credits/Offsets:**
    - **Volumetric Water Benefits (VWBs):** These include activities like watershed restoration, agricultural water efficiency improvements, and municipal water infrastructure upgrades that result in a measurable volume of water being restored to ecosystems or made available for other uses.
    - **Water Restoration Certificates (WRCs):** A specific type of VWB, often representing a volume of water restored to a dewatered ecosystem.
    - **Alliance for Water Stewardship (AWS) Water Positive:** While not a credit, this is a framework some companies (like AWS itself ) use to commit to returning more water than they consume, often through a portfolio of water stewardship projects. This might involve purchasing VWBs.

**C. Verification Standards and Bodies for Credits:**

Credibility is paramount for offset credits. Robust verification by independent third parties against recognized standards is essential.

- **Carbon Credit Standards:**
    - **Verified Carbon Standard (VCS) by Verra:** One of the largest and most well-known global standards for voluntary carbon market projects.
    - **Gold Standard:** Focuses on projects with strong sustainable development co-benefits in addition to carbon reductions.
    - **American Carbon Registry (ACR):** A leading carbon offset standard in the U.S..
    - **Climate Action Reserve (CAR):** Another prominent U.S.-based standard.
    - **Puro.earth:** Specializes in engineered carbon removal credits.
    - **United Nations Clean Development Mechanism (CDM):** While primarily for compliance markets, some CDM credits or methodologies might be relevant. Carbon TradeXchange (CTX) lists UN CDM credits.
- **Water Credit Standards/Frameworks:**
    - The market for water credits is less mature than for carbon credits, but standards are emerging.
    - **Volumetric Water Benefit Accounting (VWBA):** A methodology developed by WRI, WBCSD, and others to quantify the benefits of water stewardship activities.
    - **AWS Water Positive Methodology :** Amazon's internal methodology for its water stewardship claims.
    - Verification often involves project-specific assessments by environmental consultancies or specialized bodies focusing on water resource management.

**D. Marketplaces and Providers for Credits:**

Users seeking to offset their AI footprint can be directed to reputable marketplaces or providers.

- **Carbon Credit Marketplaces/Exchanges:**
    - **Xpansiv CBL:** A dominant spot exchange for voluntary carbon credits, connecting with major registries.
    - **AirCarbon Exchange (ACX):** Uses blockchain technology for transparent pricing and instant settlement of carbon credits.

    - **Carbon TradeXchange (CTX):** A global electronic exchange for voluntary carbon market credits, operating since 2009.
    - **Intercontinental Exchange (ICE):** Offers trading in environmental markets, including carbon.
    - **Carbonplace:** A platform often involving financial institutions for carbon credit transactions.
    - **Toucan Protocol:** Facilitates tokenization and trading of carbon credits on decentralized exchanges (blockchain-based).
- **Water Credit Providers/Platforms:**
    - **Bonneville Environmental Foundation (BEF) Water Restoration Certificates® (WRCs®):** A well-established provider of WRCs in the U.S.
    - **Water Funder Initiative (WFI) / WaterFunds:** Support local water funds that invest in watershed conservation.
    - Platforms or brokers specializing in environmental commodities may also offer water credits or VWBs.
- **Retail Offset Providers:** Many organizations offer portfolios of carbon offsets (and sometimes water projects) to individuals and businesses. The calculator should advise users to check for the underlying standards and project types offered.

The calculator should emphasize the importance of due diligence when purchasing credits, focusing on project quality, additionality, permanence (for carbon removal), leakage prevention, and co-benefits. It can provide links to the websites of major standards bodies for users to learn more.

# VII. Tool Design & User Experience Considerations

The success of the AI Carbon/Water Footprint Calculator will depend not only on the robustness of its underlying methodologies but also on its usability and the clarity of its communication.

**A. Essential User Inputs for the Calculator:**

Based on the quantification factors identified in Section III and the review of existing calculators (Section II.A), the following inputs are essential for a meaningful calculation:

1. **AI Model Selection:**
    - Dropdown list of common LLMs (GPT series, Claude series, Llama series, Gemini, Mistral, etc.), image/video models (Stable Diffusion, DALL-E, Midjourney, Sora), and potentially generic categories for "other classification model," "other translation model."
    - Option for "Custom Model" if user has specific energy/FLOPs data (advanced).
2. **Task Type:**
    - Dropdown list relevant to the selected model type (e.g., for LLMs: Simple Q&A, Long-form Text Generation, Code Generation, Summarization, Translation; for Image Models: Image Generation).
    - For agentic platforms: A general "Complex Workflow" or specific common agentic tasks.
3. **Usage Volume / Task Units:**
    - **LLMs:** Number of queries, average input tokens per query, average output tokens per query.
    - **Image Models:** Number of images generated.
    - **Video/Audio Models:** Minutes of content generated.

    - **Agentic Platforms:** Number of primary tasks completed, or duration of use.
4. **Location of Computation:**
    - Country and ideally specific cloud provider region (e.g., "USA - AWS us-east-1," "Germany - Google Cloud europe-west3"). This is crucial for selecting appropriate grid carbon intensity and potentially regional PUE/WUE data.
    - Option for "Global Average" or "Unknown" if location is not specific, which would use broader average emission factors and carry higher uncertainty.
5. **(Optional/Advanced) Hardware Specifics:**
    - If the user knows the primary hardware used (e.g., "NVIDIA H100," "NVIDIA A100," "Google TPU v5"). This allows for more precise energy calculations. If not provided, the tool can use default hardware assumptions based on the selected model and provider region.

