the brandtech group



Generative AI **Environmental Impact Study**

MAY 2025

Table of Contents

Foreword 3

Introduction 5
Rising Energy Demands and Global Impacts 6
How the Tech Industry is Responding 7
The Path Forward 8

Environmental Impact 9
Calculation Methodology
Model Training 10
Inferences 11

Marketing Use Case Calculations 12
Use Case #1: Product Page Copy
Use Case #2: Static Creative Assets 15

Insights and Best Practices 17

Estimate of Emission Repartition and Mitigation Impact 23

Conclusion and Next Steps 26

Appendix 28
Glossary
Contributors & Authors 33
Detailed Calculation Methodology by Emission Source 36

Foreword by Rebecca Sykes, The Brandtech Group

The conversation around AI and sustainability has, for too long, been built on assumptions rather than data. While many of us in the industry acknowledge the theoretical environmental impact of AI, there's been little in the way of concrete measurement or actionable insights. That's why we embarked on this project - to build a foundation of knowledge that will allow us, and the broader marketing industry, to take meaningful steps forward.

This whitepaper is not a one-and-done exercise. It marks the beginning of a sustained commitment from The Brandtech Group and our partners to ensure that generative AI is deployed responsibly. The insights here may seem intuitive to some, but the reality is that, until now, very few had even attempted to quantify the environmental footprint of AI-driven marketing and media. Without measurement, there can be no meaningful mitigation. This report is our first step in addressing that.

Our focus now is twofold: first, to deepen our understanding of AI's impact - because the complexity of this issue cannot be overstated. The AI ecosystem is vast and interconnected, spanning multiple platforms, partners, and workflows, many of which lack transparency. As AI becomes increasingly multimodal, the challenge of assessing its true footprint only grows. And second, to turn insights into action. We are committed to training our people, equipping our clients, and embedding best practices into our platforms so that sustainability is not an afterthought, but an inherent part of how AI is developed and deployed.

Our clients are deeply invested in this, too. They have their own ESG targets to meet, and they need partners who can help them navigate AI's rapid evolution with sustainability in mind. We want to be those partners. By measuring impact, sharing knowledge, and pushing for accountability, we believe we can drive real change, not just in our own work, but across the entire industry.

David Jones, founder and CEO, The Brandtech Group:

"We are committed both to leveraging the incredible opportunities Gen AI brings for our brand partners, but also to not shying away from some of the big challenges that come with it. From bias to legal questions to carbon footprint, there are several. This white paper aims to start addressing the critical issue of what the environmental impact of using Gen AI is in marketing, and how can we accurately measure that with a view to being able to reduce it. It's a first toe in the water in a very important area.."

Ludovic Moulard, Global Sustainability Director, fifty-five:

"The issue at hand is not the technology itself but how we use it: GenAI can provide incredible value for companies, administrations as well as individuals - but since it's a highly energy-intensive tech, it has to be used with caution, for valuable use cases, keeping track of environmental impacts and deploying mitigation techniques to avoid wasting our valuable resources, trashing our planet and therefore our economy. This study aims at providing the right data, insights and techniques for both GenAI end users and businesses."

Brian O'Kelley, co-founder and CEO, Scope3

"As AI transforms the marketing ecosystem, we are missing crucial data on its environmental impact," said Scope3 CEO Brian O'Kelley. "This research gives marketers a reference point to start making informed decisions about AI implementation. The companies leading the way will be those tracking emissions from day one, optimizing their AI usage, and building sustainability directly into their workflows. When marketers can see the full picture – both performance and environmental impact – they can deploy AI strategically to maximize results while minimizing waste. This is how we turn AI's enormous potential into responsible, sustainable growth."

Introduction

In January 2023, the world of artificial intelligence changed forever. ChatGPT became the fastest-growing consumer application in history[1], amassing more than 100 million users in just two months. This extraordinary adoption not only propelled OpenAI into competition with established AI leaders like Google's DeepMind but also ignited a global race to develop increasingly advanced large language models. Organizations of all kinds—tech giants, startups, private companies, and governments—scrambled to leverage this transformative technology. By the end of 2023, investment in generative AI had surged ninefold compared to 2022 and was roughly 30 times higher than in 2019[2].

Yet behind this explosive growth lies an often-overlooked reality: the substantial environmental cost of generative AI. Every interaction with a model like ChatGPT requires energy, whether it's drafting an email, debugging code, or offering advice. A single ChatGPT query consumes 2.9 watt-hours of electricity—nearly ten times the energy of a standard Google Search, which uses 0.3 watt-hours[3]. While this might seem inconsequential for one query, the cumulative impact is staggering. Consider this: a single user making 100 queries a day generates a carbon footprint equivalent to driving a car for one mile[4]. On a global scale, the daily energy demands of generative AI now rival the consumption of small cities.

Sources:
1) https://www.forbes.com/sites/cindygordon/2023/02/02/chatgpt-is-the-fastest-growing-ap-in-the-history-of-web-applications/
2) https://hai.stanford.edu/news/ai-index-state-ai-13-charts
3) https://www.goldmansachs.com/insights/articles/AI-poised-to-drive-160-increase-in-power-demand
4) https://piktochart.com/blog/carbon-footprint-of-chatgpt/

Rising Energy Demands and **Global Impacts**

The widespread adoption of AI is driving unprecedented strain on energy infrastructure. Data centers—essential for running these AI systems—are at the heart of this challenge. In 2023, data centers accounted for 1-2% of global electricity use, equivalent to powering 17 million homes. But this demand is rising rapidly. Goldman Sachs Research projects a 160% increase in data center power requirements by 2030, pushing their share of global electricity consumption to 3-4%.

The United States illustrates this trend. US data centers consumed 3% of the nation's power in 2022. By 2030, this figure is expected to rise to 8%, necessitating approximately $50 billion in new electricity generation capacity. These trends represent the most significant increase in electricity demand in a generation.

How the Tech Industry is **Responding**

To mitigate these challenges, major tech companies are making substantial investments in renewable and alternative energy sources. Google, for instance, has partnered with Kairos Power to generate 500 megawatts of carbon-free nuclear power—enough to supply electricity to 438,000 US households.[5]

However, even with these initiatives, sustainability commitments are becoming harder to maintain. Google, which previously claimed to have achieved carbon neutrality since 2007, admitted in 2023 that the rising energy demands of AI infrastructure made this achievement unsustainable. Similarly, Microsoft's goal of becoming carbon-negative by 2030 faces significant hurdles. Its emissions have increased by 29% in 2023 compared to 2020, which is the equivalent of adding 754,348 cars to the road. And the upcoming publication of 2024 numbers, with the tremendous pace of AI development during that year, may very well show worse results.

