Environmental Footprint of AI Systems: Data for Calculator Tier Population

I. Executive Summary

This report presents a detailed investigation into the operational environmental footprint of a wide array of artificial intelligence (AI) systems, with a primary focus on energy consumption (Wh/unit) and operational water usage (mL/unit). The objective is to provide verifiable or robustly estimable data to populate and refine the tiers of an AI Footprint Calculator. The research encompasses video, audio, image, and text generation modalities, as well as the overhead associated with agentic AI platforms and the contextual impact of hardware and data center infrastructure.

Key quantitative findings indicate significant energy and water consumption variability across different AI tasks and models. For **video generation**, energy estimates range from approximately 2.3 Wh/second for user-estimated OpenAI Sora outputs to potentially 80-96 Wh/second for high-resolution, complex scenes generated by diffusion models like `zeroscope_v2_xl` on H100 GPUs.[1] Scaling factors, such as a quadratic increase in energy with resolution and linear increase with duration, are critical for accurate tiering.[3] For **audio generation**, diffusion models like AudioLDM show energy consumption around 0.073 Wh/minute under specific conditions on an NVIDIA A40 GPU, while less efficient models like Tango2 can be significantly higher.[4] Simple Text-to-Speech (TTS) is anticipated to be less energy-intensive. **Image generation** models like Stable Diffusion (1024x1024px) consume approximately 0.634 Wh/image [5], with values potentially reaching 2.9 Wh/image for other configurations.[6] **Text generation** exhibits a vast range, from less than 0.1 Wh per short query for highly efficient models like LLaMA-3.2 1B to over 33 Wh per long query for complex reasoning models like DeepSeek-R1.[7]

Operational water consumption is primarily derived from energy figures, factoring in data center Power Usage Effectiveness (PUE) and Water Usage Effectiveness (WUE) for both on-site cooling (WUEsite) and off-site electricity generation (WUEsource). Representative PUE values for major cloud providers (AWS, Google Cloud, Meta) hover around 1.08-1.14 [8], while WUEsite can be as low as 0.18-0.30 L/kWh_IT for efficient

facilities.[7] However, WUEsource, reflecting the water intensity of the electricity grid, can be substantially higher (e.g., 3.142 L/kWh_grid for the US average).[7]

Significant data gaps persist, particularly concerning the operational footprint of leading commercial video generation models like OpenAI Sora and Google Veo, advanced audio platforms such as Suno and ElevenLabs, and the precise overhead of agentic AI orchestrators. Hardware choices (e.g., NVIDIA H100 vs. A100, Google TPUs) demonstrably affect energy efficiency, with newer generations often providing better performance per watt.[11]

The findings underscore the necessity of detailed, context-aware data for meaningful AI footprint assessment. This report aims to provide a foundational dataset for the AI Footprint Calculator, while also highlighting areas requiring further research and advocating for greater transparency from AI developers and infrastructure providers.

II. Video Generation: Operational Energy and Water Footprint Analysis

A. Context and High Priority

Video generation represents a rapidly advancing frontier in AI, characterized by its potential for high computational demand and, consequently, a significant environmental footprint. Due to the complexity of synthesizing dynamic, coherent visual sequences, these models are presumed to be among the most resource-intensive AI applications. However, specific, verifiable data on their operational energy and water consumption remain scarce, particularly for leading commercial offerings. This section addresses this critical data gap, which is a high priority for the AI Footprint Calculator. The objective is to establish robust estimates for operational energy consumption, typically measured in Watt-hours (Wh) per second of generated video or per video clip of varying lengths, and to derive corresponding operational water consumption in milliliters (mL) per unit.

B. Models Under Investigation

The investigation targets prominent text-to-video platforms, including:

- OpenAI Sora
- Google Gemini Veo
- Pyxa
- 1minAI (focusing on its 30-second block capability)
- Amazon Nova video (specifically Amazon Nova Reel, with its 30-second block capability and longer generation options)

Given the limited direct disclosures from commercial entities, the analysis incorporates data from open-source models and academic research, such as studies on Open-Sora [3] and benchmarks from initiatives like the ML.ENERGY Leaderboard.[2] Other relevant platforms like Runway Gen-2 [15], Pika Labs [16], and Synthesia [17] are considered where comparable data or salient characteristics inform the broader understanding of video generation footprints.

C. Energy Consumption Data & Analysis

The energy consumption of video generation is highly variable, influenced by model architecture, output resolution, duration, frame rate, and scene complexity.

A significant source of benchmarked data is the **ML.ENERGY Leaderboard**.[2] For "Diffusion text-to-video generation," it provides energy per video in Joules for models such as `zeroscope_v2_576w` and `zeroscope_v2_xl`, tested on NVIDIA H100 and A100 GPUs. For instance:

- `model: zeroscope_v2_576w, denoising_params: 0.6B, gpu: H100-80GB-HBM3, energy_video_joules: 103333, frames: 24, resolution: 576x320` This translates to 103333 J/3600 J/Wh≈28.7 Wh for a 24-frame video (typically 1 second at 24 fps).
- `model: zeroscope_v2_xl, denoising_params: 1.7B, gpu: H100-80GB-HBM3, energy_video_joules: 346038, frames: 24, resolution: 1024x576` This is approximately 96.1 Wh for a 24-frame video. These figures provide concrete data points for specific open models and hardware configurations, forming a basis for calculator tiering.

The academic study **"Carbon in Motion: Characterizing Open-Sora"** [3], while not providing absolute baseline energy values for Open-Sora running on an NVIDIA A100 GPU, offers critical insights into scaling factors:

- The carbon footprint, and by direct relation energy consumption, is dominated by the iterative diffusion denoising process.
- Energy consumption scales almost quadratically with video resolution (e.g., generating at 720p produces approximately 10.1 times the carbon/energy of generating at 240p).
- Energy consumption scales linearly with the duration of the generated video.
- The study noted that generating an 8-second video at 480p resolution took approximately 8 minutes on an NVIDIA A100 GPU. This processing time, combined with the A100's TDP (up to 400W [11]), allows for an energy estimation

for that specific task. For example, $8 \text{ min} * (400\text{W} * \text{utilization_factor}) / 60 \text{ min/hr} = \text{Wh}$. Assuming a utilization factor (e.g., 0.7, as used in other estimates [19]), this would be 8/60 hr∗(400W∗0.7)=0.133 hr∗280W≈37.3 Wh for an 8s, 480p video, or ~4.67 Wh/s.

Synthesia, an AI video generation platform often used for avatar-based content, claims an average of 0.00025 kg CO2e per minute of AI video.[17] Converting this to energy requires an assumed grid carbon intensity. If using a global average electricity grid intensity (e.g., ~475 gCO2e/kWh in 2022, though this varies widely), then 0.25gCO2e/min would equate to (0.25 g/475000 g/MWh)∗106 Wh/MWh≈0.526 Wh/min, or ~0.0088 Wh/second. This figure is substantially lower than other estimates and likely reflects a highly optimized, less computationally complex type of video generation (e.g., rendering pre-defined avatar movements and lip-syncing).

An estimate for **CogVideoX**, reported by MIT Technology Review, suggests it uses 700 times the energy of generating one high-quality image (defined as a Stable Diffusion 3 Medium 1024x1024px image at 2,282 Joules or 0.634 Wh) for a 5-second video.[5] This implies an energy consumption of 700∗0.634 Wh≈443.7 Wh for a 5-second video, which is approximately 88.7 Wh/second. This represents a significantly higher energy intensity, potentially for more complex, fully generative scenes.

A user-based estimate for **OpenAI Sora** suggests 140 Wh per minute of video, assuming one NVIDIA H100 GPU (700W TDP [11]) running for one hour to generate five minutes of video.[1] This translates to approximately 2.33 Wh/second.

For **Amazon Nova Reel**, documentation states it generates videos up to two minutes in 6-second increments at 1280x720 resolution and 24 fps.[21] The generation is asynchronous, taking approximately 90 seconds for a 6-second video and 14-17 minutes for a 2-minute video.[22] While the underlying hardware is not explicitly detailed beyond AWS infrastructure [23], if we assume it runs on instances like AWS P4d (NVIDIA A100, up to 400W TDP per GPU [24]) or P5 (NVIDIA H100, up to 700W TDP per GPU [26]), energy can be estimated. For a 6s video taking 90s on an A100-equivalent: $(90/3600) \text{ hr} * (400\text{W} * \text{utilization_factor})$. With a 0.7 utilization factor: 0.025 hr∗280W=7 Wh for a 6s video, or ~1.17 Wh/s.

Data for **Pyxa** and **1minAI** were not found in the provided materials.

D. Impact of Generation Parameters

The operational footprint of video generation is highly sensitive to several parameters:

- **Resolution**: The Open-Sora study demonstrated an almost quadratic relationship between carbon emissions (and thus energy) and video resolution.[3] For example, increasing resolution from 240p to 720p (a 3x increase in each dimension, 9x pixels) resulted in a ~10.1x increase in carbon. This is because the latent space representation often scales with pixel dimensions, leading to a quadratic increase in tensor sizes processed by the diffusion model.
- **Scene Complexity/Quality Settings**: Higher perceived quality or more complex scenes often necessitate a greater number of diffusion denoising steps. The Open-Sora study showed that the denoising phase dominates the carbon footprint, especially with more steps (e.g., 97.3% of carbon for 40 steps in a 2s, 240p video).[3] Energy consumption is expected to be proportional to the number of denoising steps.
- **Clip Length (Duration)**: The total energy for generating a video scales linearly with its duration, as each additional frame or segment requires a proportional amount of computation.[3]

E. Hardware Context

The hardware used for inference significantly impacts energy consumption.

- The ML.ENERGY benchmarks utilize NVIDIA A100-SXM4-40GB and H100 80GB HBM3 GPUs.[2]
- The Open-Sora study specifically used an NVIDIA A100 GPU.[3]
- The user estimate for OpenAI Sora was based on an NVIDIA H100.[1]
- Google's video generation models, like Veo, are likely optimized for their proprietary Tensor Processing Units (TPUs). Google has reported a 3x improvement in carbon-efficiency from TPU v4 to Trillium (TPU v6), with Trillium being over 67% more energy-efficient than TPU v5e.[12]
- Amazon Nova Reel runs on AWS infrastructure, which includes various GPU instances such as those with NVIDIA A100s (P4d) and H100s/H200s (P5).[24]

The calculator must account for these hardware differences, as an H100 generally offers better energy efficiency for many AI workloads compared to an A100, and TPUs have their own distinct performance-per-watt characteristics.

F. Water Consumption

Specific operational water consumption figures per unit of video (e.g., mL/second or mL/clip) are not directly available in the reviewed literature for video generation. Water consumption estimates will be derived by:

1. Estimating the energy consumption per video unit (Wh).
2. Applying the PUE of the data center where the model is likely hosted (see Section VII). This gives total data center energy per video unit.
3. Applying relevant WUE values:
    - WUEsite (on-site cooling water per kWh of IT energy).
    - WUEsource (water consumed for off-site electricity generation per kWh of grid electricity). The methodology outlined in "How Hungry is AI?" [7] will be adapted, using provider-specific PUE and WUE values where available.

G. Comparative Footprint (Video vs. Image/Text)

Existing data strongly suggests that video generation is orders of magnitude more resource-intensive than image or text generation on a comparable unit basis.

- The CogVideoX estimate indicates that generating 5 seconds of video consumes 700 times the energy of generating one complex image (0.634 Wh).[5] This suggests roughly 443.7 Wh for 5s video vs. 0.634 Wh for an image.
- The Open-Sora study found that the carbon per frame of video is approximately 78 times the carbon per token of text generated by an LLM.[3]

These comparisons, while using different baselines, consistently point to the substantially higher environmental cost of video synthesis.

H. Deeper Insights & Causal Relationships

The available data reveals several important characteristics of video generation's environmental footprint. There is a notable lack of transparency from major commercial video model providers like OpenAI (Sora) and Google (Veo) regarding their operational energy and water consumption. Most quantitative data stems from studies on open-source models (e.g., Open-Sora, models on ML.ENERGY) or high-level

estimations. This opacity makes comprehensive assessment and comparison challenging.