**B. Desirable Outputs:**

The calculator should provide clear, actionable, and understandable results:

1. **Total Carbon Footprint:**
    - In kgCO_2e or tCO_2e.
    - Equivalent in relatable terms (e.g., miles driven by an average car, number of trees needed to absorb, smartphone charges – as seen in ).
2. **Total Water Footprint:**
    - In Liters or m^3.
    - Equivalent in relatable terms (e.g., number of showers, bottles of water, portion of an Olympic swimming pool ).
    - Contextualization with local water stress level for the specified region, if possible (e.g., "This water use occurs in a region with [High/Medium/Low] water stress").
3. **Total Energy Consumption:**
    - In kWh or MWh.
4. **Breakdown of Impact (Pie chart or bar graph):**
    - **Operational Footprint:**
        - Carbon: Contribution from model energy use, data center overhead (PUE).
        - Water: Contribution from on-site cooling, electricity generation.
    - **Embodied Footprint (if calculated):**
        - Prorated carbon/water from hardware manufacturing.
    - This helps users understand the primary drivers of their AI activity's footprint.
5. **Equivalent Offset Credits Needed:**
    - Number of carbon credits (tonnes CO_2e).
    - Volume of water credits/restoration (Liters or m^3).
6. **Plausible Range of Estimates:**
    - Present results not as a single definitive number but as a range (e.g., "Carbon Footprint: X - Y kgCO_2e") to reflect inherent uncertainties.
7. **Actionable Recommendations (Contextual):**
    - Based on the inputs, suggest ways to reduce footprint (e.g., "Consider using Model Y which is more efficient for this task," "Generating fewer/smaller images can significantly reduce impact," "Running this workload in Region Z with lower grid carbon intensity could save X% emissions").
    - Tips for efficient AI use (e.g., from WAB Learns : plan questions, combine queries, avoid superficial uses).

**C. How to Communicate Assumptions, Data Sources, and Uncertainty to the User:**

Transparency is crucial for building trust and ensuring users understand the nature of the

estimates.

1. **Dedicated "Methodology" or "About Our Calculations" Section/Link:**
    - Clearly explain the formulas used (Section V.A).
    - List primary data sources for model consumption figures, emission factors, PUE/WUE values (e.g., "Model data based on Jegham et al., 2025; Grid intensity from IEA 2024; PUE from Google Cloud 2025 reports").
    - Detail how embodied carbon/water is estimated and its limitations.
2. **Input-Specific Assumptions:**
    - If the user doesn't provide certain inputs (e.g., hardware type), the tool should state the default assumption made (e.g., "Assuming use of NVIDIA H100 equivalent GPU for this model").
    - If a generic location is chosen, state the average emission factors used.
3. **Output-Specific Caveats:**
    - Accompany results with a clear statement about the estimate range and confidence level (e.g., "This is an estimate. Actual impact may vary. The primary sources of uncertainty for this calculation are X and Y.").
    - If proxy data was used for a specific model, this should be noted (e.g., "Data for Model Z is based on the profile of similar Model A due to lack of direct measurements for Model Z.").
4. **Visual Cues for Uncertainty:**
    - Error bars on charts.
    - Color-coding or labels indicating confidence levels (e.g., green for high confidence data, yellow for estimated, red for highly uncertain).
5. **Links to Original Data Sources:**
    - Where feasible and appropriate (especially for public datasets or reports), provide links so users can explore the source data.

**D. Mechanisms for User Feedback and Iterative Improvement of the Tool:**

Continuous improvement is vital given the dynamic nature of AI and environmental data.

1. **Feedback Form/Button:**
    - Allow users to report discrepancies, suggest new models/data sources, or provide general usability feedback.
2. **"Report an Issue" for Specific Calculations:**
    - If a user believes a specific estimate is significantly off, allow them to flag it with details.
3. **Community Forum or Discussion Area (Optional):**
    - Could foster discussion, data sharing, and collaborative improvement if resources allow for moderation.
4. **Regularly Published Update Log:**
    - Inform users about new models added, data sources updated, and methodology refinements. This builds transparency and shows the tool is actively maintained.
5. **Collaboration with Researchers and Industry:**
    - Establish channels to receive updated data from academic researchers, AI providers, and initiatives like the AI Energy Score.

By thoughtfully designing the user interface and clearly communicating the underpinnings and limitations of the calculations, the tool can empower users to make more informed and environmentally conscious decisions about their AI usage.

# VIII. Addressing Challenges & Future-Proofing

Developing and maintaining an accurate and relevant AI Carbon/Water Footprint Calculator faces several inherent challenges. Proactive strategies are needed to address these issues and ensure the tool's long-term viability and credibility.