Goldman Sachs estimates the social cost of these rising emissions at $125-140 billion (in present value), underscoring the need for urgent action.

Sources:
5) https://www.businessinsider.com/google-nuclear-power-ai-data-centers-2024-10

The Path Forward: Transparency and **Sustainability**

Despite the growing environmental impact, many companies at the forefront of the Gen AI revolution, including OpenAI, have yet to fully disclose the emissions associated with training and deploying large language models. Greater transparency is essential to understanding the true cost of this technology and driving responsible use.

At The Brandtech Group, we are witnessing firsthand how AI is transforming the marketing industry, improving processes, boosting creativity, and reducing cost. However, as very early adopters of Gen AI tech in marketing and leaders in the space, we believe we have a responsibility to understand and mitigate the environmental cost of this technology, ensuring its advancement aligns with global sustainability efforts.

Developed in collaboration with Scope3, a company dedicated to decarbonizing advertising and AI, this whitepaper explores the carbon cost of generative AI in marketing. Leveraging open-source data and independent research, it examines the environmental impact of automating marketing content creation and aims to provide practical recommendations and best practices.

This is an early first step in understanding AI's environmental impact—a complex and evolving issue that requires ongoing exploration. As Gen AI technology and our knowledge of its impact continue to evolve, The Brandtech Group remains committed to an ongoing effort to drive more sustainable AI adoption in marketing. To that end, we will publish future iterations of this report with updated data and research later in 2025.

Environmental Impact **Calculation Methodology**

In order to calculate the environmental impact of Gen AI, our developed methodology integrates two key elements of the process, which are model training and inferences.

Training refers to the process where an AI model studies patterns from large datasets to develop its capabilities. It requires significant computational resources, including GPUs (graphics processing units) and CPUs (central processing units), along with data storage for maintaining datasets and saving the model's learned parameters.

Inference is the phase where the trained model generates predictions or responses based on user inputs. This process involves utilizing the model's training to produce outputs, again relying on computational resources for efficient execution.

This report will focus on the carbon impact of these two elements. In future iterations, this methodology will incorporate water consumption used to cool servers in data rooms, as well as abiotic resource depletion which refers to the use of non-renewable metals and minerals for hardware infrastructure.

The calculation of carbon emissions is based on runtime, which is the execution time required to run AI models. For this whitepaper, runtime is measured in GPU hours but will take into account all computational components (GPUs, CPUs, and RAM). Detailed explanations are provided in the appendix.

Environmental Impact of **Model Training**

Understanding the carbon footprint of training AI models requires examining the primary contributors to energy consumption and embodied emissions.

Electricity consumption is a key factor in carbon emissions. Energy is required both for storing data during the training phase and for the actual training itself. During training, GPUs, CPUs, and RAM consume significant energy to process datasets and execute machine learning calculations, with GPUs being the largest contributors to overall energy use.

To estimate electricity consumption and related emissions, we multiply the total runtime (in GPU-hours) of training by the average power per GPU[6] (W), then apply the electricity emission factor, using the global data center average[7].

Embodied emissions are another important factor. These emissions are the greenhouse gases produced through the lifecycle of data center equipment, including production, transportation, and disposal. The global distribution of these facilities adds to embodied emissions. Semiconductor fabrication is energy-intensive, and transporting components across continents further increases the environmental footprint of data center operations.

To estimate embodied emissions, we divide the total runtime (in GPU-hours) by the Lifetime per GPU, then apply the GPU emission factor. We use Boavizta's impact factors[8], which detail the emissions linked to the manufacture and life cycle of GPUs. Detailed explanations are provided in the appendix.

DIMENSIONS	TYPE OF IMPACT	TRAINING IMPACT FORMULA
Type of task Model size Model popularity		Avg training runtime **X** Avg power (W) per GPU **X** Data center avg carbon intensity Avg training runtime **/** Lifetime per GPU **X** GPU emission factor
• *Text+img-to-text* • *80-90B params* • *>100k downloads*		*1, 775, 120 GPU-hours X 0.7kW X 0.2 kgCO2e* *1, 775, 120 GPU-hours / 3 x 365.25 x 24 X 143 kgCO2e*
		Carbon footprint for this training model = ***~ 260 tCO2e***

Sources:

6) We use a 3 year baseline for lifetime per GPU, for a utilization rate of >60%.
Peak power capacity per GPU device used adjusted for PUE – from ML CO2 & Boavizta database.
7) ADEME x NégaOctet v1.5
8) https://github.com/Boavizta/boaviztapi/issues/65#issuecomment-1604242196

Environmental Impact of **Inferences**

Our approach to evaluating the environmental impact of model inferences focuses on leveraging open-source models and examining their application in real-world marketing scenarios. Closed-source models inherently lack transparency, as they do not provide access to the underlying data or computations required for precise environmental assessments. Therefore, we rely on open-source models, which enable replicable and detailed evaluations.

We drew on Lucconi's experiment[9] to conduct our experiments, following this process:

- Download & deploy open source models to our own cloud project
- Use codecarbon.io to library to monitor activity (RAM, CPU, GPU) & runtime
- Run text-to-text and text-to-image models with various prompts to compare runtime & energy consumption by task complexity
- Output the energy per 1000 queries for each model size

Detailed calculations are provided in the appendix.

Sources:
9) https://arxiv.org/pdf/2311.16863

Marketing Use Case Calculations

For this study, we chose to assess the emissions generated by automating common marketing tasks using Gen AI models - the generation of a 1,000-character-long product page copy and the generation of 20 static creative assets for a marketing campaign.

To assess these use cases, we built a calculator (which can be accessed here) relying on the methodology described above and the numbers taken from our experiments. The calculator assesses the energy consumption (kWh) and GHG emissions (kgCO2e) of executions of a list of Gen AI tasks, deployed with a given model version and type of output. Results from our experiments are also available within the calculator.

In the use cases below, we used Pencil Pro, a model aggregator for AI content generation, as a data source to understand typical workflows and to refine our hypotheses based on live cases.

Use Case #1: Product Page Copy

Marketers require detailed, up-to-date product page descriptions on their web and e-commerce sites across their entire product inventory. Product page descriptions are usually designed with a specific template, including a predefined text length range and bullets for the product title and description. Creating high-performing content for new SKUs can be very time-consuming, which is why using LLMs to generate product copy across different languages makes sense.