The operational inference phase for video generation is exceptionally computationally demanding. For diffusion-based models, which are common, the iterative denoising process is the primary energy consumer.[3] The relationship between generation parameters and energy consumption is quite direct:

- Increasing output resolution quadratically increases the amount of data to be processed per frame, leading to a corresponding quadratic rise in energy demand.[3]
- Extending video duration linearly increases the total number of frames, thus scaling total energy linearly.[3]
- Achieving higher visual quality or depicting more complex scenes typically requires more denoising steps or more computationally intensive model components, directly increasing energy use per frame.[3]

The choice of hardware also plays a critical role. Newer, more powerful GPUs like the NVIDIA H100 or specialized ASICs like Google TPUs are generally more energy-efficient per operation than older hardware, assuming software is optimized. However, the absolute power draw of these high-performance chips can still be substantial.

The significant energy requirements for video generation have broad implications. If AI-driven video creation becomes widespread for entertainment, marketing, education, or personal use, the cumulative energy demand could place considerable strain on electricity grids and contribute significantly to carbon emissions, especially if the energy is not sourced from renewables. The associated water consumption, linked directly to energy use via data center cooling and indirectly via electricity generation processes, would also escalate. This situation underscores the urgent need for breakthroughs in energy-efficient video model architectures, continued hardware optimization, and a rapid transition to renewable energy sources for the data centers hosting these intensive workloads. The current data, although varied, consistently points towards video generation being at the higher end of the AI energy consumption spectrum. For example, the ML.ENERGY benchmark for `zeroscope_v2_xl` (1024x576, 24 frames) on an H100 GPU indicates approximately 96.1 Wh for a 1-second clip.[2] In contrast, a user estimate for OpenAI's Sora is around 2.33 Wh/second [1], and the CogVideoX estimate

reaches approximately 88.7 Wh/second.[5] This wide variance (from ~2 Wh/s to nearly 100 Wh/s for "high-quality" video) suggests that the AI Footprint Calculator must employ nuanced tiering for video generation. These tiers should consider not only resolution and length but also factors indicative of model efficiency or generative complexity (e.g., simple avatar animation versus complex cinematic scene generation). The extremely low energy figure from Synthesia (potentially <0.01 Wh/second [17]) likely represents a distinct category of highly optimized, less computationally intensive video tasks, such as avatar lip-syncing, rather than full scene generation from text.

I. Proposed Table for Section II

Table II.A: Estimated Operational Energy and Water Footprint of Text-to-Video Generation Models/Platforms

Model/Platform	Tier Category (Example)	Est. Wh/second	Est. Wh/10-sec clip (1080p, 24fps equiv.)	Derived mL/10-sec clip (Total: Direct+Indirect)	Resolution Output Context	Hardware Context (GPU/TPU)	Key Assumptions (e.g., Denoising Steps, Quality)	Data Source/Methodology	Confidence Level

Zeroscope_v2_xl	Short-Form HD (Open Source)	~96.1	~961	*Calculated*	1024x576 (24 frames)	H100 80GB HBM3	ML.ENERGY benchmark defaults	ML.ENERGY [2]	High
Zeroscope_v2_576w	Short-Form SD (Open Source)	~28.7	~287	*Calculated*	576x320 (24 frames)	H100 80GB HBM3	ML.ENERGY benchmark defaults	ML.ENERGY [2]	High
Open-Sora (derived)	Short-Form SD (Research)	~4.67	~46.7 (for 480p equiv.)	*Calculated*	480p	NVIDIA A100	8 min gen time for 8s video, 0.7 util. factor	"Carbon in Motion" [3] + Estimation	Medium
OpenAI Sora (estimate)	Cinematic HD (Comme	~2.33	~23.3	*Calculated*	1080p (assumed)	NVIDIA H100	User estimate based on H100 hours	User Estimate [1]	Low

	rcial Est.)						/vide o min		
CogVideoX (estimate)	Cinematic HD (Research Est.)	~88.7	~887	*Calculated*	High (not specified)	Not specified	700x image energy (0.634 Wh/image)	MIT Tech Review / Willison [5]	Low-Medium
Amazon Nova Reel	Short-Form HD (Commercial Est.)	~1.17	~11.7 (for 720p)	*Calculated*	1280x720 (24 fps)	AWS GPU (e.g., A100-equiv)	90s gen time for 6s video, 0.7 util. factor for 400 W TDP	AWS Docs [21] + Estimation	Low-Medium
Synthesia	Avatar Animation	~0.0088	~0.0888	*Calculated*	Not specified	Cloud-based	Based on 0.00025 kgCO2e/min, 475 gCO2e/kWh grid	Synthesia Report [17] + Grid Assumption	Medium (for task)

							intensity		
Generic Low-Res Tier	Video Tier 1 (e.g., ≤15s, 720p)	1.0 - 5.0	10-50	*Calculated*	~720p	Mixed (A100/H100/TPU)	Moderate complexity, standard parameters	Synthesis of Nova Reel, Open-Sora, Sora (low end)	Medium
Generic High-Res Tier	Video Tier 2 (e.g., >15s, ≥1080p)	20 - 100+	200-10000+	*Calculated*	≥1080p	H100 / Future Gen	High complexity, many steps	Synthesis of Zeroscope_xl, CogVideoX, Sora (high end)	Medium

Note: "Calculated" water usage will be populated using the methodology in Section VIII, applying relevant PUE/WUE values based on likely hosting infrastructure.

This table structure aims to consolidate diverse data points, clearly state assumptions, and provide a foundation for the calculator's video generation tiers. The "Tier Category" helps contextualize the wide range of energy values observed.

III. Audio Generation: Operational Energy and Water Footprint Analysis

A. Context and Priority

Audio generation encompasses a diverse set of AI applications, including Text-to-Speech (TTS), music synthesis (from simple loops to complex compositions with vocals), and advanced voice cloning. While perhaps not as overtly computationally intensive as high-fidelity video generation, these tasks still require significant processing, and data on their specific environmental footprint is necessary for

comprehensive calculator tiering. This section aims to find or estimate operational energy (Wh per minute of generated audio) and derive operational water consumption (mL per minute), differentiating by the complexity and type of audio generated.

B. Models Under Investigation

The research includes:

- **Suno**: A prominent platform for AI music generation.
- **Google's audio synthesis capabilities**: Including Google Cloud Text-to-Speech (which historically leveraged WaveNet technology [29]) and newer music generation models like Lyria.[30]
- **ElevenLabs**: Known for its advanced voice cloning and TTS services.
- **Open-source diffusion models for audio**: As benchmarked in the "Diffused Responsibility: Analyzing the Energy Consumption of Generative Text-to-Audio Diffusion Models" paper (arXiv:2505.07615), which includes AudioLDM, AudioLDM2, Make-an-Audio, Make-an-Audio-2, Stable Audio Open, Tango, and Tango2.[4]

C. Energy Consumption Data & Analysis

Direct energy consumption figures for commercial audio generation platforms are largely unavailable. However, academic research and platform-specific processing details offer pathways for estimation.

The study **"Diffused Responsibility" (arXiv:2505.07615)** [4] provides the most detailed energy benchmarks for open-source text-to-audio (TTA) diffusion models. Experiments were conducted on a single NVIDIA A40 GPU, with energy measurement via the CodeCarbon toolkit. Key findings include:

- Energy consumption (reported in kWh for processing entire test datasets like AudioCaps - 4895 clips, ~10s each; and Clotho - 1045 clips, 15-30s each) varies significantly by model and the number of inference steps.
- **AudioLDM** was consistently the most energy-efficient model. For the AudioCaps test set (~816 total minutes of audio), at 100 inference steps, its energy consumption can be estimated. Figure 1a in the paper shows AudioLDM using roughly 0.001 kWh for the AudioCaps set at 100 steps. This translates to: $(0.001 \text{ kWh} / 815.83 \text{ min}) * 1000 \text{ Wh/kWh} \approx 0.001226 \text{ Wh/min}$.
- **Tango2**, one of the more energy-intensive models, consumed approximately 0.0175 kWh for the AudioCaps set at 100 inference steps. This is: $(0.0175 \text{ kWh} / 815.83 \text{ min}) * 1000 \text{ Wh/kWh} \approx 0.02145 \text{ Wh/min}$.

- Energy consumption shows a linear relationship with the number of inference steps for all tested models. Increasing batch size reduces per-sample energy, most significantly between batch sizes 1 and 2.

For **Google WaveNet (Cloud Text-to-Speech)**, an older but influential model, it was reported to generate 1 second of speech in just 50 milliseconds when running on Google Cloud TPU infrastructure.[29] While direct power draw for this specific task on a TPU chip is not provided, the high speed suggests optimization for efficiency in TTS tasks. If a representative power draw for a TPU core handling such a task could be estimated (e.g., a fraction of a TPU v4/v5e chip's power), energy per second could be calculated. For instance, if a TPU core dedicated to this task drew 20W (a hypothetical value), then $20W * 0.05s = 1 Ws = 1 Joule = 1/3600 Wh/s$ of speech, or 0.0167 Wh/minute.

Google Lyria, for music generation, can produce 30-second instrumental clips in 10-20 seconds on Vertex AI.[30] Similar to WaveNet, energy estimation depends on assumed hardware power. Google generally highlights TPU carbon efficiency improvements for its AI workloads.[12]

For **OpenAI's TTS models**, like `gpt-4o-mini-tts`, pricing is $0.60 per million tokens, which OpenAI estimates as roughly 1.5 cents per minute of audio.[32] While this is a financial cost, it indirectly reflects computational resource use. If we assume a portion of this cost relates to energy (e.g., using the ~50% estimate from Soham (2024) cited in [33] for general LLM API calls, though this is a stretch for specific services), and know the cloud provider's electricity cost, a rough energy estimate might be attempted, but with very low confidence.

No specific operational energy or water consumption data for **Suno** or **ElevenLabs** inference was found in the provided materials.[34] Suno is noted for generating full 3 minute 33 second songs in under 2 minutes, indicating fast generation.[40]

D. Differentiation by Audio Type

The energy footprint of audio generation is expected to vary significantly based on the nature of the audio:

- **Simple Text-to-Speech (TTS)**: Likely the least energy-intensive, especially with optimized models like WaveNet. The primary computation involves converting text tokens to acoustic features and then to a waveform. Its energy footprint

might be analogous to short text generation tasks if the token-to-sound process is highly efficient.
- **Short, simple instrumental music loops/beds**: Diffusion models like AudioLDM, as benchmarked in "Diffused Responsibility" [4], are relevant here. Energy will depend on model size, number of inference steps (affecting quality and detail), and loop duration.
- **Complex, multi-track musical compositions with vocals (e.g., Suno-like outputs)**: These are expected to be more energy-intensive. This could involve more complex underlying models, longer generation sequences to maintain coherence, or even multiple specialized models working in concert (e.g., one for instrumentation, one for vocals, one for mixing). Data is particularly scarce here.
- **Advanced voice cloning (inference)**: Once a voice model is cloned (a training/fine-tuning process), the inference stage (generating speech using the cloned voice) is likely comparable in energy consumption to high-quality, expressive TTS. The complexity lies in accurately reproducing the nuances of the target voice.

E. Hardware Context

- The "Diffused Responsibility" study used an **NVIDIA A40 GPU**.[4]
- Google's audio synthesis services (WaveNet, Lyria) are stated to run on **Google Cloud TPU infrastructure**.[12] The specific generation of TPU or type of AWS/Azure instance for other commercial platforms is generally not disclosed.