**A. Strategies for Dealing with Data Opacity, Lack of Standardization, and "Greenwashing":**

- **Data Opacity:** The lack of transparent, granular data from many AI providers, especially for proprietary models and the embodied footprint of hardware, is a primary challenge.
    - **Strategy:**
        1. **Prioritize Peer-Reviewed Research and Public Benchmarks:** Rely on verifiable data from academic studies (e.g., Jegham et al. , Morrison et al. , Strubell et al. , Patterson et al. ) and public benchmarks (e.g., AI Energy Score ) as the foundation.
        2. **Develop Robust Estimation Models:** For proprietary models or data gaps, use the "intelligent estimation" techniques outlined in Section V.C (model analogies, extrapolation, proxy data).
        3. **Clearly Label Data Sources and Confidence:** Mark estimates derived from non-direct or less certain sources with lower confidence levels and provide detailed explanations of the assumptions made.
        4. **Advocate for Transparency:** The tool itself, by highlighting data gaps, can contribute to the broader call for greater transparency from AI developers and hardware manufacturers.
- **Lack of Standardization:** Methodologies for measuring and reporting AI's environmental impact are not yet standardized across the industry or academia. This makes comparing data from different sources difficult.
    - **Strategy:**
        1. **Adopt Widely Recognized Protocols:** Where possible, align calculation methodologies with established protocols like the GHG Protocol for carbon and emerging frameworks for water footprinting (e.g., VWBA).
        2. **Be Explicit About Methodology:** Document the calculator's own methodology comprehensively and transparently (as discussed in Section VI.C), allowing users and experts to understand how estimates are derived.
        3. **Contribute to Standardization Efforts:** Engage with initiatives like the AI Energy Score or other relevant industry/academic groups working towards standardization.
- **"Greenwashing":** The risk of AI providers making sustainability claims that are not substantiated or that obscure the full impact (e.g., focusing on operational efficiency while ignoring embodied costs or absolute growth in consumption).
    - **Strategy:**
        1. **Critical Evaluation of Provider Data:** Do not take provider claims at face value. Cross-reference with independent research and benchmarks.
        2. **Focus on Measurable Metrics:** Base calculations on quantifiable data (kWh, Liters, gCO_2eq/kWh, PUE, WUE) rather than qualitative sustainability statements.
        3. **Educate Users:** Provide context on different types of sustainability claims (e.g., RECs vs. 24/7 CFE matching, operational vs. lifecycle impacts) to help

users critically assess information.

4. **Highlight Full Lifecycle:** Emphasize that operational footprint is only part of the story, and include embodied impacts where possible, even if estimated with higher uncertainty.

**B. Process for Updating the Tool with Data for New AI Models, Hardware, and Research Findings:**

The AI landscape evolves rapidly, so a static database will quickly become obsolete.

- **Strategy:**
    1. **Dedicated Research/Data Team/Process:** Allocate resources for ongoing monitoring of academic publications (e.g., arXiv, ACM Digital Library, IEEE Xplore, JMLR, NeurIPS, ICLR proceedings), industry reports, provider disclosures, and new benchmark releases (MLPerf, AI Energy Score).
    2. **Structured Database Design:** Design the backend database to be easily updatable with new models, hardware specs, efficiency figures, and emission factors. Include versioning for data points.
    3. **Regular Update Cycle:** Implement a defined schedule for reviewing new information and updating the calculator's database (e.g., quarterly or biannually, aligning with major conference publications or benchmark releases like the AI Energy Score's biannual updates ).
    4. **Automated Alerts/Scrapers (Potential):** Explore tools to automatically flag new relevant research papers or data releases.
    5. **Community Input Mechanism:** Utilize the user feedback mechanisms (Section VI.D) to gather leads on new data or models.
    6. **Partnerships/Collaborations:** Forge relationships with research institutions or initiatives that track AI sustainability data.

**C. Incorporating Regional Differences in Environmental Impact Beyond Just Grid Carbon Intensity:**

While grid carbon intensity is a primary factor, other regional differences are important.

- **Strategy:**
    1. **Water Stress Levels:**
        - Integrate regional water stress data (e.g., from WRI Aqueduct Water Risk Atlas or similar sources).
        - Develop a methodology to weight water consumption based on local scarcity. For example, a "Water Stress Index" could be applied as a multiplier to the calculated operational water footprint, resulting in a "Weighted Water Impact Score." This would better reflect the true environmental consequence of water use in water-scarce regions.
    2. **Data Center PUE/WUE Variations:**
        - Use region-specific PUE and WUE values for cloud providers where available (e.g., Google provides PUE by campus ; AWS provides PUE by region ).
        - If specific regional data for a provider is unavailable, use national averages or provider fleet averages with appropriate caveats.
    3. **Cooling Technology Prevalence:**
        - Research the dominant data center cooling technologies in different regions. Regions with colder climates might rely more on free air cooling, reducing water use for cooling, while hotter/arid regions might use more water-intensive evaporative cooling. This can inform default WUE_{site} assumptions.

4. **Renewable Energy Mix and Policies:**
    - Beyond average grid intensity, consider the actual renewable energy penetration and policies in a region, which can affect the likelihood of AI workloads being powered by cleaner energy, especially if 24/7 CFE matching is not employed by the provider.
5. **E-waste Regulations and Recycling Infrastructure:**
    - While harder to quantify per task, regional differences in e-waste management capabilities could be contextually noted if discussing the disposal phase of hardware.

By implementing these strategies, the AI Carbon/Water Footprint Calculator can strive to be a dynamic, transparent, and increasingly accurate tool, empowering users to understand and mitigate the environmental impact of their AI activities despite the challenges of a rapidly evolving and often opaque field.