An AI content-generation system, such as Pencil Pro, lets users generate and edit text content over different large language models with custom assistants designed for e-commerce use cases. For an e-commerce manager, a typical workflow could look like this:

1. Select the e-commerce assistant and pick a LLM, such as GPT 4o or Gemini 1.5 Pro
2. Provide e-commerce guidelines specific to the brand, including expressions not to be used, tone of voice, and specific examples of successful existing product page descriptions
3. Provide details about the product featured in the description page, such as the product ID and metadata. Typically this includes information from a Product Inventory Management (or PIM) system
4. Generate a first version of the product description
5. Refine the first version and fine-tune manually if necessary
6. Iterate to the next product
7. Export to the PIM or Content Management System (or CMS) for publication

We utilize the following hypotheses in our carbon emission calculation:

A. Selecting the LLM GPT-4o mini, hosted in the US. This represents average grid carbon intensity
B. Choose inputs of 250 tokens for ~1,000 characters, plus 100 tokens for brand context, such as tone of voice
C. Choose outputs of 300 words, including ~250 words for the description
D. Choose one (1) AI rewrite for another 300 words
E. Redo, without the need to re-provide steps 1 and 2

Results from **Calculator**

Inference Carbon Footprint Calculator									
Description	Task	Models	Output size	Custom Output Size	Number of queries	Total electricity consumption (kWh)	Electricity emissions (kgCO2e)	Embodied emissions (kgCO2e)	Total emissions (kg)
				Number of words for Text, number of pixels for Images	*Write the number of repetitions for your task*	*Wh x Output size x Number of queries*	*Total Wh x electricity carbon intensity / 1000 (g to kg)*	*Hardware emissions x Runtime / Lifetime*	
1. Generate ecommerce ideal product page content									
1 product page characteristics with a total of ~300 words. Average of 4 versions to get a result meeting expectations	txt-to-txt	GPT-4o mini		300	4	0.009	0.004	0	**0.004**
Same thing for an inventory of 1,000 products created in a single language as output	txt-to-txt	GPT-4o mini		300	4, 000	8.79	3.53	0.31	**3.84**
Same thing for an inventory of 1,000 products created with 25 different cultural / language variations as output	txt-to-txt	GPT-4o mini		300	100, 000	219.75	88.34	7.74	**96.08**
Same thing for an inventory of 300, 000 products created with 50 different cultural / language variations as output	txt-to-txt	GPT-4o mini		300	60, 000, 000	131.850	53, 003.7	4, 643.96	**57, 647.66**

From our calculations, we can see that generating a single product page generates very little carbon impact, with a total of 4 gCO2e generated.

For a company creating 1,000 products every year, the impact of using Gen AI for these product descriptions is a few kgCO2e, roughly the impact of a meat-based meal. If that company wanted to automate further by creating multiple versions to match language or cultural specificities — 25 variations for instance — the resulting impact is around 100 kgCO2e which equals around 1,500 hours of video streaming.

For the biggest retailers, it is estimated that 300,000 new items could be released every year, published in 50 different versions[10]. If Gen AI was used for the related product pages, it would result in over 50 tCO2e emitted. That is roughly 30 round trips from Paris to New York.

__50 tCO2e__ emissions is roughly 30 round trips from Paris to New York

Sources:
10) https://www.mobiloud.com/blog/shein-statistics

Use Case 2:
Static Creative Assets for Digital Marketing Campaign

When running digital campaigns, marketers need to adapt their master creative assets to a multitude of platforms and ad formats, as well as customize content for their target audience to maximize effectiveness. This process has historically been tedious, time consuming, and expensive. As a result, marketers often run campaigns with subpar creative material, which hinders performance.

We will focus on a typical digital campaign that requires 20 static variants from a master creative asset to run across multiple social media and programmatic environments. This use case does not cover building the master asset, which encompasses other environmental impacts that are not assessed here.

AI platforms offer a convenient way to accelerate and optimize this process by leveraging text and generation models. A standard workflow, using a platform such as Pencil Pro, looks like the following, alongside the hypotheses utilized in our GHG emission calculation. Typical roles for this workflow include a strategist, AI art director and/or AI creator, copywriter, producer, and project manager.

1) Provide brand details (tone of voice, campaign brief, target audience, etc) as content for the models to optimize output quality
   - Our hypothesis is an input of 100 words to cover this

2) Use a LLM such as GPT 4o or Gemini 1.5 Pro to identify specific audience needs that can be leveraged for creative adaptation and copywriting
   - Hypothesis: 200 words of input, 500 words of output

3) Design a template for specific target media platforms
   - Not assessed (impact should be marginal)

4) Generate base text assets (call-to-action, headline, etc) that will be used to generate more creative variants with the template
   - Hypothesis: 100 words of input, 100 words of output

5) Iterate to finalize the base text assets
   - Same hypothesis as step 4

6) Generate base image assets that will be used to generate more creative variants with the template

- Hypothesis: 50 words of input, output = a 1024px square image with a model such as Imagen, Dall-E or Stable Diffusion 3.5

7) Iterate to finalize the base image assets

- Hypothesis: 30 iterations (3 base image assets, run 10 times to get the a suitable result)

8)Use a feed to associate each base asset with the template to generate multiple variants of the ad

- N/A (simple association of elements for ad rendering)

9) Validate the final ads

- N/A

10) Export to the DAM, ad server, media platform...

- Not assessed

Results from **Calculator**

Inference Carbon Footprint Calculator									
Description	Task	Models	Output size	Custom Output Size	Number of queries	Total electricity consumption (kWh)	Electricity emissions (kgCO2e)	Embodied emissions (kgCO2e)	Total emissions (kg)
				Number of words for Text, number of pixels for Images	*Write the number of repetitions for your task*	*Wh x Output size x Number of queries*	*Total Wh x electricity carbon intensity / 1000 (g to kg)*	*Hardware emissions x Runtime / Lifetime*	
1. Generate digital marketing creatives									
Step 2. Use a LLM such as GPT4-o or Gemini 1.5 Pro to identify specific audience needs that can be leveraged for creative adaptation and copywriting	txt-to-txt	GPT-4o mini		500	1	0.004	0.001	0	**0.002**
Step 4 & 5. Generate base text assets (call-to action, headline...) that will be used to generate more creative variants with the template, then iterate	txt-to-txt	GPT-4o mini		100	3	0.002	0.001	0	**0.001**
Step 3. & 7. - Generate base image assets (background image...) that will be used to generate more creative variants with the template, with 30 variations	txt-to-img	Stable Diffusion 3.5 Large Turbo			30	0.121	0.049	0.001	**0.050**
Step 3. & 7. for 100 master creatives used in a year	txt-to-img	Stable Diffusion 3.5 Large Turbo			3, 000	12.1	4.9	0.1	**5**
Step 3. & 7. for 100 master creatives, with multiple variants for 1,000 different audiences	txt-to-img	Stable Diffusion 3.5 Large Turbo			3, 000, 000	12, 127	4, 875	105	**4, 980**

Our calculations show us that the environmental impact of this B2B use case remains rather small for a single asset at 2 gCO2e. Even for a set of 100 master creative assets for which we'd apply this recipe, the emissions equal <5 kgCO2e, about the impact of a meat-based meal.