F. Water Consumption

Direct water consumption figures (mL/minute) for audio generation are not available. These will be derived by:

1. Estimating energy consumption (Wh/minute) for different audio types and models.
2. Applying appropriate PUE and WUE (site and source) values based on likely hosting infrastructure.[7]

G. Deeper Insights & Causal Relationships

The available data, though limited for commercial platforms, indicates that diffusion-based audio generation models exhibit energy consumption patterns similar to those in image and video generation: energy scales with model complexity (parameters, architecture) and the intensity of the generation process (e.g., number of inference

steps).[4] Specialized TTS models, such as Google's WaveNet, appear to be highly optimized for speed, suggesting a potentially lower per-unit energy consumption for basic speech synthesis tasks.[29] A significant lack of transparency regarding operational energy and water data persists for major commercial audio generation platforms like Suno and ElevenLabs.

The causal chain for energy consumption in audio generation involves several factors. Increased audio complexity—such as moving from single-speaker TTS to multi-instrumental music with vocals—necessitates more sophisticated models or longer, more intricate generation sequences, thereby directly increasing computational load and energy use. Similarly, the desired length of the audio output linearly affects the total computations and thus total energy. For diffusion models, a higher number of inference steps, often correlated with improved audio quality or detail, directly translates to more processing and higher energy consumption.[4] The efficiency of the underlying hardware (e.g., specialized TPUs for Google's services versus general-purpose GPUs like the A40) will also critically influence the energy per minute of audio.

The proliferation of easily accessible AI audio generation tools for music, voiceovers, and other applications could lead to a substantial new category of AI-driven energy and water consumption. The environmental impact will be amplified if users frequently generate long-form content (e.g., full songs, podcast episodes, audiobooks) or iterate extensively to achieve desired results. The calculator must, therefore, differentiate between low-energy TTS and potentially much higher-energy complex music generation, even if precise figures for all commercial models remain elusive. Providing ranges based on open-source benchmarks (like those from "Diffused Responsibility" [4]) and clearly stating the assumptions regarding audio complexity will be crucial for user guidance. For instance, the ~0.0012 Wh/min for AudioLDM (simple instrumental) versus ~0.021 Wh/min for Tango2 (also instrumental, but less efficient architecture or different training) already shows a ~17.5x difference within the same broad category benchmarked on the same hardware.[4] More complex tasks like full song generation with vocals are likely to be even higher.

H. Proposed Table for Section III

Table III.A: Estimated Operational Energy and Water Footprint of Audio Generation Models/Tasks

Model/Platform/Task Type	Est. Wh/minute	Derived mL/minute (Total: Direct+Indirect)	Hardware Context (GPU/TPU)	Basis for Estimate	Key Assumptions (Complexity, Length, Parameters)	Confidence Level
Google Cloud TTS (WaveNet-like Estimate)	~0.017	*Calculated*	Google Cloud TPU	Processing speed (1s audio/50ms) + hypothetical 20W TPU core power draw [29]	Basic TTS, efficient model	Low
AudioLDM (Simple Loop/Short Audio)	~0.0012	*Calculated*	NVIDIA A40	"Diffused Responsibility" [4] (0.001 kWh / ~816 min audio @ 100 steps)	10s clips, 100 inference steps	Medium
Tango2 (Simple Loop/Short Audio)	~0.0215	*Calculated*	NVIDIA A40	"Diffused Responsibility" [4] (0.0175 kWh / ~816 min audio @ 100 steps)	10s clips, 100 inference steps	Medium

Suno Music (Complex Composition Est.)	0.1 - 1.0+	*Calculated*	Cloud GPU (unspecified)	Extrapolation: Higher than Tango2, reflecting complexity/length; Wide range due to lack of data	Full song (e.g., 3 min), vocals, multiple instruments	Very Low
ElevenLabs Voice Cloning (Inference Est.)	0.02 - 0.2	*Calculated*	Cloud GPU (unspecified)	Analogy to high-quality TTS / expressive audio generation; Wide range	High-fidelity speech, per minute output	Very Low
Tier: Basic TTS	0.01 - 0.05	*Calculated*	Mixed (TPU/GPU)	Synthesis of WaveNet estimate & efficient text models	Clear, short speech segments	Low-Medium
Tier: Simple Instrumental	0.001 - 0.05	*Calculated*	GPU (e.g., A40)	Based on AudioLDM/Tango2 range from $4^{}$ for short clips, 100-200 steps	Short loops, moderate complexity	Medium
Tier: Complex Music/Vocals	0.05 - 1.0+	*Calculated*	GPU (High-end)	Extrapolation based on increased complexity over simple instrumental	Full songs, vocals, rich	Low

				s; high uncertainty	instrumentation	

Note: "Calculated" water usage will be populated using the methodology in Section VIII, applying relevant PUE/WUE values. Ranges for commercial platforms are highly speculative due to lack of data.

IV. Image Generation: Refining Tier Data and Specific Model Footprints

A. Context

The generation of static images from text prompts is a relatively mature application of generative AI, with more available benchmarks compared to video or audio. However, refinement of data is necessary for specific new models and to better differentiate the calculator's tiers, such as "Standard Image Generation" versus "High-Detail/Complex Image Generation." This section aims to consolidate energy (Wh/image) and derive water (mL/image) data, focusing on the impact of resolution, diffusion steps, and model features.

B. Models Under Investigation

The investigation includes:

- Image generation capabilities of Google Gemini.
- Image generation by Anthropic Claude models (if publicly confirmed and data is available).
- Alibaba Qwen series (for image generation).
- Amazon Titan Image Generator.
- Widely benchmarked open-source models like Stable Diffusion (various versions).

C. Energy Consumption Data & Analysis

The **ML.ENERGY Leaderboard** is a crucial source for "Diffusion text-to-image generation" benchmarks.[2] It provides:

- **Model Name**: e.g., `runwayml/stable-diffusion-v1-5`, `stabilityai/stable-diffusion-xl-base-1.0`.
- **Denoising Parameters (Billions)**.
- **GPU Model**: NVIDIA A100-SXM4-40GB, H100 80GB HBM3.
- **Energy per image (Joules)**.
- **Resolution**.

Examples from ML.ENERGY [2]:

- `runwayml/stable-diffusion-v1-5` (0.9B params) on H100, 512x512: 30,271 J/image (≈0.0084 Wh/image).
- `stabilityai/stable-diffusion-xl-base-1.0` (2.6B params) on H100, 1024x1024: 104,858 J/image (≈0.0291 Wh/image).

Simon Willison, summarizing an MIT Technology Review article, reported that **Stable Diffusion 3 Medium** (1024x1024 pixels) consumes 2,282 Joules per image.[5] This equates to 2282/3600≈0.634 Wh/image. This figure is notably higher than the ML.ENERGY figures for SDXL 1.0 at the same resolution, suggesting differences in model versions, specific parameters (like diffusion steps), or measurement methodology.

A report by Greenly.earth cites image generation at 2.9 kWh per 1,000 images, which is 2.9 Wh/image.[6] The same source claims image generation is 60 times more energy-intensive than text generation. The specific model or conditions for this 2.9 Wh/image figure are not detailed in the snippet but would be valuable context.

A Reddit user described their consumer hardware (potentially a GPU like an NVIDIA 3060/4070 [47]) drawing ~200W for ~7.5 seconds per image. This calculates to (200W∗7.5s)/3600 s/hr≈0.417 Wh/image.[47] This offers a data point for non-datacenter hardware, highlighting that energy consumption can be significant even on consumer-grade equipment.

Epoch AI qualitatively notes that image generation requires "substantially more computational power than text processing" [48], aligning with the Greenly.earth comparison.

Specific operational energy or water data for image generation via **Google Gemini's native capabilities, Anthropic Claude's potential image features, Alibaba Qwen (image models like Qwen-VL [49]), or Amazon Titan Image Generator** (part of Amazon Bedrock, with custom model training costing $0.005 per image [50]) were not found in the provided research snippets in terms of Wh/image for inference. Amazon Nova's image generation capabilities are mentioned [51], but without energy figures.

D. Differentiating Tiers: Standard vs. High-Detail/Complex

The AI Footprint Calculator aims to differentiate between "Standard Image Generation" (Tier 1) and "High-Detail/Complex Image Generation" (Tier 2). Several factors influence this:

- **Resolution**: The ML.ENERGY data includes resolution, and it is evident that higher resolutions (e.g., 1024x1024 vs. 512x512) require more energy. For instance, `stable-diffusion-xl-base-1.0` at 1024x1024 used ~3.46 times the energy of `stable-diffusion-v1-5` at 512x512 on the same H100 GPU, though model parameter count also differs.[2] An increase from 1MP to 4MP to 16MP would substantially increase pixel count and thus computational load. The exact scaling relationship (e.g., linear or quadratic with pixel count for diffusion models) needs further empirical data but a significant increase is expected.
- **Diffusion Steps**: A higher number of diffusion steps generally leads to better image quality and detail but involves more iterations of the denoising neural network, thus increasing computational work and energy per image. This is a common configurable parameter in many image generation tools.
- **Model Features and Complexity**: Advanced features like in-painting, out-painting, detailed upscaling, or generating images with highly intricate details or specific artistic styles often rely on larger models or more intensive computational processes, contributing to higher energy use. The parameter count of the denoising model itself (e.g., 2.6B for SDXL vs. 0.9B for SDv1.5 in ML.ENERGY data) is a strong indicator of baseline energy.

Based on this, Tier 1 (Standard) could represent images around 0.25-1 megapixel (e.g., 512x512 to 1024x1024), using a moderate number of diffusion steps (e.g., 20-50) with models like Stable Diffusion v1.5. Tier 2 (High-Detail/Complex) could encompass larger resolutions (>1-4MP), high diffusion step counts (>50-100), or the use of larger models like SDXL or features requiring additional processing.

E. Water Usage for Image Generation

No direct mL/image figures were found. Water consumption will be derived from the energy per image (Wh/image) data using appropriate PUE and WUE (site and source) values for the likely hosting infrastructure.[7] To refine estimates beyond a single data center example, a range of WUEs from different providers or geographical regions should be considered, or assumptions clearly stated if an average is used. The variability in data center water efficiency (e.g., WUEsite from 0.18 L/kWh for AWS to

1.20 L/kWh for some Chinese data centers [7]) will directly impact derived water footprints.

F. Deeper Insights & Causal Relationships

The energy consumed per image is fundamentally tied to the model architecture (particularly the size and complexity of the denoising U-Net in diffusion models) and the specific generation parameters chosen by the user (resolution, number of diffusion steps). Open-source models like Stable Diffusion have more publicly available energy benchmarks (e.g., via ML.ENERGY [2]) compared to proprietary image generation services offered by major cloud providers, where such data is often not disclosed.

The causal links are clear:

- Increasing image resolution (e.g., from 512x512 to 1024x1024) directly multiplies the number of pixels the model must generate. For diffusion models, this typically increases the size of the latent space representation and the computational load on the U-Net at each step, leading to higher energy per image.
- Increasing the number of diffusion steps means the core denoising network performs more iterative refinements. This linearly increases the total computation and, consequently, the energy consumed for a given image resolution and model.

The widespread availability and ease of use of AI image generation tools could lead to substantial cumulative energy consumption, even if the per-image cost for standard images is relatively modest (e.g., in the range of 0.01 Wh to a few Wh). If millions of users generate numerous images daily, the aggregate energy demand can become significant. Furthermore, the trend towards higher resolutions and more detailed outputs for professional or artistic applications will inherently push the per-image energy cost upwards. The use of credit systems by some commercial services can obscure the actual energy footprint from the end-user, making informed sustainable choices more difficult. The calculator's tiering, informed by data such as that from ML.ENERGY, is therefore essential for reflecting this variability and promoting awareness.