# IX. Conclusions & Recommendations

This research has underscored the significant and growing environmental footprint of Artificial Intelligence, driven by the substantial energy and water demands of model training, inference, and the underlying hardware and data center infrastructure. While AI offers transformative potential, its current trajectory of resource consumption poses considerable challenges to global sustainability goals. The development of a user-friendly AI Carbon and Water Footprint Calculator, as envisioned by the user query, is a critical step towards empowering individuals and organizations to understand, measure, and ultimately mitigate these impacts.

**Key Conclusions from the Research:**

1. **Data Availability and Transparency are Major Hurdles:** A persistent theme is the opacity of data from commercial AI providers regarding the specific energy and water consumption of their models and the embodied footprint of their hardware. While cloud providers offer tools to track customer-side operational carbon, these often lack the granularity needed for precise AI service footprinting, and water footprinting tools are less mature. Open models and academic research provide valuable data points, but a comprehensive picture requires more industry transparency.
2. **Inference is a Dominant and Growing Concern:** While training large models has a massive upfront environmental cost, the cumulative impact of inference, performed billions of times daily, is emerging as the dominant long-term environmental burden for widely deployed models.
3. **Significant Variability in Footprints:** The environmental footprint of AI is not monolithic. It varies drastically based on:
    - **Model Choice:** Newer, optimized, or smaller models can be significantly more efficient than older or larger counterparts for specific tasks.
    - **Task Type:** Generative tasks (especially image and video) are orders of magnitude more resource-intensive than analytical or discriminative tasks.
    - **Hardware:** The efficiency of GPUs/TPUs used for computation plays a major role. Newer hardware generations offer substantial operational energy savings per computation, but their embodied footprint from manufacturing must also be considered due to rapid refresh cycles.
    - **Data Center Efficiency:** PUE and WUE values, along with cooling methods, significantly influence the overhead impact.

    - **Geographical Location:** Grid carbon intensity and local water stress levels are critical regional differentiators.
4. **Embodied Footprint is Significant but Underreported:** The energy and water consumed, and carbon emitted, during the manufacturing of AI hardware (GPUs, CPUs, servers) represent a substantial portion of the total lifecycle impact, yet this data is among the least transparent.
5. **Agentic AI Presents New Calculation Complexities:** Estimating the footprint of agentic platforms requires aggregating the impacts of multiple, often dynamic, LLM calls and tool invocations, plus the orchestrator's own processing overhead, introducing significant new layers of complexity and uncertainty.
6. **Existing Calculators Have Limitations:** Current AI footprint calculators vary widely in scope, usability, accuracy, and transparency, with no single tool comprehensively addressing all aspects of the user query.

**Recommendations for the AI Carbon/Water Footprint Calculator Development:**

1. **Prioritize a Phased Approach to Complexity and Scope:**
    - **Phase 1 (Core Functionality):** Focus on operational carbon and water footprint for inference of major, well-benchmarked LLMs and image generation models. Utilize the best available academic data (e.g., Jegham et al. , Luccioni et al. ) and provider-level PUE/grid intensity data.
    - **Phase 2 (Expanded Scope):** Incorporate embodied carbon/water for hardware (based on estimates like Morrison et al. ), add more model types, and refine methodologies for agentic platform estimation (potentially using simplified proxies initially). Introduce training footprint estimates for context.
    - **Phase 3 (Advanced Features):** Explore integration of real-time or more granular grid data, local water stress indices, and more sophisticated uncertainty modeling.
2. **Embrace "Intelligent Estimation" and Transparently Communicate Uncertainty:**
    - Given data gaps, the calculator *must* rely on estimations, model analogies, and proxy data.
    - For every output, provide a plausible range (min/max/average) rather than a single number.
    - Clearly disclose all assumptions, data sources (with dates), and the confidence level for each estimate. A dedicated, easily accessible "Methodology" section is essential.
3. **User-Centric Design:**
    - **Essential Inputs:** Keep initial user inputs minimal and intuitive (model, task, volume, general location). Offer advanced options for users with more detailed information (specific hardware, precise token counts).
    - **Actionable Outputs:** Provide results in clear units (kgCO_2e, Liters, kWh) with relatable equivalencies. Offer concrete, context-specific recommendations for footprint reduction.
    - **Educational Component:** Briefly explain *why* certain choices (e.g., image generation vs. text query) have different impacts.
4. **Robust and Agile Data Management:**
    - Establish a rigorous process for continuously monitoring new research, benchmarks (MLPerf, AI Energy Score), and provider disclosures.
    - Design the backend database for frequent and easy updates of model consumption figures, hardware specs, PUE/WUE values, and grid carbon intensities.
5. **Address Water Footprint Comprehensively:**

    - Go beyond simple operational water use. Incorporate both on-site cooling (WUE_{site}) and off-site electricity generation water (WUE_{source}).
    - Strive to integrate regional water stress indicators to provide a more accurate reflection of the environmental impact of water consumption.
6. **Develop Specific Methodologies for Agentic Platforms:**
    - This is a frontier area. Initial approaches might involve users estimating the number/type of sub-tasks an agent performs, or using high-level proxies based on task complexity and duration, acknowledging higher uncertainty.
7. **Incorporate Embodied Impacts Strategically:**
    - Use available academic estimates for embodied carbon and water in GPUs and servers.
    - Clearly explain that these are often amortized estimates and subject to significant uncertainty due to lack of manufacturer transparency.
8. **Carbon and Water Credit Information:**
    - Translate calculated footprints into potential offset volumes.
    - Provide information on reputable credit types, verification standards (Verra, Gold Standard for carbon; emerging VWB frameworks for water), and marketplaces, emphasizing due diligence.
9. **Engage with the Community and Advocate for Transparency:**
    - Implement user feedback mechanisms for continuous improvement.
    - The calculator itself can serve as an advocacy tool by highlighting data gaps and encouraging greater transparency from AI providers and hardware manufacturers.