Gen AI, however, unlocks the possibility to generate highly-targeted versions of assets. In these situations, if a marketer intended to target, for example, 1,000 different audiences, they could reach ~5 tCO2e a year, equivalent to manufacturing approximately 15 laptops or 100,000 search executions.

Insights and **Best Practices**

Key Factors that Influence Environmental Impact

Based on the experiments published by Hugging Face[11] and the outputs of our experiments on model inference (detailed here), we note that, in general, the main determining factors of direct energy and carbon impact are, in order:

1. **The type of task** — text-to-text generation, text-to-image generation, classification, multi-purpose, etc
2. **The output size** — number of words, pixels, etc
3. **The carbon intensity of electricity powering the hosting data center**
4. **The model size** — number of parameters

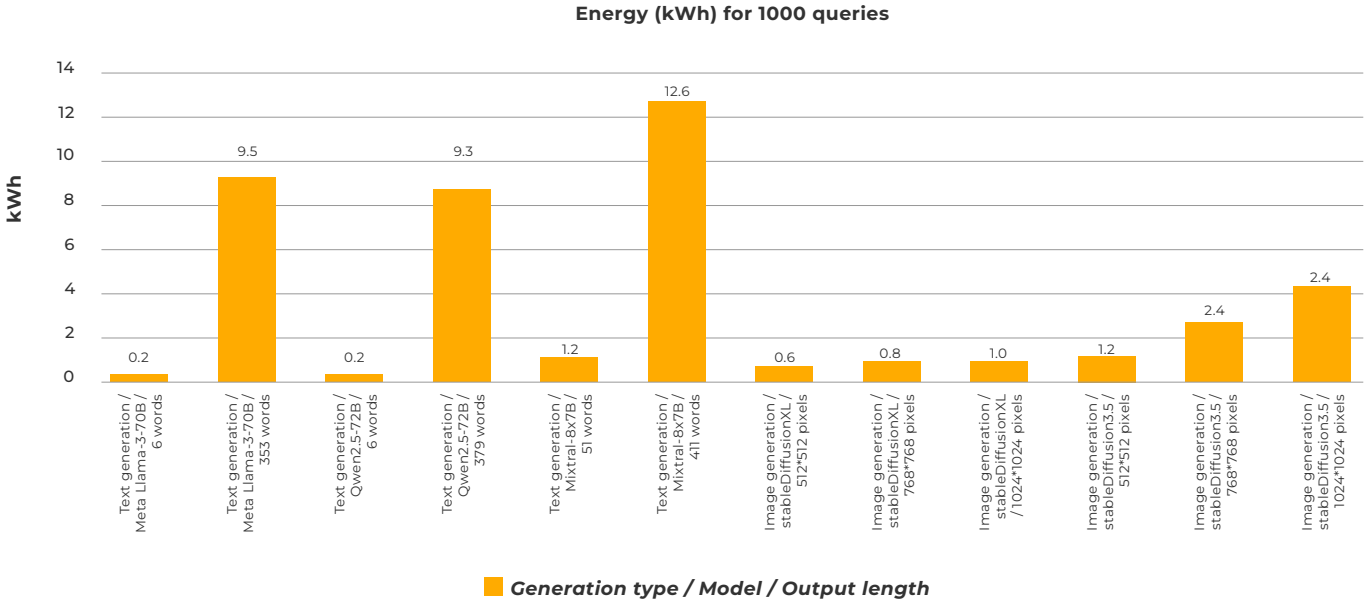


Chart 1: Bar chart illustrating the energy consumption by experiment

Type of Task

The complexity and size of inputs, particularly when dealing with text-based tasks, affect how much energy is required for processing. Text-to-text tasks, for instance, will generally require far less energy than text-to-image tasks due to fewer resources needed for generating text outputs, depending on the chosen model.

Sources:
11) https://arxiv.org/pdf/2311.16863

Output Size

The size and format of outputs—such as, text length or image resolution—will directly influence energy consumption. Larger or more complex outputs, such as high-resolution images or longer text, will require more computational power and generate more emissions.

Carbon Intensity

Depending on the energy mix of the data center used, the carbon intensity may vary widely from at least a 1 to 10 ratio. Selecting the right data center location, when possible, will have a long-term impact on GHG emissions.