G. Proposed Table for Section IV

Table IV.A: Estimated Operational Energy and Water Footprint of Image Generation Models/Tiers

Model/Platform/Tier	Est. Wh/image	Derived mL/image (Total: Direct+Indirect)	Resolution (MP)	Key Parameters (e.g., Diffusion Steps)	Hardware Context (GPU)	Data Source/Methodology	Confidence Level
Stable Diffusion v1.5 (ML.ENERGY)	~0.0084	*Calculated*	0.26 (512x512)	Benchmark default	H100 80GB HBM3	ML.ENERGY 2	High
Stable Diffusion XL 1.0 (ML.ENERGY)	~0.0291	*Calculated*	1.05 (1024x1024)	Benchmark default	H100 80GB HBM3	ML.ENERGY 2	High
Stable Diffusion 3 Medium (MIT Tech Rev)	~0.634	*Calculated*	1.05 (1024x1024)	Not specified	Not specified	Willison/MIT Tech Review 5	Medium
General Image Gen (Greenly.earth)	~2.9	*Calculated*	Not specified	Not specified	Not specified	Greenly.earth 6	Low
Consumer GPU Estimate (Reddit)	~0.417	*Calculated*	Not specified	User settings	Consumer GPU	User Report 47	Low

					(~200 W)		
Tier 1: Standard Image Gen	0.01 - 0.5	*Calculated*	~0.25 - 1 MP	~20-50 steps	A100/ H100 or equiv.	Synthesis of ML.ENERGY (smaller models), Consumer GPU data	Medium
Tier 2: High-Detail/ Complex Image Gen	0.5 - 3.0+	*Calculated*	>1 - 4+ MP	>50-100 steps, advanced features	H100 or equiv.	Synthesis of ML.ENERGY (larger models), SD3M, Greenly	Medium-Low
Google Gemini Image Gen (API Estimate)	*Est. Req.*	*Calculated*	User-defined	API defaults	Google TPU	Estimation needed (e.g., via API cost, comparison)	Low
Amazon Titan Image Gen (API Estimate)	*Est. Req.*	*Calculated*	User-defined	API defaults	AWS GPU	Estimation needed	Low

Note: "Calculated" water usage will be populated using the methodology in Section VIII. "Est. Req." indicates that an estimation methodology needs to be developed for these commercial models based on available proxies if direct data is not found.

V. Text Generation: Enhancing Data for Advanced Models and Tier 3

A. Context

While basic text generation might have a relatively lower footprint per unit compared to image or video, the sheer volume of text AI interactions and the rise of highly capable, computationally intensive Large Language Models (LLMs) necessitate robust data, especially for "Tier 3: Advanced Text Generation." This tier should represent models like Anthropic Claude 3 Opus, specific intensive tasks using Google Gemini Advanced, and large versions of Alibaba Qwen and Amazon Titan text models. The focus is on precise energy (Wh/1k tokens or Wh/complex query) and water (mL/unit) data.

B. Models Under Investigation

- Anthropic Claude 3 Opus
- Google Gemini Advanced (for computationally intensive tasks)
- Alibaba Qwen (large versions, e.g., Qwen2-72B, Qwen3-235B-MoE)
- Amazon Titan text models (large versions, e.g., Titan Text Premier)
- Other models benchmarked in relevant studies (e.g., GPT-4 variants, LLaMA variants, DeepSeek models).

C. Energy Consumption Data & Analysis

The paper **"How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference" (arXiv:2505.09598)** [7] is a foundational resource. It provides energy consumption (Wh per query) for ~30 LLMs, differentiating by prompt length (short: 100 input/300 output tokens; medium: 1k/1k; long: 10k/1.5k tokens). Key relevant findings include:

- **GPT-4o (Mar '25)**: 0.421 Wh (short), 1.214 Wh (medium), 1.788 Wh (long).
- **Claude-3.7 Sonnet**: 0.836 Wh (short), 2.781 Wh (medium), 5.518 Wh (long).
- **Claude-3.7 Sonnet ET (Extended Thinking)**: 3.490 Wh (short), 5.683 Wh (medium), 17.045 Wh (long).
- **DeepSeek-R1 (Reasoning Model)**: 23.815 Wh (short), 29.000 Wh (medium), 33.634 Wh (long).
- **o3 (OpenAI Reasoning Model)**: 7.026 Wh (short), 21.414 Wh (medium), 39.223 Wh (long).
- **GPT-4.1 nano (Most Efficient in Study)**: 0.103 Wh (short), 0.271 Wh (medium), 0.454 Wh (long).
- **LLaMA-3.2 1B**: 0.070 Wh (short), 0.218 Wh (medium), 0.342 Wh (long). This study did not explicitly list Claude 3 Opus, specific Gemini Advanced models, Alibaba Qwen, or Amazon Titan models in its table of 30.

Epoch AI's analysis of ChatGPT [19] estimates:

- Typical GPT-4o query: ~0.3 Wh.

- Long input (10k tokens): ~2.5 Wh.
- Very long input (100k tokens): ~40 Wh.
- OpenAI's `o1` and `o3` reasoning models are suggested to consume "substantially more energy."

Regarding specific target models for Tier 3:

- **Anthropic Claude 3 Opus**: Described as Anthropic's most capable model, excelling in reasoning, math, and coding.[56] It is priced higher than Sonnet ($15/M input, $75/M output tokens for Opus vs. $3/$15 for Sonnet [58]). The "How Hungry is AI?" paper benchmarks Claude-3.7 Sonnet and Sonnet ET. Given Opus's positioning and pricing, its energy consumption is likely higher than Sonnet ET, potentially in the range of 10-25 Wh for complex queries, but this is an estimation.
- **Google Gemini Advanced**: This typically implies use of models like Gemini 1.5 Pro or Ultra. Gemini 1.5 Pro is described as a compute-efficient MoE model.[59] While direct Wh/query figures are absent, research indicates AI-augmented search (powered by such models) could increase energy demand 60-70x over conventional search (0.04 Wh) [60], suggesting a range of ~2.4-2.8 Wh per advanced search query. The "How Hungry is AI?" paper does not include Gemini models.
- **Alibaba Qwen (large versions)**: The Qwen3 series includes large dense models (up to 32B) and MoE models (e.g., Qwen3-235B-A22B activating 22B parameters).[61] MoE architectures claim significant energy savings (e.g., Qwen3-235B MoE -90% energy vs. dense model of same performance; Qwen-Max MoE 4x FLOP reduction vs. 175B dense).[62] Qwen3-32B is priced at $0.0003/1k tokens.[62] Without direct benchmarks, estimating energy for large Qwen models is challenging but their MoE versions are designed for efficiency.
- **Amazon Titan text models (large versions)**: Titan Text Premier is a large model.[64] Pricing for Titan Text Express is $0.0008/1k input and $0.0016/1k output tokens.[66] The "How Hungry is AI?" paper does not include Titan models. Energy estimates would need to be based on comparisons to AWS-hosted models with known energy footprints [7] or relative pricing, with low confidence.

The energy per 1k tokens can be derived from Wh/query figures if total tokens (input + output) for that query type are known. For example, for GPT-4o short prompt (400 total tokens) at 0.421 Wh, this is 0.421 Wh/0.4k tokens≈1.05 Wh/1k tokens. For DeepSeek-R1 long prompt (11.5k total tokens) at 33.634 Wh, this is 33.634 Wh/11.5k tokens≈2.92 Wh/1k tokens.

D. Specific Water Footprint for LLMs

The most robust methodology for LLM water footprint is from **"How Hungry is AI?"** [7]:

- Formula: `Water (L) = (E_query / PUE) * WUE_site + E_query * WUE_source`
- **PUE values used**: OpenAI/Azure: 1.12; Anthropic/AWS: 1.14; DeepSeek/China: 1.27.
- **WUEsite (on-site cooling, L/kWh_IT) values used**: OpenAI/Azure: 0.30; Anthropic/AWS: 0.18; DeepSeek/China: 1.20.
- **WUEsource (off-site electricity generation, L/kWh_grid) values used**: US average (for Azure/AWS): 3.142; China average (for DeepSeek): 6.016.
- **Resulting water per query**:
    - GPT-4.1 nano: < 2 mL/query (across prompt sizes).
    - DeepSeek-R1: > 150 mL/query (consistently).
    - The paper's case study for GPT-4o (0.43 Wh/short query) at 700 million queries/day projects an annual water consumption equivalent to the drinking needs of 1.2 million people, highlighting the impact of scale.

Other less precise estimates include 500 mL per ChatGPT query or per 5-50 prompts [35], 16 oz (~473 mL) per prompt string [68], or 519 mL per 100-word email by ChatGPT-4.[6] A refined estimate suggests 500 mL per 300 queries if considering only direct on-site cooling water.[70] The [7] methodology, accounting for both direct and indirect water based on specific infrastructure parameters, is preferred for its rigor.

E. Deeper Insights & Causal Relationships

The energy consumption per query for LLMs demonstrates extreme variability, spanning more than two orders of magnitude.[7] This variance is primarily driven by:

- **Model Architecture and Size**: Larger models and those designed for complex reasoning (like OpenAI's o-series or DeepSeek-R1) are significantly more energy-intensive per query.

- **Mixture-of-Experts (MoE) Architecture**: Models like Qwen's MoE variants claim substantial energy reductions compared to dense models of similar capability by activating only a fraction of their parameters per token.[62] This is a critical architectural trend for efficiency.
- **Nature of the Query**:
    - **Input/Output Length**: Longer prompts and completions directly increase the number of tokens processed, leading to higher energy per query.[7]
    - **Task Complexity**: Queries requiring multi-step reasoning, extensive knowledge retrieval, or chain-of-thought processes inherently demand more computation than simple lookups or short text continuations. This is evident in the higher energy use of "reasoning models".[7]

The water footprint of LLM inference is inextricably linked to energy consumption and the specifics of the data center infrastructure (PUE, cooling methods determining WUEsite) and the regional electricity grid (determining WUEsource). Even if a model is energy-efficient per query, high-volume deployment in a region with a water-intensive electricity grid or inefficient data center cooling can result in a substantial water footprint.

The implications of these findings are profound. As LLMs become more powerful and capable of tackling increasingly complex tasks (often termed "advanced" or "reasoning" tasks), there's a risk of a disproportionate surge in energy consumption per query if efficiency improvements do not keep pace with capability enhancements. The calculator's Tier 3 for "Advanced Text Generation" must capture this wide energy spectrum. Users choosing models for demanding applications need to be aware that selecting a more "intelligent" or "reasoning-focused" model could entail a significantly larger environmental footprint per interaction. The emergence of MoE architectures offers a promising path to mitigate this, potentially allowing for high capability with managed energy costs. Highlighting models that achieve a favorable balance of performance and energy efficiency [7] is essential.