By adopting these recommendations, the AI Carbon/Water Footprint Calculator can become a valuable resource for raising awareness, promoting responsible AI development and use, and contributing to a more sustainable AI ecosystem. The journey will be iterative, requiring ongoing research and adaptation as the field of AI and our understanding of its environmental impacts continue to evolve.

## Works cited

1. Explained: Generative AI's environmental impact | MIT News ..., https://news.mit.edu/2025/explained-generative-ai-environmental-impact-0117 2. The Environmental Impact of Artificial Intelligence - AI - Greenly, https://greenly.earth/en-gb/leaf-media/data-stories/the-environmental-impact-of-artificial-intelligence 3. AI's Environmental Impact: Making an Informed Choice - Marmelab, https://marmelab.com/blog/2025/03/19/ai-carbon-footprint.html 4. arxiv.org, https://arxiv.org/pdf/2503.05804 5. Paper page - Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models - Hugging Face, https://huggingface.co/papers/2304.03271 6. A Call for 'Green AI' in California: Evaluation of Existing and Alternative Regulations, https://framerusercontent.com/assets/GclY6SDou2V1Hy3Gn317N4Oqc.pdf 7. How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference, https://www.researchgate.net/publication/391741710_How_Hungry_is_AI_Benchmarking_Energy_Water_and_Carbon_Footprint_of_LLM_Inference 8. AI for Faculty: AI Footprints - WAB Learns - Western Academy of Beijing, https://learn.wab.edu/innovation/ai/faculty/ethics/aifootprint 9. Calculate Your AI Green Footprint in 60 Seconds - Optim.ai, https://optimaitech.org/servicios/impact/ai-environmental-offset/ai-green-impact-calculator/ 10.

AI Carbon Footprint Calculator | Deloitte UK,
https://www.deloitte.com/uk/en/services/consulting/content/ai-carbon-footprint-calculator.html 11.
Powering AI: The environmental footprint of AI today and tomorrow - YouTube,
https://www.youtube.com/watch?v=KHD3aNBT_Gw 12. AI Energy Score - A Standardized
Approach to Evaluating AI Model ...,
https://www.sustainableaicoalition.org/ai-energy-score-a-standardized-approach-to-evaluating-ai
-model-energy-efficiency/ 13. How Salesforce is Tracking AI's Impact with AI Energy Score ...,
https://energydigital.com/technology-and-ai/tracking-the-impact-of-ai-salesforces-ai-energy-scor
e 14. Announcing AI Energy Score Ratings - Hugging Face,
https://huggingface.co/blog/sasha/announcing-ai-energy-score 15. AI Energy Score: Initiative to
establish comparable energy efficiency ratings for AI models. - GitHub Pages,
https://huggingface.github.io/AIEnergyScore/ 16. AI Energy Score Leaderboard - a Hugging
Face Space by ..., https://huggingface.co/spaces/AIEnergyScore/Leaderboard 17.
AIEnergyScore/Leaderboard · How do these numbers compare to ...,
https://huggingface.co/spaces/AIEnergyScore/Leaderboard/discussions/3 18. AI's
Environmental Impact: Calculated and Explained - Arbor.eco,
https://www.arbor.eco/blog/ai-environmental-impact 19. Carbon Footprint | Google Cloud,
https://cloud.google.com/carbon-footprint 20. Sustainable AI: 3 tools to measure the
environmental impact of ML ...,
https://dev.to/audaciatechnology/sustainable-ai-3-tools-to-measure-the-environmental-impact-of-
ml-solutions-2hgk 21. Updated Carbon Methodology for the AWS Customer Carbon ...,
https://aws.amazon.com/blogs/aws-cloud-financial-management/updated-carbon-methodology-f
or-the-aws-customer-carbon-footprint-tool/ 22. Viewing your carbon footprint - AWS Billing,
https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/what-is-ccft.html 23. Azure
Carbon Optimization | Microsoft Learn,
https://learn.microsoft.com/en-us/azure/carbon-optimization/overview 24. View and analyze
emission data and insights - Azure Carbon Optimization | Microsoft Learn,
https://learn.microsoft.com/en-us/azure/carbon-optimization/view-emissions 25. Carbon
Footprint reporting methodology - Google Cloud,
https://cloud.google.com/carbon-footprint/docs/methodology 26. Customer Carbon Footprint
Tool - AWS,
https://aws.amazon.com/aws-cost-management/aws-customer-carbon-footprint-tool/ 27.
Sustainability Tools - Sustainable Cloud Computing - AWS,
https://aws.amazon.com/sustainability/tools/ 28. Sustainability | Google Cloud,
https://cloud.google.com/sustainability 29. Azure Sustainability—Sustainable Technologies |
Microsoft Azure, https://azure.microsoft.com/en-us/explore/global-infrastructure/sustainability
30. Reports - Amazon Sustainability, https://sustainability.aboutamazon.com/reports 31.
Microsoft Sustainability Solutions: Dynamics 365 & Azure.,
https://www.compusoftadvisors.com/microsoft-sustainability-solutions/ 32. AWS Customer
Carbon Footprint Tool Methodology Assurance - Amazon Sustainability,
https://sustainability.aboutamazon.com/aws-customer-carbon-footprint-tool-methodology-assura
nce.pdf 33. Sustainable AI: 3 tools to measure the environmental impact of ML solutions -
Audacia,
https://audacia.co.uk/technical-blog/tools-for-measuring-the-environmental-impact-of-ml 34.
(PDF) The HCI GenAI CO2ST Calculator: A Tool for Calculating the Carbon Footprint of
Generative AI Use in Human-Computer Interaction Research - ResearchGate,
https://www.researchgate.net/publication/390405126_The_HCI_GenAI_CO2ST_Calculator_A_T
ool_for_Calculating_the_Carbon_Footprint_of_Generative_AI_Use_in_Human-Computer_Intera