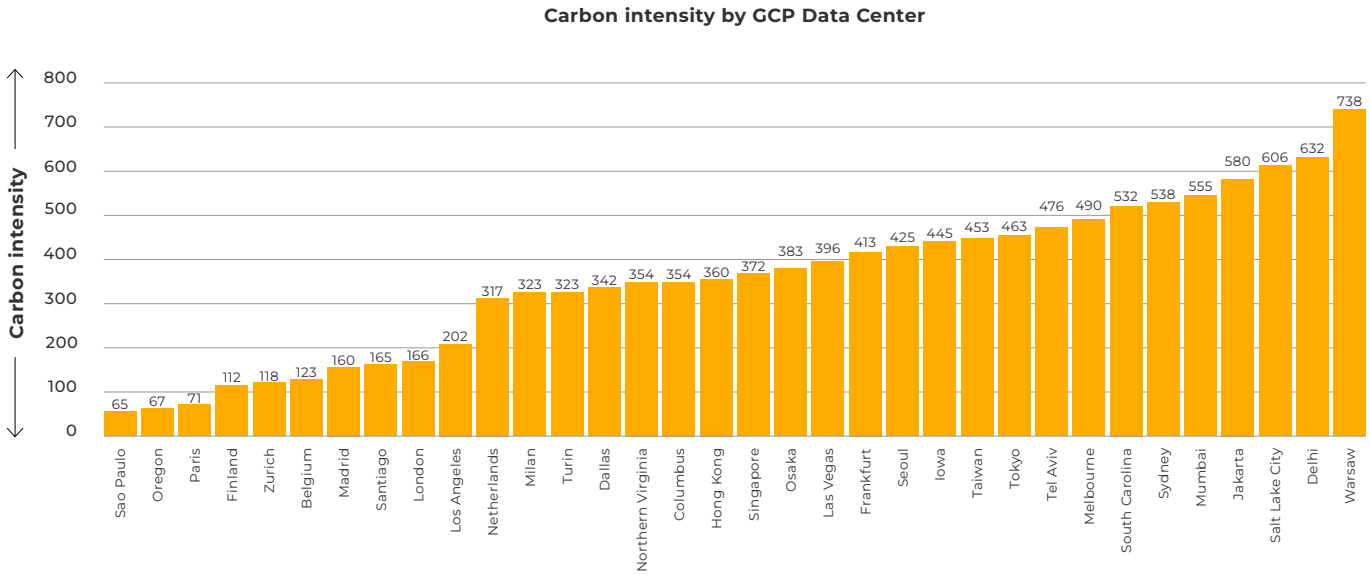


Chart 2: Bar chart illustrating the carbon intensity of data centers by location

Model Size

Larger models with more parameters require more computational resources, thus increasing energy consumption - even if not proportionally to the size of the model. For example, models with over 100 billion parameters - such as ChatGPT-4o - are more resource-intensive than smaller ones - such as Google Gemma 2 or GPT-4o mini.

Model Types, Output Length, and their Environmental Impact

Smaller LLMs (<10 billion parameters) found in Hugging Face's paper are rare use cases; most inferences are executed by bigger models. Additionally, without proper fine-tuning on specific use cases, some small models actually cannot provide outputs above 100 words. This is why we chose to calculate the impact of other model types for text generation. The result shows a higher energy consumption for inference for bigger models, however not directly proportional to the number of parameters.

Models vary in the number of words, or tokens, generated by the same query.

Additionally, models vary in the number of words, or tokens, generated for the same query. Therefore, the impact of each model varies in energy consumption, as well as the energy consumed for a given number of words.