F. Proposed Table for Section V

Table V.A: Estimated Operational Energy and Water Footprint of Advanced Text Generation Models (Tier 3 Focus)

Model	Provider	Est. Wh/1k tokens (derived)	Est. Wh/complex query (e.g., long prompt)	Derived mL/complex query (Total: Direct+Indirect)	Hardware Context (GPU/TPU, Cloud)	Basis for Estimate & Key Assumptions (Prompt Length, Task Type)	Confidence Level
GPT-4o (Mar '25)	OpenAI	~1.05 (short); ~0.61 (medium); ~0.16 (long)	0.421 (S); 1.214 (M); 1.788 (L)	~1.8 (S); ~5.2 (M); ~7.7 (L)	DGX H200/H100, Azure	7: Short (0.4k tok), Med (2k), Long (11.5k). Water: PUE 1.12, WUEs 0.30/3.142 L/kWh.	High
Claude-3.7 Sonnet	Anthropic	~2.09 (S); ~1.39 (M); ~0.48 (L)	0.836 (S); 2.781 (M); 5.518 (L)	~3.4 (S); ~11.4 (M); ~22.6 (L)	DGX H200/H100, AWS	7: Prompt lengths as above. Water: PUE 1.14, WUEs 0.18/3.142 L/kWh.	High
Claude-3.7 Sonnet ET	Anthropic	~8.73 (S); ~2.84 (M); ~1.48 (L)	3.490 (S); 5.683 (M); 17.045 (L)	~14.3 (S); ~23.3 (M); ~70.0 (L)	DGX H200/H100, AWS	7: As above. "Extended Thinking" implies more	High

						computation.	
DeepSeek-R1	Deepseek	~59.5 (S); ~14.5 (M); ~2.92 (L)	23.815 (S); 29.000 (M); 33.634 (L)	~206.8 (S); ~251.8 (M); ~292.0 (L)	DGX H800, Deepseek DC	[7]: As above. Reasoning model. Water: PUE 1.27, WUEs 1.20/6.016 L/kWh.	High
OpenAI o3	OpenAI	~17.6 (S); ~10.7 (M); ~3.41 (L)	7.026 (S); 21.414 (M); 39.223 (L)	~30.2 (S); ~92.1 (M); ~168.7 (L)	DGX H200/H100, Azure	[7]: As above. Reasoning model.	High
Claude 3 Opus (Est.)	Anthropic	*2.0 - 6.0*	*10 - 25*	*Calculated (40-100 mL)*	DGX H200/H100, AWS	Estimate: Higher than Sonnet ET due to capability/price. For complex reasoning, long context.	Low-Medium

Gemini Advanced (Est.)	Google	*0.5 - 2.5*	*2.0 - 10.0*	*Calculated (8-40 mL)*	Google TPU	Estimate: Based on 60-70x conventional search (0.04Wh) or comparison to similar capable models.	Low
Qwen3-235B-A22B MoE (Est.)	Alibaba	*0.1 - 0.5*	*0.5 - 2.0* (for equiv. 22B dense task)	*Calculated (China DC)*	Alibaba Cloud GPU	Estimate: Based on 90% reduction vs. comparable dense model energy, high parameter but MoE. 62	Low
Titan Text Premier (Est.)	Amazon	*1.0 - 4.0*	*5 - 20*	*Calculated*	AWS GPU	Estimate: Based on pricing relative to Claude Sonnet on AWS, for large context/complex tasks.	Low

Note: (S)=Short, (M)=Medium, (L)=Long prompts as per.[7] Wh/1k total tokens derived. Water calculated using [7] methodology with provider-specific PUE/WUE. Estimates for Opus, Gemini, Qwen, Titan are speculative and require further validation.

VI. Agentic AI Platforms: Quantifying Orchestrator Overhead

A. Context

Agentic AI systems represent a significant evolution from simple request-response models. These platforms employ an orchestrating layer—often an LLM itself—to manage complex user interactions by understanding intent, decomposing tasks, planning sequences of actions, invoking tools or other AI models, and synthesizing results. This orchestration process inherently involves computational work, contributing to the overall environmental footprint. This section aims to investigate the typical patterns of LLM calls made by the orchestrator and to quantify, or establish a methodology for estimating, this computational overhead.

B. Platforms and Frameworks Under Review

The investigation considers specific platforms like ChatGPT Operator, Proxy by Convergence AI, and Nelima, as well as general agentic frameworks and concepts such as AutoGPT, BabyAGI, and ReAct, which illustrate common operational patterns.

C. LLM Call Patterns in Agentic Orchestration

Agentic workflows typically involve multiple LLM interactions for a single complex user query.[71] These calls serve distinct purposes within the orchestration layer:

1. **Task Understanding and Decomposition**: The orchestrator LLM first processes the user's request to understand the overarching goal. For complex goals, it then decomposes the task into smaller, manageable sub-tasks or steps.[71] This planning phase itself often requires one or more LLM calls.
2. **Tool/Agent Selection and Invocation**: For each sub-task, the orchestrator may need to decide if an external tool (e.g., a search engine, database API, code interpreter) or another specialized AI agent is required.[74] This can involve an LLM call to select the appropriate tool and another to format the query or command for that tool.
3. **Observation Interpretation and Re-planning (Iterative Reasoning)**: After a tool executes or a sub-agent responds, the orchestrator receives the output (an "observation"). It then makes an LLM call to interpret this observation, assess progress towards the goal, and decide the next action. This forms an iterative

loop, common in frameworks like ReAct (Reason-Act).[78] Each iteration of "Thought" in ReAct can be an LLM call.

4. **Result Synthesis**: Once all necessary information is gathered and sub-tasks are completed, the orchestrator may make a final LLM call to synthesize the individual results into a coherent answer for the user.
5. **Self-Reflection and Correction**: Advanced agentic systems might include self-reflection steps, where an LLM call is used to evaluate the agent's own plan or previous actions and make corrections, adding to the call count.[80]

The number of LLM calls made by the orchestrator for a single complex user task can therefore range from a few to potentially dozens, depending on the task's complexity, the number of tools or sub-steps required, and the agent's reasoning capabilities.[78]

D. Token Counts and Computational Overhead

Each LLM call made by the orchestrator has an associated token count for both the prompt fed to it and the completion it generates.

- **Orchestrator Prompts**: These can be substantial, often including system instructions defining the agent's role and capabilities, descriptions of available tools, the history of the current interaction (user queries, previous agent actions and observations), and the specific reasoning task at hand (e.g., "Given this goal and current state, what is the next best action?").
- **Orchestrator Completions**: These are the "thoughts" of the agent, the selected actions, formatted tool calls, or the synthesized final plans.

The computational overhead of the agentic layer is, therefore, the cumulative energy and water footprint of all these internal LLM calls made by the orchestrator. While direct measurements of "Wh per orchestrator task" are not readily available, the research points to this overhead being a function of the number and token-length of these internal calls.

Studies like "History-based Agent Optimization Kit (HAOK)" [81] highlight that unoptimized agents can lead to "frequent calls to large language models and high token consumption" due to repetitive planning or code generation. Similarly, the "Performant LLM Agentic Framework (PAF)" [82] notes that planning phases in agent systems can require "additional LLM calls, which adds computational overhead," and aims to reduce these. These studies implicitly confirm that the orchestrator's LLM calls are a source of significant computational load.

E. Deeper Insights & Causal Relationships

The adoption of agentic AI introduces a multiplier effect on LLM usage. A single user query can trigger a cascade of internal LLM calls within the orchestrator, each contributing to the overall environmental footprint. This trade-off—increased LLM call volume for enhanced capabilities like tool use and complex problem-solving—is central to agentic AI. The "intelligence" or autonomy of an agent often correlates with the number of reasoning steps (and thus LLM calls) it performs.

The complexity of the user's goal is a primary driver: more intricate tasks necessitate more decomposition, more tool interactions, and potentially more reflection cycles, all leading to a higher number of LLM calls by the orchestrator. The verbosity of the orchestrator's internal "thoughts" or reasoning traces also directly impacts the token count of its LLM calls, thereby affecting their individual energy costs.

If agentic AI systems become widely adopted, this orchestrator overhead could significantly amplify the total demand for LLM inference. Even if the individual LLMs used by the orchestrator are relatively efficient per call, the sheer volume of calls for complex tasks could result in a substantial cumulative footprint. The architectural design of the agentic framework itself—for example, a ReAct-style iterative loop versus a more structured multi-stage planner, or the use of a single powerful LLM as orchestrator versus a committee of smaller, specialized LLMs—will profoundly influence the magnitude of this overhead. Current research is actively exploring ways to optimize agentic workflows to minimize redundant calls and improve overall efficiency.[81]

For the AI Footprint Calculator, representing this orchestrator overhead is crucial. It might be modeled as an additive component or a multiplier based on the estimated average number of internal LLM calls anticipated for different complexities of agentic tasks. For example, a simple agentic task might involve 2-3 orchestrator calls, while a highly complex research task could involve 10-20 or more. The token length of these internal calls would also need to be estimated (e.g., short to medium, as defined in Section V).

F. Proposed Table for Section VI

Table VI.A: Estimated Orchestrator LLM Call Overhead for Agentic AI Platforms per Complex User Task

Agentic Platform/Framework Type	Est. Orchestrator LLM Calls per User Task (Range)	Est. Avg. Token Count per Orchestrator Call (Input+Output)	Resulting Est. Orchestrator Energy Overhead (Wh/task) (Derived from Section V data)	Basis & Assumptions for Call/Token Estimates	Confidence Level
ReAct-based Agent (e.g., LangChain ReAct)	3 - 10+	500 - 2000 tokens	*Calculated*	Iterative Thought-Action-Observation loop; complexity-dependent iterations.[78] Assumes medium-complexity LLM for orchestration.	Medium-Low
Multi-Agent System (e.g., AutoGen-like)	5 - 15+	1000 - 3000 tokens (incl. inter-agent comms)	*Calculated*	Task decomposition among specialized agents, planning, communication, and synthesis steps.[73]	Low

Simple Tool-Using Agent	2 - 5	500 - 1500 tokens	*Calculated*	Initial planning/tool selection, tool call formatting, result interpretation [74].	Medium
Complex Planning/Research Agent (e.g., AutoGPT-like)	10 - 50+	1000 - 4000+ tokens	*Calculated*	Deep task decomposition, multiple tool uses, web searches, file operations, self-correction loops.[84]	Low

Note: "Calculated" energy overhead will be derived by multiplying the estimated number of calls by the estimated energy per call (based on token count and LLM type used for orchestration, using data from Table V.A). These are broad estimates due to high variability.

VII. Hardware and Infrastructure Context: Impact on Environmental Footprint

A. Context

The environmental footprint values for AI model inference derived in the preceding sections are intrinsically linked to the underlying hardware executing the computations and the infrastructure of the data centers housing this hardware. This section examines the impact of specific hardware choices (e.g., different GPU and TPU generations) and data center operational efficiencies—Power Usage Effectiveness (PUE) and Water Usage Effectiveness (WUE)—on the final energy and water consumption figures.

B. Hardware Impact: GPU and TPU Efficiency

The choice of AI accelerator significantly influences energy consumption per operation.

- **NVIDIA H100 vs. A100 GPUs**: The NVIDIA H100 GPU generally offers improved performance and energy efficiency compared to its predecessor, the A100.[11] The H100 has a Thermal Design Power (TDP) of up to 700W for the SXM version and around 350W for the PCIe version.[11] In contrast, the A100's TDP is up to 400W (SXM) and 250-300W (PCIe).[11] Cloud instances leverage these: AWS P5 instances use H100/H200 GPUs [26], while P4d instances use A100 GPUs.[24] Models benchmarked on an A100 (e.g., Open-Sora study [3], some ML.ENERGY results [2]) will yield different energy-per-task figures than if run on an H100, assuming optimal software utilization. The AI Footprint Calculator may need to incorporate hardware-specific efficiency factors or provide ranges.
- **Google TPUs**: Google's custom-designed Tensor Processing Units are optimized for their AI workloads. A significant trend is the generational improvement in efficiency: a 3x improvement in carbon-efficiency (a proxy for energy efficiency at constant grid intensity) was observed from TPU v4 to Trillium (TPU v6).[12] Furthermore, Trillium is reported to be over 67% more energy-efficient than its predecessor, TPU v5e.[28] Given that operational electricity accounts for over 70% of a TPU's lifetime emissions [12], these efficiency gains are crucial. Google models like Veo, Gemini, Lyria, and WaveNet are typically run on TPUs.
- **AMD Instinct MI300X**: While no specific performance-per-watt data for the MI300X was found in the provided snippets, it stands as a key competitor to NVIDIA's H100, and its relative energy efficiency would be a critical factor for AI workloads deployed on AMD hardware.
- **General GPU Power Consumption**: AI servers equipped with multiple GPUs are inherently power-intensive, consuming 10 to 15 times more energy than traditional CPU-only servers.[34] This underscores the concentration of energy demand in the AI-specific components of data centers.

C. Data Center Efficiency: PUE (Power Usage Effectiveness)

PUE measures the ratio of total energy consumed by a data center to the energy delivered to the IT equipment. A PUE of 1.0 is ideal, meaning all energy goes to IT load.