ction_Research 35. For Google's sustainability reporting, AI is a 'game changer' - Trellis Group, https://trellis.net/article/google-ai-sustainability-report-writing/ 36. Energy, Water & Data: Could Google Make AI Sustainable?, https://sustainabilitymag.com/articles/energy-water-data-could-google-make-ai-sustainable 37. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models, https://www.networkdee.org/library-v2/making-ai-less-thirsty-uncovering-and-addressing-the-secret-water-footprint-of-ai-models/RHJJKCL6 38. Power usage effectiveness – Google Data Centers, https://www.google.com/about/datacenters/efficiency/ 39. Sustainability - Google AI, https://ai.google/applied-ai/sustainability/ 40. AWS Cloud - Amazon Sustainability, https://sustainability.aboutamazon.com/products-services/aws-cloud 41. AWS AI powers new water projects in Spain - Amazon Europe, https://www.aboutamazon.eu/news/sustainability/aws-ai-powers-new-water-projects-in-spain 42. How AI is helping Amazon buildings conserve water and improve energy efficiency around the world, https://www.aboutamazon.com/news/sustainability/amazon-ai-buildings-water-energy-efficiency 43. AI for Sustainability: ESG Transformation with Microsoft Cloud, https://www.compunnel.com/drive-sustainability-progress-and-business-transformation-with-ai/ 44. Sustainability | Azure global infrastructure experience - Microsoft Datacenters, https://datacenters.microsoft.com/globe/powering-sustainable-transformation/ 45. Meta's 2024 Sustainability Report | Smart Energy Decisions, https://www.smartenergydecisions.com/research/metas-2024-sustainability-report/ 46. The Meta dilemma: Invest billions in AI but find ways to cut ..., https://trellis.net/article/the-meta-dilemma-invest-billions-in-ai-but-find-ways-to-cut-emissions-too/ 47. images.nvidia.com, https://images.nvidia.com/aem-dam/Solutions/documents/FY2024-NVIDIA-Corporate-Sustainability-Report.pdf 48. A No Brainer: How AI's Energy and Water Footprints Threaten Climate Progress, https://www.foodandwaterwatch.org/wp-content/uploads/2025/03/FSW_0325_AI_Water_Energy.pdf 49. Making AI Less “Thirsty”: Uncovering and Addressing the Secret Water Footprint of AI Models - arXiv, https://arxiv.org/html/2304.03271v5 50. AI's hidden thirst or how much water does Artificial Intelligence really drink? - Solveo, https://solveo.co/ais-hidden-thirst-or-how-much-water-does-artificial-intelligence-really-drink/ 51. Addressing Water Stress in the Age of AI-Driven Data Centers by Reem Kseibati - DSpace@MIT, https://dspace.mit.edu/bitstream/handle/1721.1/158889/Kseibati-reemzk-msred-cre-2025-Thesis.pdf?sequence=-1&isAllowed=y 52. (PDF) Generate Impressive Videos with Text Instructions: A Review ..., https://www.researchgate.net/publication/378310049_Generate_Impressive_Videos_with_Text_Instructions_A_Review_of_OpenAI_Sora_Stable_Diffusion_Lumiere_and_Comparable_Models 53. Energy-Efficient AI Development using Python: Methods and Case Studies - PhilArchive, https://philarchive.org/archive/DIYEAD 54. Manuscript BINGO!: A Novel Pruning Mechanism to Reduce the Size of Neural Networks - arXiv, https://arxiv.org/pdf/2505.09864 55. Increasing the Energy Efficiency of AI Models - Cohere, https://cohere.com/research/papers/efficient-ai.pdf 56. Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model - Journal of Machine Learning Research, https://www.jmlr.org/papers/volume24/23-0069/23-0069.pdf 57. Summary Report: White Paper on Global Artificial Intelligence ..., https://ai-sustainability.pubpub.org/pub/fz5j9nzu 58. 1 Introduction - arXiv, https://arxiv.org/html/2304.03271v4 59. Artificial Intelligence Index Report 2025,