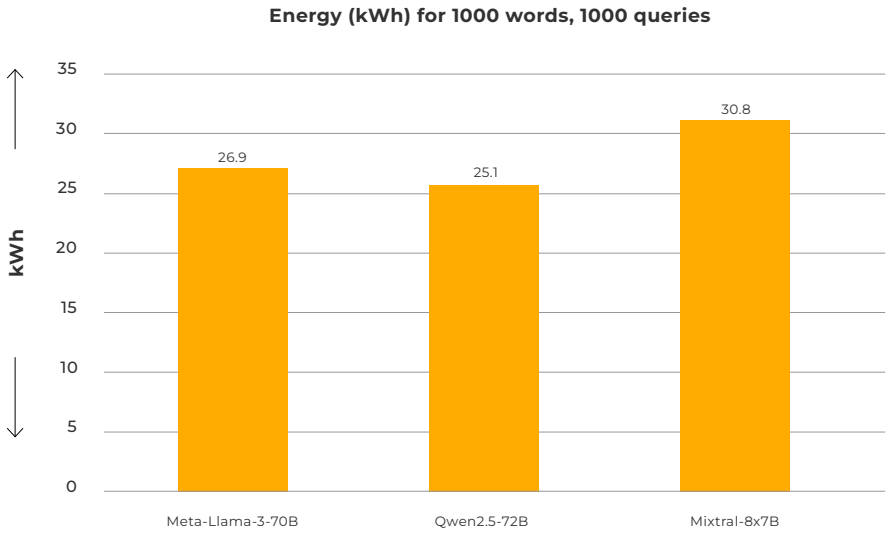


Chart 3: Bar chart illustrating energy consumed per word generated, by model

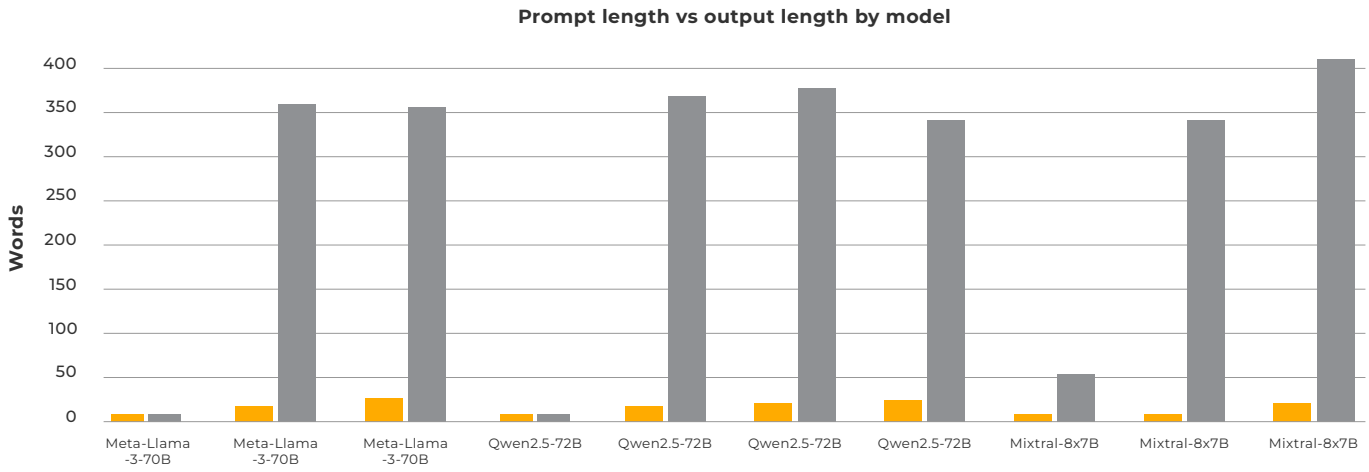


Chart 4a: Bar chart illustrating prompt length vs output length by model

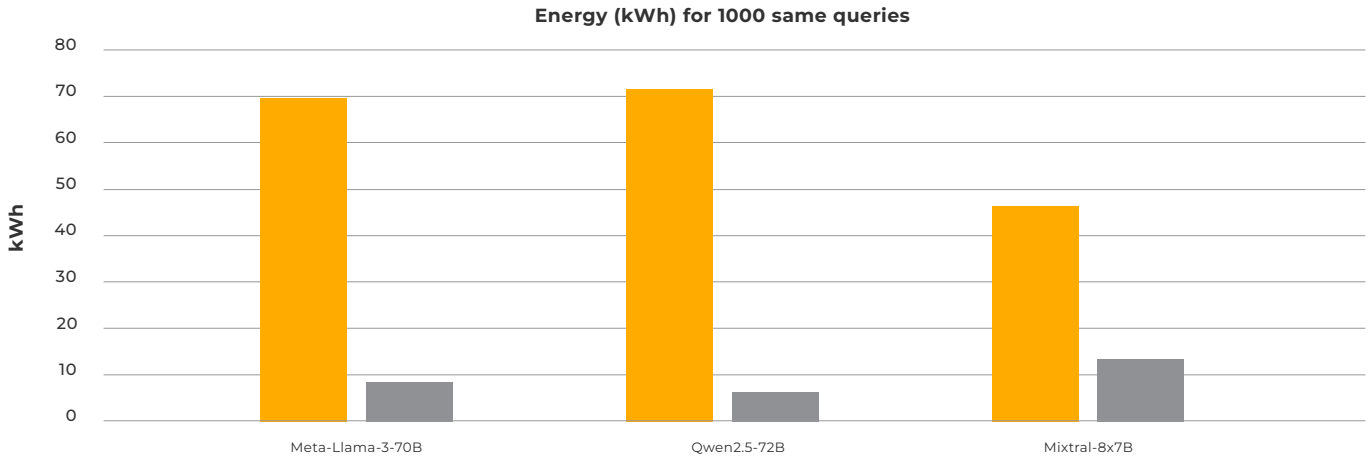


Chart 4b: Bar chart illustrating energy consumed for the same query, by model size

Best Practices and Recommendations

To help mitigate the environmental impact of their Generative AI use cases, marketers can use the following best practices, in order of influence:

Use Case Selection, Impact Assessment, and Team Upskilling

As with any technology, but most especially for generative AI, it is critically important to carefully select use cases before deploying AI tools and systems. A marketer should determine a use case is both valuable and worth the use of generative AI models instead of other tools and technologies with lower requirements.

In the use cases assessed in this report, we determined the critical impact comes when high numbers of executions are required—for example, when scaling up a strategy or deploying highly-targeted content.

The calculator we created (cf Marketing Use Case Calculations section) based on this study is designed for this use. We will continue to improve this tool based on new data points, new findings, and feedback.

Additionally, upskilling teams on how to be proficient and efficient with generative AI is also significant in limiting the number of iterations. For instance, when generating assets with AI, focusing on building the right prompt—which can also be done with an LLM—before generating images will save time and reduce the number of energy-hungry image generations.

Self-Hosted Solutions in Low-Carbon Environments

When possible, we recommend a preference of self-hosted solutions versus ready-to-use platforms.

Our next recommendation is to prefer self-hosted solutions versus ready-to-use platforms. By using open-source or private models deployed to a configurable cloud platform, a marketer would be able to:

- Select a low-carbon cloud location (ref: Chart 2)
- Fine-tune models based on brand-specific needs to reduce the number of iterations

For instance, on average, running Gen AI executions in France instead of Australia would divide by five the related carbon footprint.

This type of setup will also unlock, with most cloud platforms or tools, the use of batch executions with batch API. When compatible with a use case, asynchronous batch executions will launch inferences when and where GPU resources can be powered with lower carbon electricity, reducing pressure on the power grid.

Smaller Outputs

In many situations, LLM outputs are more verbose than is necessary. Energy consumption and runtime (therefore, embodied emissions) are strongly correlated to output size. Whenever relevant, asking for smaller outputs in prompts or directly stating the required output size will result in reduced impact.

Adding “be concise” or “limit your answer to approximately 100 words” is a good practice to mitigate the impact and cost of inference. For image generation, smaller image size (512 vs 1024) can help as well, especially in early iterations and experimentation, as the downgrade in quality may not work for final creatives.

Smaller Models

Start a new chat for unrelated questions – longer chats use more energy to proccess.

It is possible to further reduce impact by utilizing smaller models with fewer parameters when possible. Many common tasks do not require the most recent generation of models, with the highest number of parameters, to get a satisfying level of quality.

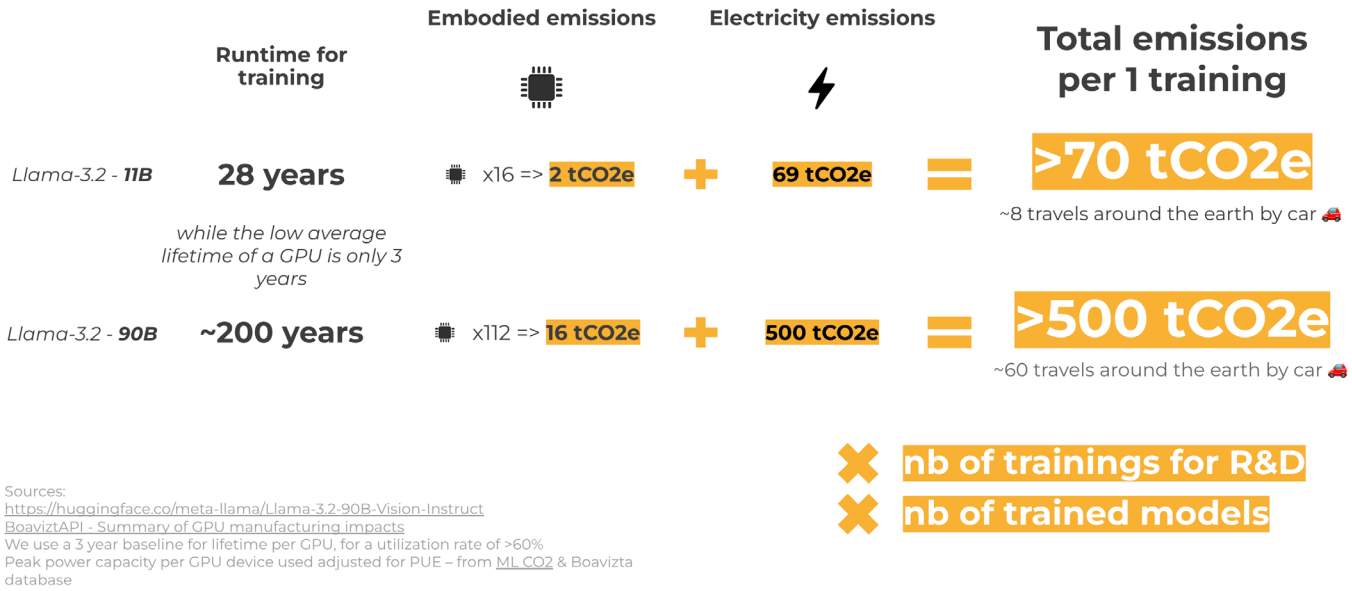
For instance, when generating a 1024 pixel image with Stable Diffusion’s models, we noticed that using the XL version required 4 times less energy than the more recent 3.5 version (ref: Source Data).