- **AWS**: Their latest data center designs target a PUE of 1.08.[8]

- **Google Cloud**: Reported a fleet-wide average PUE of 1.10 in 2023.[9] Some individual Google data centers achieve even lower PUEs, such as 1.04 or 1.05.
- **Microsoft Azure**: While their 2024 Environmental Sustainability Report discusses PUE improvements [93], a specific fleet-wide average was not in the provided snippets. The "How Hungry is AI?" study used an assumed PUE of 1.12 for Azure-hosted models.[7]
- **Meta**: Reported an average PUE of 1.08 across its operational data centers in 2023.[10] The PUE is a critical multiplier; the energy consumed by the AI model (IT energy) must be divided by the PUE to estimate the actual energy drawn from the grid by the data center for that workload, or multiplied by PUE if starting from IT energy to get total facility energy. For clarity in this report, energy values for AI tasks are typically IT energy; PUE is applied when calculating total data center draw or related water/carbon.

D. Data Center Efficiency: WUE (Water Usage Effectiveness)

WUE measures a data center's water efficiency, typically as liters of water consumed per kilowatt-hour of IT equipment energy (L/kWh_IT).

- **AWS**: Is committed to being water positive by 2030 and was 41% towards this goal by the end of 2023.[8] AWS utilizes reclaimed water in over 25% of its data centers.[8] The "How Hungry is AI?" study [7] used a WUEsite (on-site cooling water for IT energy) of 0.18 L/kWh_IT for AWS.
- **Google Cloud**: Does not state a fleet-wide WUE in L/kWh_IT in the provided snippets.[9] They report that in 2023, 83% of freshwater withdrawal was from low or medium-risk watersheds, and roughly one-third of their data center campuses used non-potable water sources or air-cooling methods.[9] The "How Hungry is AI?" study [7] used a WUEsource (indirect water for electricity generation, US average) of 3.142 L/kWh_grid for Google-hosted (via Azure in that study, but applicable for US-based GCP) models.
- **Microsoft Azure**: Their 2024 report states new data centers designed for AI workloads will consume zero water for cooling.[93] Microsoft aims to be water positive by 2030. One source reports a global WUE of 0.49 L/kWh for Microsoft,

and 0.1 L/kWh for its EMEA zone.[97] The "How Hungry is AI?" study [7] used a WUEsite of 0.30 L/kWh_IT for Azure.

- **Meta**: Reported a WUE of 0.18 L/kWh_IT in 2023.[10]
- **General Context**: The industry average WUE is often cited around 1.8 L/kWh [97], though this likely refers to total facility energy and may not distinguish direct/indirect use consistently. Highly efficient data centers can achieve WUEsite values below 0.2 L/kWh_IT. The distinction between on-site water for cooling (WUEsite) and off-site water for electricity generation (WUEsource) is vital for accurate footprinting, as highlighted in.[7]

E. Deeper Insights & Causal Relationships

Hardware efficiency and data center infrastructure design are pivotal in mitigating the environmental footprint of AI. The trend towards specialized AI accelerators (NVIDIA GPUs, Google TPUs, AWS custom silicon like Trainium [8]) reflects an industry-wide effort to improve performance per watt for AI-specific computations. These hardware-level efficiencies directly reduce the "IT equipment energy" component of the overall footprint.

Simultaneously, cloud providers are making substantial investments in improving data center PUE and WUE. Lower PUE values directly reduce the total energy overhead (for cooling, power distribution, lighting) associated with running the IT equipment. This, in turn, lowers the operational carbon emissions if the grid's carbon intensity remains constant. Similarly, a lower WUEsite indicates more efficient on-site water use for cooling per unit of IT energy, alleviating direct water stress in the data center's locality.

However, these factors are interconnected. For instance, some cooling technologies that reduce water use (lowering WUEsite) might be less energy-efficient, potentially increasing PUE. Conversely, highly efficient liquid cooling solutions, including direct-to-chip or immersion cooling [8], can handle higher-TDP chips more effectively, potentially allowing for denser server deployments and better overall energy efficiency, even if they involve fluid circulation.

The rapid evolution of AI hardware, while bringing performance and efficiency gains per operation, also fuels a quick refresh cycle for data center equipment. This raises concerns about increased e-waste and the embodied energy and resources in manufacturing new generations of chips and servers.[103]

The geographical location of data centers is also critical. Colder climates can leverage free cooling for more extended periods, improving PUE. Locations with access to low-carbon electricity grids reduce the operational carbon footprint. Similarly, areas with abundant, sustainably managed water resources, or access to non-potable water for cooling, are preferable from a water stewardship perspective.[9]

A key consideration is that while newer hardware like the H100 or advanced TPUs are more energy-efficient per FLOP or per inference operation compared to their predecessors [8], the concurrent growth in AI model size, complexity, and usage volume means that overall energy consumption from the AI sector continues to rise.[103] Efficiency gains at the component and facility level are crucial, but they may not be sufficient to counteract the escalating demand if model complexity and deployment scale massively. The AI Footprint Calculator should ideally allow users to select or understand the impact of different hardware generations and PUE/WUE assumptions on their estimated footprint.

F. Proposed Tables for Section VII

Table VII.A: Illustrative Impact of Hardware on Energy Consumption for Representative AI Tasks

AI Task Representative (Parameters)	**Hardware 1 (e.g., NVIDIA A100) Est. Wh/unit**	**Hardware 2 (e.g., NVIDIA H100) Est. Wh/unit**	**Hardware 3 (e.g., Google TPU v5e Est.) Wh/unit**	**Percentage Difference (H1 vs H2)**	**Data Source / Basis for Comparison**
Image Gen (SDXL 1024px, ~50 steps)	*From ML.ENERGY (A100)*	*From ML.ENERGY (H100)*	N/A	*Calculated*	ML.ENERGY data [2]

Text Gen (LLaMA-3-70B, Med. Prompt 2k tok)	*Est. from* [7] *data if on A100*	*Est. from* [7] *data (H100 assumed)*	*Est. based on relative TPU efficiency*	*Calculated*	[77], relative efficiency claims (e.g., H100 vs A100 from vendor specs or benchmarks if available for this specific task)
Video Gen (Zeroscope_v2_xl, 1s 1024x576)	*From ML.ENERGY (A100)*	*From ML.ENERGY (H100)*	N/A	*Calculated*	ML.ENERGY data [2]

Note: This table would require careful selection of comparable tasks and models from available benchmarks or robust estimations.

Table VII.B: PUE and WUE Values for Major Cloud Providers/Infrastructure (2023-2025 Estimates)

Cloud Provider / Infrastructure Type	**Reported/Est. PUE (Annual Average)**	**Reported/Est. WUEsite (L/kWh_IT) (On-site cooling)**	**Reported/Est. WUEsource (L/kWh_grid) (Off-site electricity gen.)**	**Year/Source**	**Notes**

AWS	~1.08 - 1.14	~0.18	~3.142 (US Avg)	PUE: [8]; WUEsite/source: [77] (for Anthropic/Meta on AWS)	AWS aiming for PUE 1.08 in new designs. WUEsite is low. WUEsource depends on grid.
Google Cloud	~1.10	*Not explicitly stated as L/kWh_IT*	~3.142 (US Avg)	PUE: [9]; WUEsource: [77] (US Avg for Azure, applicable)	Focus on non-potable water & air cooling. TPU efficiency gains are significant. [7] uses 0.30 L/kWh_IT for Azure, potentially similar for GCP.
Microsoft Azure	~1.12 - 1.15	~0.30 (US); ~0.10 (EMEA)	~3.142 (US Avg)	PUE: [77] (for OpenAI on Azure), general estimates; WUEsite: [7]; WUEsource: [77]	New AI DCs aim for zero water cooling.[93]

Meta	~1.08	~0.18	~3.142 (US Avg, if applicable)	PUE/WUEsite : 10	Very efficient PUE and WUEsite reported.
Generic Data Center (Older/Average)	~1.5 - 1.8	~1.8 (total facility, not just IT)	Grid dependent	98	Represents less optimized infrastructure.
DeepSeek Data Center (China Est.)	~1.27	~1.20	~6.016 (China Avg)	77	Higher WUE values reflecting regional averages.

These tables provide essential parameters for the AI Footprint Calculator to adjust baseline IT energy figures based on hardware and data center context.

VIII. Cross-Cutting Investigation: Operational Water Footprint

A. Context

This section consolidates all research findings pertaining to the operational water footprint of AI systems. It distinguishes between direct on-site water consumption (primarily for cooling IT equipment) and indirect water consumption (associated with the energy generation lifecycle supplying power to data centers). It also examines common water sources, cooling methodologies, and the increasingly critical issue of water scarcity in regions hosting data centers.

B. Direct vs. Indirect Water Consumption: The WUE Framework

The most comprehensive framework for understanding AI's water footprint is presented in "How Hungry is AI?" (arXiv:2505.09598).7 This study proposes the formula:

$$Water(L) = (Equery / PUE) \cdot WUEsite + Equery \cdot WUEsource$$

Where:

- Equery is the IT energy consumed per query (or AI task unit) in kWh.
- **PUE** (Power Usage Effectiveness) accounts for data center energy overhead. (Equery/PUE) represents the IT energy component if Equery was total facility energy, or Equery is IT energy and total facility energy is Equery $\cdot$ PUE. The formula as written implies Equery is IT energy, and the first term (Equery/PUE) is incorrect if Equery is IT energy; it should be Equery $\cdot$ WUEsite for direct on-site water related to IT energy, and the second term is for indirect water for the IT energy. Assuming Equery is the IT energy:
    - **Direct On-site Water**: Equery $\cdot$ WUEsite
        - WUEsite (L/kWh_IT): Liters of water consumed on-site for cooling per kWh of IT energy.
    - **Indirect Off-site Water**: (Equery $\cdot$ PUE) $\cdot$ WUEsource_grid *OR* if WUEsource is already per IT kWh (accounting for grid losses and PUE implicitly or if applied to total facility energy): Equery $\cdot$ WUEsource_IT. The [7] paper clarifies WUEsource is water for off-site electricity generation (L/kWh of electricity generated). So, the indirect water is (Equery $\cdot$ PUE) $\cdot$ WUEsource_grid.

The study in [7] uses the following specific values:

- **Microsoft Azure (US)**: PUE 1.12; WUEsite 0.30 L/kWh_IT; WUEsource 3.142 L/kWh_grid (US average).
- **AWS (US)**: PUE 1.14; WUEsite 0.18 L/kWh_IT; WUEsource 3.142 L/kWh_grid (US average).
- **DeepSeek Data Centers (China)**: PUE 1.27; WUEsite 1.20 L/kWh_IT; WUEsource 6.016 L/kWh_grid (China average).

David Mytton's analysis [107], citing older 2014 data (from Shehabi et al. 2016 based on 2003 US power data), indicated that water for electricity generation (indirect) was significantly higher (7.6 L/kWh_electricity) than on-site cooling water (1.8 L/kWh_total_DC_site_energy). The 1.8 L/kWh figure for on-site cooling is an average for total data center energy, not just IT energy, and is higher than recent figures from efficient hyperscalers.

A report by FoodandWaterWatch [106] mentions AI servers consuming up to 2.4 gallons/kWh (approx. 9.08 L/kWh) for on-site cooling, a figure considerably higher than those reported by major cloud providers for their efficient facilities and used in.[7] This

discrepancy highlights the variability and importance of using context-specific WUE values.

The Bird & Bird article [101] reinforces that electricity generation is generally the most water-intensive part of the chain, with on-site cooling being the data center's direct consumption point.