https://cdn.ymaws.com/techcouncilofdelaware.org/resource/resmgr/research/hai_ai_index_report_2025.pdf 60. Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings - arXiv, https://arxiv.org/html/2501.08219v1 61. Power Usage and Energy Efficiency - llm-tracker, https://llm-tracker.info/_TOORG/Power-Usage-and-Energy-Efficiency 62. Full article: Computational Power and Subjective Quality of AI ..., https://www.tandfonline.com/doi/full/10.1080/10447318.2024.2422755 63. ChatGPT Energy Consumption Visualized - BEUK, https://www.businessenergyuk.com/knowledge-hub/chatgpt-energy-consumption-visualized/ 64. [2505.09598] How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference - arXiv, https://www.arxiv.org/abs/2505.09598 65. [OC] Will AI really drink us dry? Water use per 10 k tokens across popular LLMs (arXiv:2505.09598) : r/ChatGPT - Reddit, https://www.reddit.com/r/ChatGPT/comments/1kpjtif/oc_will_ai_really_drink_us_dry_water_use_per_10_k/ 66. How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference - Paper Detail - Deep Learning Monitor, https://deeplearn.org/arxiv/605068/how-hungry-is-ai?-benchmarking-energy,-water,-and-carbon-footprint-of-llm-inference 67. A Survey on Inference Engines for Large Language Models: Perspectives on Optimization and Efficiency - arXiv, https://www.arxiv.org/pdf/2505.01658 68. How Can OpenAI Reduce Water Usage in Server Cooling? - Use cases and examples, https://community.openai.com/t/how-can-openai-reduce-water-usage-in-server-cooling/1195434 69. Prompt engineering and its implications on the energy consumption of Large Language Models - arXiv, https://arxiv.org/html/2501.05899v1 70. [Literature Review] Prompt engineering and its implications on the energy consumption of Large Language Models - Moonlight, https://www.themoonlight.io/en/review/prompt-engineering-and-its-implications-on-the-energy-consumption-of-large-language-models 71. Under the radar? | Ada Lovelace Institute, https://www.adalovelaceinstitute.org/report/under-the-radar/ 72. Introducing Gemini: our largest and most capable AI model - Google Blog, https://blog.google/technology/ai/google-gemini-ai/ 73. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context - arXiv, http://arxiv.org/pdf/2403.05530 74. [2501.08219] Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings - arXiv, https://arxiv.org/abs/2501.08219 75. the rapid development of artificial intelligence models makes AI Faster and Cheaper, https://mpgone.com/the-rapid-development-of-artificial-intelligence-models-makes-ai-faster-and-cheaper/ 76. Mistral 7B - Stanford CRFM, https://crfm.stanford.edu/fmti/May-2024/company-reports/Mistral_Mistral%207B.html 77. ACLM: Developing a Compact Arabic Language Model - iajit, https://iajit.org/downloadfile/5211 78. Towards Sustainable Web Agents: A Plea for Transparency and Dedicated Metrics for Energy Consumption - arXiv, https://arxiv.org/html/2502.17903v1 79. towards sustainable web agents:aplea for transparency and dedicated metrics for energy consumption - arXiv, https://arxiv.org/pdf/2502.17903? 80. aclanthology.org, https://aclanthology.org/P19-1355.pdf 81. The Hidden Cost of AI: Unraveling the Power-Hungry Nature of Large Language Models - Preprints.org, https://www.preprints.org/frontend/manuscript/30dc8badac9e44da699113e5b5cd6737/download_pub 82. LLMCarbon - arXiv, https://arxiv.org/pdf/2309.14393 83. LLMCARBON: MODELING THE END-TO-END CAR- BON FOOTPRINT OF LARGE LANGUAGE MODELS, https://par.nsf.gov/servlets/purl/10508034 84. LLMCarbon: Modeling the End-To-End Carbon Footprint of Large Language ModelsThis work was supported in part by CCF-2105972, and NSF CAREER AWARD CNS-2143120. - arXiv, https://arxiv.org/html/2309.14393v2 85. UC