However, some tasks are not compatible with small models. It is impossible, for example, to generate a product description or any type of long text with the 1B-parameter version of the BLOOM model. Marketers need to identify the right compromise between capabilities and energy efficiency. In most cases, this should also help reduce execution costs. From a broader industry perspective, we must also avoid encouraging the continuous training of new generations of models, which is a major part of the current environmental impact of generative AI. We will publish more data on this aspect in future iterations of this report.

Estimate of Emission Repartition and **Mitigation Impact**

To highlight the importance of these best practices, we took the example of the Llama 3 open-source model, and estimated the GHG emission reductions of each mitigation opportunity. For this model, we estimated based on their public traffic data — 600M monthly active users — that a minimum of 78B annual queries would be made with this model (that is, without considering the usage of Llama deployed to a private environment for professional use). Starting from there, we estimated the emissions generated by Llama 3 for the training phase:

Training a model accounts for tons of CO2e and thousands of models are **published and trained every month**

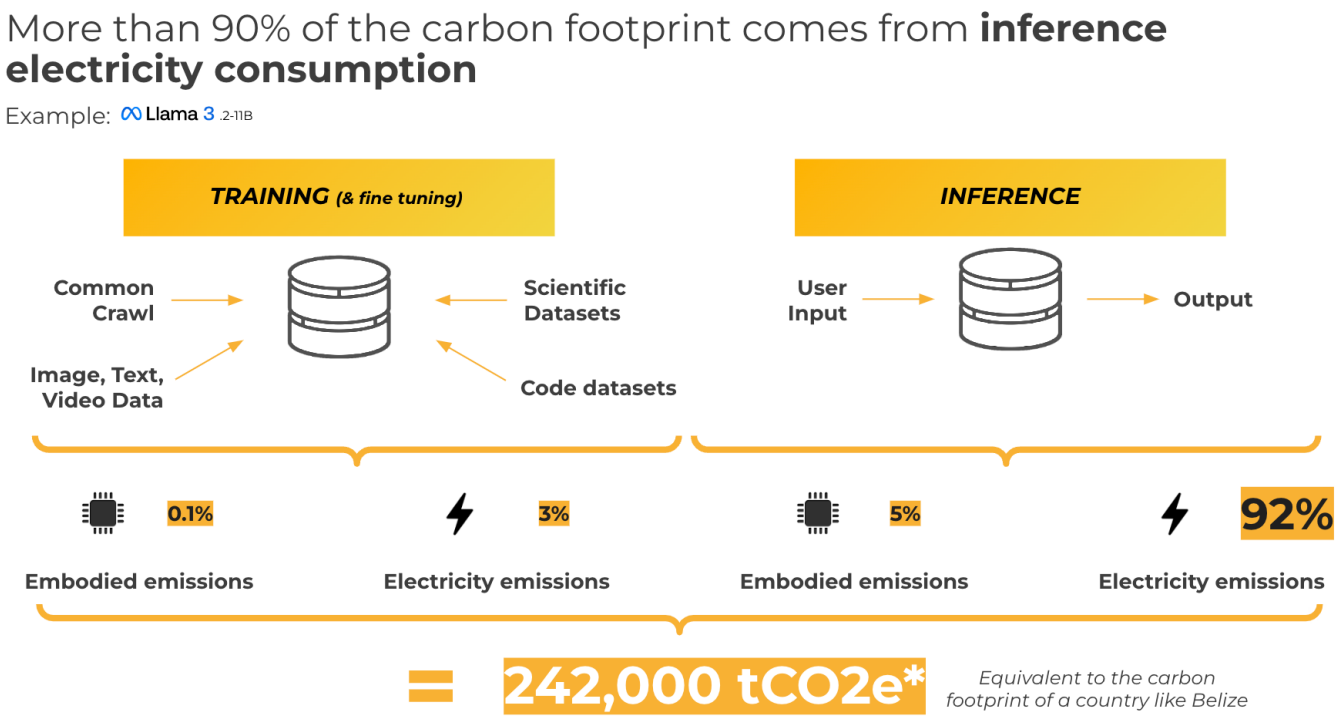


Sources:
https://huggingface.co/meta-llama/Llama-3.2-90B-Vision-Instruct
BoaviztAPI - Summary of GPU manufacturing impacts
We use a 3 year baseline for lifetime per GPU, for a utilization rate of >60%
Peak power capacity per GPU device used adjusted for PUE – from ML CO2 & Boavizta database

Sources:
12) https://ai.meta.com/blog/future-of-ai-built-with-llama/

From a broader industry perspective, it's clear that we must also avoid promoting the continuous training of new model generations, as this is a major source of generative AI's greenhouse gas emissions.

We also used our calculator to estimate emissions due to inference and were able to calculate the repartition between the various emission sources. We can see that, for such a widespread model as Llama 3, over 90% of GHG emissions come from inference electricity consumption:

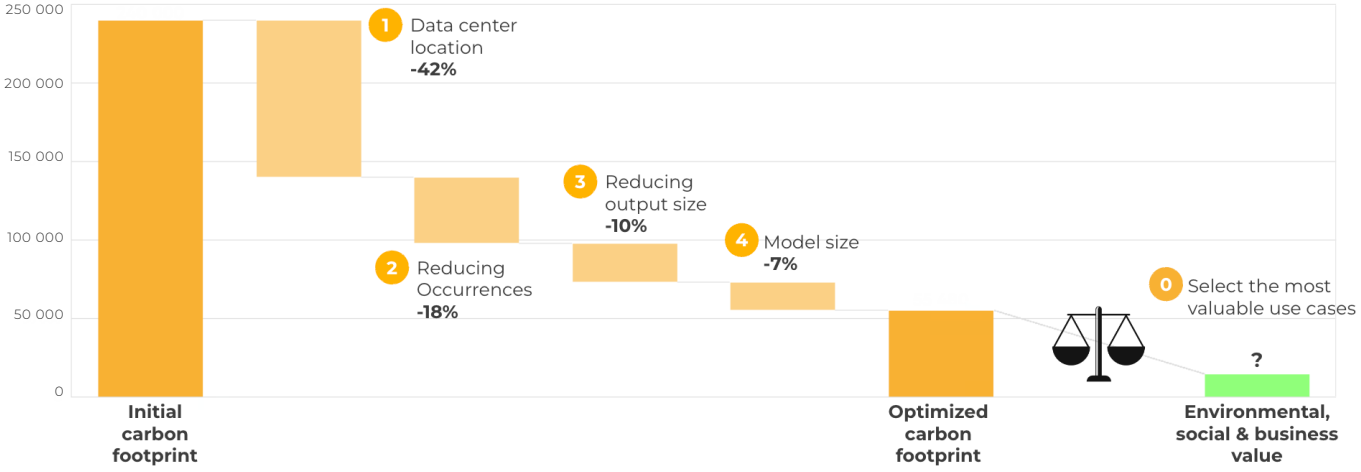


*For 1 year of model lifetime

By implementing the best practices detailed earlier, we can estimate a reduction of GHG emissions by a factor of four. This is encouraging as this reduction, achieved through a few practical and actionable steps, mean that both individual users and organizations leveraging Gen AI have a direct opportunity to reduce their environmental impact. However, it remains essential to assess that the value derived from the application of Gen AI ultimately outweighs the associated environmental impact.