C. Common Water Sources and Cooling Methods

Data centers primarily utilize:

- **Potable Water**: Often the default source from municipal supplies.[100]
- **Reclaimed/Non-Potable Water**: Increasingly adopted by major providers to reduce reliance on freshwater. Google uses reclaimed or non-potable water at over 25% of its data center campuses.[9] AWS uses reclaimed water as an intake source for cooling at 20 of its data centers.[95]
- **Seawater**: An alternative used by Google in some locations (e.g., Hamina, Finland).[108]

Common cooling methods involving water include:

- **Evaporative Cooling**: Utilized in cooling towers and adiabatic economizers, where water evaporation dissipates heat. This is effective but leads to direct water consumption.[99]
- **Chilled Water Systems**: Water is cooled by chillers and circulated to absorb heat from IT equipment.[100]
- **Emerging Technologies**:
    - **Direct-to-Chip (DTC) Liquid Cooling**: Coolant is circulated directly over processors. This can reduce the need for large-scale air cooling and associated water use for air handling.[8]
    - **Immersion Cooling**: Servers are submerged in dielectric fluid. This can be highly efficient and reduce water needs for traditional cooling.[101]
    - **Zero Water Cooling for AI DCs**: Microsoft's new AI data centers are designed to consume zero water for cooling.[93]

D. Water Scarcity Considerations

The substantial water demand of data centers, particularly those running AI workloads, is a growing concern, especially in water-stressed regions:

- Reports indicate data centers impacting local water supplies, such as Google's facility in The Dalles, Oregon, which was found to use a significant portion of the town's water.[105]
- In 2021, 20% of US data centers were drawing water from moderately to highly stressed watersheds.[106] By 2023, this concern was amplified.[109]
- The global data center industry's water use is projected to grow significantly, making it a fast-growing industrial water consumer.[95] In response, leading tech companies (AWS, Google, Meta, Microsoft) have pledged to become "water positive" by 2030, aiming to replenish more water than they consume operationally.[8] These initiatives involve water restoration projects and improving on-site efficiency.

E. Deeper Insights & Causal Relationships

The water footprint of AI is a multifaceted issue, with direct on-site consumption for cooling IT equipment and often larger indirect consumption tied to the electricity supply chain. The indirect water footprint (WUEsource) can dominate, especially in regions where the electricity grid relies on water-intensive generation methods like coal or gas power plants (which use water for cooling and steam cycles). For example, the [7] study shows WUEsource values (3.142 L/kWh for US, 6.016 L/kWh for China) that are an order of magnitude larger than the WUEsite values for efficient hyperscale data centers (0.18-0.30 L/kWh_IT).

This implies that even with highly efficient on-site cooling (low WUEsite), the total water footprint of an AI workload is heavily influenced by the PUE of the data center and the water intensity of the electricity grid powering it. A higher PUE means more total energy is drawn from the grid per unit of IT energy, thus amplifying the indirect water footprint. Therefore, efforts to decarbonize the grid, particularly by shifting to low-water electricity sources like wind and solar PV (which have minimal water consumption during generation), are critical for reducing AI's indirect water impact.

The "water positive" commitments by major cloud providers are noteworthy. However, the efficacy and true impact of these initiatives depend on the nature of replenishment projects and whether they genuinely benefit the local watersheds affected by data

center operations or the watersheds impacted by electricity generation. Transparency in reporting WUE, clearly distinguishing between WUEsite and WUEsource (and the basis for WUEsource calculations, e.g., regional grid mix), is essential for accurate assessments.

The rapid growth of AI workloads is set to exacerbate these water challenges. As AI models become larger and more computationally intensive, they drive up IT energy demand, which in turn increases both direct cooling needs (if water-based cooling is used) and indirect water consumption via the electricity supply chain. The development of zero-water cooling systems for AI data centers, as pursued by Microsoft [93], and the adoption of advanced liquid cooling technologies [8] are crucial innovations to mitigate direct water use. However, the indirect water footprint remains a significant challenge tied to broader energy system decarbonization and water resource management.

F. Proposed Table for Section VIII

Table VIII.A: Summary of Operational Water Footprint Data and Factors for Representative AI Tasks

AI Modality/Task Example (Source Table)	Est. IT Energy (Wh/unit)	Assumed PUE	Assumed WUEsite (L/kWh_IT)	Assumed WUEsource_grid (L/kWh_grid)	Calc. Direct Water (mL/unit) $EIT \cdot WUEsite$	Calc. Indirect Water (mL/unit) $(EIT \cdot PUE) \cdot WUE_{source\_grid}$	Total Calc. Water (mL/unit)	Key Assumptions (Provider/Region for PUE/WUE)
Video: Zeroscope_v2_xl (1s @ 1024x576)	~96.1	1.14 (AWS)	0.18	3.142 (US Avg)	$96.1 \cdot 10^{-3} \cdot 0.18 \approx 0.017$ L = 17 mL	$(96.1 \cdot 10^{-3} \cdot 1.14) \cdot 3.142 \approx 0.345$ L = 345 mL	~362 mL	Hosted on AWS (US) [7]

(Table II.A)								
Audio: Tango 2 (1 min @ 100 steps) (Table III.A)	~0.0215	1.12 (Azure)	0.30	3.142 (US Avg)	$0.0215 \cdot 10^{-3} \cdot 0.30 \approx 0.000006$ L = 0.006 mL	$(0.0215 \cdot 10^{-3} \cdot 1.12) \cdot 3.142 \approx 0.000076$ L = 0.076 mL	~0.082 mL	Hosted on Azure (US)[7] (assuming A40 on Azure)
Image: SDXL 1.0 (1024 px) (Table IV.A)	~0.0291	1.14 (AWS)	0.18	3.142 (US Avg)	$0.0291 \cdot 10^{-3} \cdot 0.18 \approx 0.000005$ L = 0.005 mL	$(0.0291 \cdot 10^{-3} \cdot 1.14) \cdot 3.142 \approx 0.000104$ L = 0.104 mL	~0.109 mL	Hosted on AWS (US)[7] (assuming ML.ENERGY H100 on AWS)
Text: GPT-4o (Short Query) (Table V.A)	0.421	1.12 (Azure)	0.30	3.142 (US Avg)	$0.421 \cdot 10^{-3} \cdot 0.30 \approx 0.000126$ L = 0.126 mL	$(0.421 \cdot 10^{-3} \cdot 1.12) \cdot 3.142 \approx 0.00148$ L = 1.48 mL	~1.606 mL	Hosted on Azure (US)[7]

Text: DeepSeek-R1 (Long Query) (Table V.A)	33.634	1.27 (China)	1.20	6.016 (China Avg)	$33.634 \cdot 10^{-3} \cdot 1.20 \approx 0.04036$ L = 40.36 mL	$(33.634 \cdot 10^{-3} \cdot 1.27) \cdot 6.016 \approx 0.2568$ L = 256.8 mL	~297.16 mL	Hosted in DeepSeek DC (China)[7]

Note: This table applies the [7] water footprinting methodology. EIT is the IT energy per unit. Direct water is $EIT \cdot WUE_{site}$. Indirect water is $(EIT \cdot PUE) \cdot WUE_{source\_grid}$. Total water is their sum. Values are illustrative and depend on precise energy per unit and assumed infrastructure.

IX. Synthesis of Findings for AI Footprint Calculator Tiers

This section consolidates the quantitative data gathered and analyzed in the preceding sections (II-VIII) to propose representative operational energy (Wh/unit) and operational water (mL/unit) ranges for the defined tiers of the AI Footprint Calculator. The goal is to provide a structured set of values that reflect the current understanding of AI's environmental impact, acknowledging the variability and confidence levels associated with the data.

Methodology for Tier Population:

1. **Review Detailed Data**: Data from Tables II.A, III.A, IV.A, V.A, VI.A, and VIII.A are reviewed.
2. **Map to Calculator Tiers**: Relevant models, tasks, and benchmarks are mapped to the predefined calculator tiers.
3. **Determine Representative Ranges**: For each tier, a range (minimum-maximum, or a central estimate with uncertainty) for energy (Wh/unit) and water (mL/unit) is established. This considers the spread of available data points.
4. **Document Key Assumptions**: The defining characteristics for each tier (e.g., video resolution, audio complexity, prompt length) that correspond to the derived footprint ranges are explicitly stated.
5. **Assign Confidence Level**: Based on the quality, quantity, and directness of the underlying data for each tier.

A. Text Generation Tiers

- **Tier 1: Basic Text Generation (e.g., simple queries, short completions, low token count)**
    - **Energy**: 0.05 - 0.2 Wh per query (approx. <500 total tokens)
    - **Water**: 0.02 - 0.8 mL per query
    - **Assumptions**: Small models (e.g., LLaMA-3.2 1B, GPT-4.1 nano for short prompts), simple tasks.
    - **Supporting Data**: [7] (GPT-4.1 nano, LLaMA-3.2 1B short prompts).
    - **Confidence**: High.[7]
- **Tier 2: Standard Text Generation (e.g., moderate queries, summarization, moderate token count)**
    - **Energy**: 0.2 - 1.5 Wh per query (approx. 500 - 2000 total tokens)
    - **Water**: 0.8 - 6.0 mL per query
    - **Assumptions**: Medium-sized efficient models (e.g., GPT-4o, Claude-3.7 Sonnet for medium prompts), standard tasks.
    - **Supporting Data**: [7] (GPT-4o medium, Claude-3.7 Sonnet medium), Epoch AI (GPT-4o typical).
    - **Confidence**: High (for models benchmarked).
- **Tier 3: Advanced Text Generation (e.g., complex reasoning, long document processing, very high token count, specialized reasoning models)**
    - **Energy**: 1.5 - 40+ Wh per query (approx. >2000 total tokens, or complex reasoning tasks)
    - **Water**: 6.0 - 300+ mL per query
    - **Assumptions**: Large models (e.g., Claude-3.7 Sonnet ET, DeepSeek-R1, OpenAI o3 for long prompts or reasoning), computationally intensive tasks. The upper end reflects very long inputs (100k tokens [19]) or highly iterative reasoning models.
    - **Supporting Data**: [7] (Claude-3.7 Sonnet ET long, DeepSeek-R1 long, o3 long), Epoch AI (100k token input). Estimates for Claude 3 Opus, Gemini Advanced would fall here.
    - **Confidence**: Medium (benchmarked models); Low (for specific commercial models like Opus/Gemini Adv. requiring estimation).

B. Image Generation Tiers

- **Tier 1: Standard Image Generation (e.g., ~0.25-1MP resolution, standard diffusion steps)**
    - **Energy**: 0.01 - 0.5 Wh per image

    - **Water**: 0.04 - 2.0 mL per image [7]
    - **Assumptions**: Models like Stable Diffusion v1.5, moderate resolutions (512x512 to 1024x768), standard steps (20-50).
    - **Supporting Data**: ML.ENERGY (SD v1.5 [2]), consumer GPU estimate.[47]
    - **Confidence**: Medium.
- **Tier 2: High-Detail/Complex Image Generation (e.g., >1MP resolution, high diffusion steps, advanced features)**
    - **Energy**: 0.5 - 3.0+ Wh per image
    - **Water**: 2.0 - 12.0+ mL per image (derived)
    - **Assumptions**: Models like SDXL 1.0, higher resolutions (1024x1024 and up), >50-100 diffusion steps, or use of features like advanced upscaling.
    - **Supporting Data**: ML.ENERGY (SDXL 1.0 [2]), Stable Diffusion 3 Medium estimate [5], Greenly.earth estimate.[6]
    - **Confidence**: Medium-Low (wider range due to variability in "complexity").

C. Video Generation Tiers

- **Tier 1: Short/Simple Video (e.g., <15s, ≤720p, simple animation/avatar, low complexity)**
    - **Energy**: 0.01 - 5.0 Wh per second (leading to 0.15 - 75 Wh per 15s clip)
    - **Water**: 0.06 - 300 mL per 15s clip (derived, highly dependent on specific energy and infrastructure)
    - **Assumptions**: Highly optimized tasks like Synthesia avatar generation [17], or efficient open models at lower resolutions/complexity (e.g., scaled Open-Sora estimates for 480p, Amazon Nova Reel 720p estimate).
    - **Supporting Data**: Synthesia [17], Amazon Nova Reel estimate, Open-Sora (scaled).[3]
    - **Confidence**: Low-Medium (due to wide range and reliance on estimates for commercial platforms).
- **Tier 2: Long/Complex Video (e.g., >15s, ≥1080p, cinematic quality, high scene complexity)**
    - **Energy**: 20 - 100+ Wh per second (leading to 600 - 3000+ Wh per 30s clip)
    - **Water**: 2400 - 12000+ mL per 30s clip (derived)

    - **Assumptions**: High-fidelity models like OpenAI Sora (user est. [1]), advanced diffusion models like Zeroscope_v2_xl [2], CogVideoX estimate.[5] High resolution, many denoising steps.
    - **Supporting Data**: ML.ENERGY (Zeroscope_v2_xl), CogVideoX estimate, Sora user estimate.
    - **Confidence**: Low-Medium (data points exist but vary widely, commercial model data is scarce).