Riverside - eScholarship.org,
https://escholarship.org/content/qt79c880vf/qt79c880vf.pdf?t=rwo5l6 86. arXiv:2406.16893v1 [cs.CL] 15 May 2024, https://arxiv.org/pdf/2406.16893? 87. Investigating Energy Efficiency and Performance Trade-offs in LLM Inference Across Tasks and DVFS Settings | Request PDF - ResearchGate,
https://www.researchgate.net/publication/388029347_Investigating_Energy_Efficiency_and_Performance_Trade-offs_in_LLM_Inference_Across_Tasks_and_DVFS_Settings 88. Climate Implications of Diffusion-based Generative Visual AI Systems and their Mass Adoption - The House of Ethics,
https://www.houseofethics.lu/wp-content/uploads/2023/08/ICCC-2023_paper_60.pdf 89. Assessing the Carbon Footprint of OpenAI Models and Developing Strategies to Reduce It,
https://www.researchgate.net/publication/390241311_Assessing_the_Carbon_Footprint_of_OpenAI_Models_and_Developing_Strategies_to_Reduce_It 90. Generative artificial intelligence - Wikipedia, https://en.wikipedia.org/wiki/Generative_artificial_intelligence 91. From Google Gemini to OpenAI Q* (Q-Star): A Survey on Reshaping the Generative Artificial Intelligence (AI) Research Landscape - MDPI, https://www.mdpi.com/2227-7080/13/2/51 92. Midjourney AI: Analysis of Model - AmigoChat, https://amigochat.io/blog/midjourney-model-comparison 93. Towards Physical Plausibility in Neuroevolution Systems - Estudo Geral,
https://estudogeral.uc.pt/bitstream/10316/110787/1/MsCThesis_GabrielCortes.pdf 94. Hire the best Artificial Intelligence Engineers in Argentina - Upwork,
https://www.upwork.com/hire/artificial-intelligence-engineers/ar/ 95. Climate Scientists For Hire | Freelancer, https://www.freelancer.com/freelancers/skills/climate-sciences 96. Contemporary Issues in Industry 5.0 - OAPEN Library,
https://library.oapen.org/bitstream/handle/20.500.12657/98557/9783031747793.pdf?sequence=1&isAllowed=y 97. AI in Museums - Reflections, Perspectives and Applications - OAPEN Library,
https://library.oapen.org/bitstream/id/3ecbc4ec-2dac-4e05-881a-10414c20f7f2/9783839467107.pdf 98. Towards an Energy Consumption Index for Deep Learning Models: A Comparative Analysis of Architectures, GPUs, and Measurement Tools - PMC - PubMed Central,
https://pmc.ncbi.nlm.nih.gov/articles/PMC11820128/ 99. arXiv:2104.11757v2 [cs.CY] 3 May 2021, https://par.nsf.gov/servlets/purl/10257015 100. Opinion: AI's Impact on the Environment — the Good, Bad and the Ugly,
https://redlineproject.news/2025/03/20/opinion-ais-impact-on-the-environment-the-good-bad-and-the-ugly/ 101. A Survey of Sustainability in Large Language Models: Applications, Economics, and Challenges - arXiv, https://arxiv.org/pdf/2412.04782? 102. Analyzing the Energy Consumption of Generative Text-to-Audio Diffusion Models - arXiv,
https://arxiv.org/html/2505.07615v1 103. ServiceNow Yokohama Release: What's New and How to Prepare - Perspectium, https://www.perspectium.com/blog/servicenow-yokohama-release/ 104. LevelFields — Best AI Stocks to Watch in 2025: From Nvidia to Tempus AI,
https://www.levelfields.ai/news/best-ai-stocks-to-watch-in-2025 105. CXODX Magazine Feb 2025 by Leap Media Solutions - Issuu,
https://issuu.com/leapmediasolutions/docs/cxodx_magazine_feb_2025 106. Multimodal Large Language Models for Image, Text, and Speech Data Augmentation: A Survey - arXiv,
https://arxiv.org/html/2501.18648v2 107. Agentic AI in Energy Sector: Pioneering Autonomous Energy Intelligence - XenonStack, https://www.xenonstack.com/blog/agentic-ai-energy-sector 108. Agentic AI vs. Generative AI - IBM,
https://www.ibm.com/think/topics/agentic-ai-vs-generative-ai 109. CarbonCall: Sustainability-Aware Function Calling for Large Language Models on Edge Devices - arXiv,

https://arxiv.org/html/2504.20348v1 110. On the Limitations of Compute Thresholds as a Governance Strategy. - arXiv, https://arxiv.org/html/2407.05694v1 111. Energy Considerations of Large Language Model Inference and Efficiency Optimizations - OpenReview, https://openreview.net/pdf?id=aXNty1YGe0 112. Geography for AI sustainability and sustainability for GeoAI - Research Collection, https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/731159/GeographyforAIsustainabilityandsustainabilityforGeoAI.pdf?sequence=2&isAllowed=y 113. Gemini: A Family of Highly Capable Multimodal Models, https://assets.bwbx.io/documents/users/iqjWHBFdfxIU/r7G7RrtT6rnM/v0 114. From Google Gemini to OpenAI Q* (Q-Star): A Survey of Reshaping the Generative Artificial Intelligence (AI) Research Landscape - arXiv, https://arxiv.org/pdf/2312.10868 115. Geography for AI sustainability and sustainability for GeoAI - Taylor & Francis Online, https://www.tandfonline.com/doi/full/10.1080/15230406.2025.2479796?src= 116. Revisit the environmental impact of artificial intelligence: the overlooked carbon emission source? - Frontiers Journals of Higher Education Press, https://journal.hep.com.cn/fese/CN/10.1007/s11783-024-1918-y 117. Green Prompting - arXiv, https://arxiv.org/html/2503.10666v1 118. Optimizing Large Language Models: Metrics, Energy Efficiency, and Case Study Insights, https://arxiv.org/html/2504.06307v1 119. Optimizing Large Language Models: Metrics, Energy Efficiency, and Case Study Insights - arXiv, https://arxiv.org/pdf/2504.06307 120. (PDF) Artificial Intelligence and Electricity A System Dynamics ..., https://www.researchgate.net/publication/386582343_Artificial_Intelligence_and_Electricity_A_System_Dynamics_Approach/download 121. Aqueduct | World Resources Institute, https://www.wri.org/aqueduct 122. AI for a Greener Future: Its Power is in Our Hands | NVIDIA Technical Blog, https://developer.nvidia.com/blog/ai-for-a-greener-future-its-power-is-in-our-hands/ 123. The 6 Best Carbon Credit Trading Platforms: Maximize Your Impact - ClimateSort, https://climatesort.com/carbon-credit-trading-platforms/ 124. Blockchain Verified Carbon Credit Traceability → Scenario - Prism → Sustainability Directory, https://prism.sustainability-directory.com/scenario/blockchain-verified-carbon-credit-traceability/ 125. Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, and other SoTA Large Language Models - arXiv, https://arxiv.org/html/2502.18505v1 126. Boost ROI & Trust: How AI and Automation are Transforming Sustainable Growth - YouTube, https://www.youtube.com/watch?v=5GFaIjJD7ck