D. Audio Generation Tiers

- **Tier 1: Simple Audio (e.g., Basic TTS, short simple instrumental loops)**
    - **Energy**: 0.001 - 0.05 Wh per minute
    - **Water**: 0.004 - 0.2 mL per minute (derived)
    - **Assumptions**: Efficient TTS models (WaveNet-like [29]), efficient diffusion models like AudioLDM at low inference steps.[4]
    - **Supporting Data**: "Diffused Responsibility" (AudioLDM) [4], Google WaveNet speed (energy estimated).
    - **Confidence**: Medium (for benchmarked open models/TTS); Low (for broader commercial TTS).
- **Tier 2: Complex Audio (e.g., High-quality expressive TTS, complex multi-track music with vocals, advanced voice cloning inference)**
    - **Energy**: 0.05 - 1.0+ Wh per minute
    - **Water**: 0.2 - 4.0+ mL per minute (derived)
    - **Assumptions**: More intensive diffusion models [4], or estimations for platforms like Suno or ElevenLabs for complex outputs.
    - **Supporting Data**: "Diffused Responsibility" (Tango2).[4] Estimates for Suno/ElevenLabs are highly speculative.
    - **Confidence**: Low (significant data gaps for commercial complex audio).

E. Analytical/Classification Tasks (1 Tier)

- **Context**: This tier was not a primary focus of the detailed research prompt but some data exists. Greenly.earth [6] provides:
    - Text classification: 0.002 kWh/1000 queries = 0.002 Wh/query.
    - Image classification: 0.007 kWh/1000 queries = 0.007 Wh/query.

    - Object detection: 0.038 kWh/1000 queries = 0.038 Wh/query.
- **Energy**: 0.002 - 0.04 Wh per query/item.
- **Water**: 0.008 - 0.16 mL per query/item (derived).
- **Assumptions**: Standard classification/detection models, likely smaller than large generative models.
- **Confidence**: Low-Medium (based on limited data points from one secondary source).

F. Agentic AI Platforms (Orchestrator Overhead)

- **Context**: The overhead is the sum of the orchestrator's internal LLM calls.
- **Energy**: Additive per complex user task. Estimated 2-15+ LLM calls by the orchestrator, each call consuming energy based on its token length and the orchestrator LLM's efficiency (from Text Gen Tiers).
    - Example: 5 orchestrator calls, each a "medium prompt" using a Tier 2 LLM (e.g., 1 Wh/call) = 5 Wh orchestrator overhead.
- **Water**: Derived from the orchestrator's energy overhead.
- **Assumptions**: Number of calls and token length per call vary greatly with task complexity and agent design.[78]
- **Supporting Data**: Qualitative descriptions of agentic workflows [71], ReAct framework.[78]
- **Confidence**: Low (highly dependent on specific agent implementation and task).

This synthesis reveals that while some tiers (e.g., Text Generation Tier 1 & 2, parts of Image Generation) have reasonably robust data from benchmarks, others (especially advanced Video/Audio, and Agentic Overhead) rely more on estimation, scaling from related tasks, or are subject to wide ranges due to lack of transparency from commercial providers.

X. Identified Data Gaps and Recommendations for Future Research

This investigation into the environmental footprint of AI systems has successfully collated a range of quantitative data and qualitative insights. However, it has also highlighted significant areas where data remains scarce, of low confidence, or altogether absent. Addressing these gaps is crucial for enhancing the accuracy and comprehensiveness of the AI Footprint Calculator and for fostering a more sustainable AI ecosystem.

A. Critical Data Gaps

The following are key areas where specific, verifiable operational energy and water footprint data are lacking:

1. **Proprietary Commercial Video Generation Models**:
    - **OpenAI Sora and Google Gemini Veo**: Despite their high profile and potential impact, no official or independently benchmarked operational energy (Wh/second or Wh/clip) or water consumption (mL/unit) data for inference has been publicly released by the providers. Current estimates are speculative or user-derived.[1]
    - **Pyxa, 1minAI**: No specific environmental footprint data was found for these user-requested platforms within the reviewed materials.
2. **Advanced Commercial Audio Generation Platforms**:
    - **Suno (Music Generation)**: While known for generating complex, full-length songs, its operational energy and water footprint per minute of generated audio remains undisclosed.[34]
    - **ElevenLabs (Voice Cloning & TTS)**: Similarly, specific inference energy/water data for its advanced voice cloning and expressive TTS services is not publicly available.[37]
3. **Specific Large Language Models (LLMs) for Tier 3 Text Generation**:
    - **Anthropic Claude 3 Opus**: While its capabilities are touted [56], direct operational energy/water benchmarks are missing. Estimates rely on comparisons to other Claude models [7] and pricing structures.
    - **Google Gemini Advanced (specific tasks with 1.5 Pro/Ultra)**: Quantitative data for computationally intensive tasks using Google's most capable Gemini models is needed beyond general efficiency statements.[28]
    - **Alibaba Qwen (Large Versions)**: While MoE architecture promises efficiency [62], specific Wh/1k token benchmarks for large Qwen models (e.g., Qwen2-72B, Qwen3-235B) are not available in the reviewed academic literature.
    - **Amazon Titan Text Models (Large Versions)**: Similar to other large proprietary models, direct operational energy/water figures for models like Titan Text Premier are lacking.[64]
4. **Detailed Impact of Generation Parameters**:

- **Video/Image**: While resolution and duration impacts are somewhat characterized (e.g., Open-Sora study [3]), the precise quantitative impact of varying diffusion steps, specific quality settings, or advanced features (in-painting, out-painting, complex upscaling) on energy consumption across a range of models needs more systematic benchmarking.

5. **Agentic AI Platform Orchestrator Overhead**:
   - Quantitative data on the exact number and token complexity of LLM calls made by the orchestrator layer of platforms like ChatGPT Operator, Proxy by Convergence AI, or Nelima for various complex user interactions is scarce. Current understanding relies on qualitative descriptions of agentic frameworks like ReAct.[78]
6. **Comparative Hardware Benchmarks**:
   - Energy consumption benchmarks for **AMD Instinct MI300X GPUs** performing AI tasks comparable to those benchmarked on NVIDIA H100/A100 or Google TPUs are needed for a more complete hardware impact assessment.
7. **Standardized Data Center WUE Figures**:
   - While PUE is relatively well-reported, comprehensive, standardized, and regularly updated **WUEsite (L/kWh_IT for on-site cooling)** and **WUEsource (L/kWh_grid for electricity generation, specific to regional grid mixes)** figures from all major cloud providers (AWS, Azure, Google Cloud) are essential. This data needs to be clearly tied to the regions and specific data center types hosting intensive generative AI workloads. The methodology in [7] provides a good template, but relies on some national averages where provider data is missing.

B. Recommendations for Future Research and Calculator Refinement

1. **Advocacy for Transparency**: Encourage AI model developers and cloud providers to publicly disclose operational energy and water consumption data for their services, using standardized metrics and methodologies. This is the most direct path to improving data quality.
2. **Standardized Benchmarking Initiatives**: Support and expand independent benchmarking efforts like ML.ENERGY [2] and the AI Energy Score [111] to cover a wider range of commercial and open-source models (especially for video and audio), hardware types (including AMD GPUs and newer TPUs), and generation parameters. Benchmarks should explicitly report energy (Wh or J per relevant

unit), hardware used, key model parameters, PUE, and ideally WUEsite/WUEsource context.

3. **Targeted Empirical Studies**: Conduct dedicated research to measure the operational footprint of high-priority models and platforms where data is currently lacking. This could involve:
    - Developing methodologies to estimate energy consumption based on API usage, processing time, and inferred hardware, similar to the approach in "How Hungry is AI?" [7], but extended to video and audio services.
    - Empirical studies on agentic AI platforms to quantify orchestrator LLM calls (number and token characteristics) for representative complex tasks.
4. **Refinement of Proxy Metrics and Estimation Models**:
    - For the AI Footprint Calculator, where direct data is unavailable, develop and clearly document robust proxy metrics or estimation models. For instance:
        - Scaling laws based on model parameters (e.g., denoising parameters for diffusion models), FLOPs, or relative performance benchmarks, while explicitly stating limitations and confidence levels. The quadratic scaling of video resolution from Open-Sora [3] is an example.
        - Using pricing data (e.g., cost per token, cost per minute of audio/video) as a rough proxy for computational intensity, combined with estimated energy costs per FLOP on typical hardware, but this should be treated with extreme caution and low confidence.
5. **Dynamic Calculator Updates and User Contributions**:
    - Design the AI Footprint Calculator to be easily updatable as new data becomes available.
    - Clearly indicate the confidence level and source for each data point within the calculator.
    - Consider a mechanism for users or researchers to contribute new, verifiable data points, subject to a validation process.
6. **Focus on Water Footprint Granularity**: Promote research and reporting that distinguishes between direct on-site water use (WUEsite) and indirect water use from electricity generation (WUEsource), and that considers local water stress conditions and water source types (potable, reclaimed, etc.) for data center locations.
7. **Lifecycle Assessment**: While this report focuses on operational footprint, encourage broader research into the full lifecycle impacts of AI, including embodied energy and water in hardware manufacturing and disposal (e-waste [103]).

By systematically addressing these data gaps and implementing robust estimation methodologies where necessary, the AI Footprint Calculator can become an increasingly valuable tool for promoting awareness and driving sustainability in the development and deployment of artificial intelligence.

X. Conclusion

The proliferation of generative AI technologies presents both transformative opportunities and significant environmental challenges. This report has sought to provide a data-driven foundation for understanding and quantifying the operational energy and water footprint of a diverse range of AI models and tasks, with the explicit goal of populating and refining an AI Footprint Calculator.

The analysis reveals a complex and rapidly evolving landscape. While some areas, such as text and image generation using certain open-source models, have emerging benchmarks for energy consumption [2], substantial data gaps persist for many cutting-edge commercial offerings, particularly in the high-demand modalities of video and complex audio generation. The operational energy for generating a single second of high-definition video, for instance, can range from a few Watt-hours to nearly 100 Wh, depending on the model, complexity, and hardware.[1] Similarly, advanced text generation involving complex reasoning can be orders of magnitude more energy-intensive per query than simpler text completion tasks.[7]

Water consumption, intrinsically linked to energy use through data center PUE and regional WUE (both for on-site cooling and off-site electricity generation), follows these energy trends. The methodology detailed in "How Hungry is AI?" [7] provides a robust framework for estimating this, highlighting that indirect water use for electricity generation can often exceed direct on-site cooling water.

A critical overarching theme is the urgent need for greater transparency from AI model developers and cloud infrastructure providers. Without access to standardized, verifiable operational footprint data, users and developers cannot make fully informed decisions regarding the environmental impact of their AI choices. The AI Footprint Calculator, by consolidating available data and clearly indicating areas of estimation and uncertainty, can serve as a vital tool in this context.

The impact of hardware generations (e.g., NVIDIA A100 vs. H100, Google TPU advancements) and data center efficiencies (PUE, WUE) is undeniable, offering pathways to mitigate some of the escalating energy and water demands. However,

these efficiencies may be outpaced by the sheer growth in model complexity and usage volume if not coupled with conscious design choices and a shift towards sustainable energy sources.

This report provides a snapshot of current knowledge and robust estimates where possible. It underscores that the environmental footprint of AI is not a fixed attribute but a dynamic interplay of model architecture, task complexity, hardware efficiency, data center practices, and energy sourcing. Continued research, standardized benchmarking, and a collective commitment to transparency are paramount for navigating a future where AI innovation aligns with environmental stewardship. The AI Footprint Calculator is a positive step in this direction, empowering users with the information needed to make more sustainable choices in the age of artificial intelligence.