But ultimately the use case value should **outweigh environmental impacts**

Tons of CO2e for initial scenario, reductions and optimized scenario



The hypotheses and calculations behind these numbers are detailed in this file.

Conclusion and Next Steps

What is clear through the course of this study is that the largest contributor of Gen AI's environmental impact comes from scaled use cases, as well as the fact that the AI industry is training many large models at the same time.

If AI models are carefully leveraged, for value-added use cases that cannot be addressed with other technologies, and best practices are followed, the impacts remain relatively small. However, most tend to overuse AI technology when there are other lower-energy, lower-carbon alternatives.

In other words, the problem does not lie in the technology but in the way we use it. Assessing the level of environmental impact of projects before deciding to launch a pilot or scaled deployment is crucial. We aim to enable this through our open-source calculator, which we will continue to develop as more data points become available across new use cases.

More importantly, measuring and displaying these impacts directly within the UI of the tools and platforms leveraging Gen AI should become standard to help drive behavior change and avoid overconsumption.

We are still in the early stages of understanding the true impact of AI technology. Its value to business cannot be overstated, but we must find the right balance between innovation and responsibility. Asking the right questions early, taking early steps to measure, and pushing for visibility from AI platforms and models are necessary for building a more sustainable industry.

More to Come

New Case Studies:
We plan to assess additional use cases, such as video generation, and release regular updates of our findings.

More Data Points:
We will conduct more experiments for other types of models to complete our data sources.

Better Coverage of Impact Sources:
As stated in our methodology, parts of our calculations rely on global averages and some sources of impact—such as electricity emissions due to idle equipment in data centers—need to be added.

Additional Impact Metrics:
We will work on measuring more metrics, such as water consumption and withdrawal (in litres or gallons) and abiotic resource depletion (in Sb equivalent).

Additional sources

https://www.weforum.org/stories/2024/07/generative-ai-energy-emissions/

https://www.epa.gov/greenvehicles/greenhouse-gas-emissions-typical-passenger-vehicle

https://www.lemonde.fr/economie/article/2024/07/02/apres-microsoft-google-voit-ses-emissions-de-co2-bondir-a-cause-de-l-ia_6246155_3234.html

https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

https://www.statista.com/statistics/1467248/greenhouse-gas-emissions-of-microsoft/]

https://www.eia.gov/energyexplained/use-of-energy/electricity-use-in-homes.php



Glossary

Generative AI: A type of artificial intelligence that uses learning models to create new content, such as images, text or music, by learning patterns from training data. It produces outputs by identifying and mimicking the structures within the data it has analyzed.

Large Language Model (LLM): A powerful AI model trained on extensive text data to understand and generate human-like language. LLMs use deep learning to recognize patterns, context, and meaning in language, allowing them to perform tasks like answering questions, generating text, and translating languages with high accuracy.

Embodied Emissions: Refers to the greenhouse gases produced throughout the lifecycle of data center equipment, including production, transportation, and disposal. These emissions represent a significant upfront environmental cost spread over the equipment's operational life.

Electricity Emissions: Refers to greenhouse gas emissions generated by the production of electricity, which vary depending on the energy source (e.g., fossil fuels, renewables).

Abiotic Resources (Refer To LCA): Non-living resources, such as minerals, metals and fossil fuels, which are extracted from the earth to be used in the production of goods and services. Within the framework of Life Cycle Assessment (LCA), these resources are taken into account to assess overall environmental impact.

Life Cycle Assessment (LCA): A method for evaluating the environmental impacts of a product or service throughout its life cycle, from raw material extraction and production to use, transport, and disposal.

Model Training: The process where an AI model learns patterns from large datasets to develop its capabilities. It relies on significant computational resources, including GPUs and CPUs, along with data storage for maintaining datasets and saving the model's learned parameters.

Inference: The phase where the trained model is used to generate predictions or responses based on user inputs. This process involves utilizing the model's learned parameters to produce outputs, often relying on computational resources like GPUs or CPUs for efficient execution.

Data Storage: The process of organizing, securing, and managing data in systems or devices, essential for training AI models and their deployment in real-world applications.

Model Training Iterations: The repetitive cycles in which the algorithm adjusts its internal parameters according to the errors observed in the training data, with the aim of improving model performance.

Rebound Effects: A phenomenon in which certain environmental benefits from more resource-efficient management or technical developments are significantly reduced by an increase in consumption or a change in use.

AI Ecosystem: Refers to the set of technologies, tools, infrastructures and players that interact to create, deploy and manage AI systems, including models, data, platforms and users.

Model Fine-tuning: Process of adjusting its parameters on a specific dataset after general training, in order to improve its performance on particular tasks or in a specific domain.

Model Versioning: Tracking and managing the different versions of an AI model as it evolves, keeping track of modifications and ensuring optimal management of updates.

Parameter Count (In Models): The number of internal variables a model uses to make predictions. Higher parameter counts typically indicate more complex models capable of handling intricate data relationships.

Quantization: A model compression technique for AImodels, which reduces the precision of numerical parameters, decreasing model sizewhile maintaining performance.

Inpainting: An AI-based technique for restoring or completing missing parts of an image by using the surrounding context as a guide.

Runtime: The total execution time of AI models; in the case of this whitepaper, measured in GPU hours, representing the computational effort needed to run the models.

GPU Parallelization: A technique that uses the computing power of multiple graphics processing units (GPUs) to execute tasks simultaneously, speeding up complex calculations. When several GPUs are used, they can work in parallel to quickly process large amounts of data. However, parallelization across multiple GPUs can pose problems, including making it difficult to manage and visualize the exact number of units used to process a specific request. This can make it difficult to monitor resources and optimize GPU utilization, especially when tasks are not perfectly distributed between different processors.

Graphics Processing Units (GPUs): Processors specialized in parallel computing, used primarily for intensive machine learning calculations, in particular for training AI models.

Central Processing Units (CPUs): The main processors in a computer, responsible for executing program instructions, although they are generally less efficient than GPUs for machine learning tasks.

Random Access Memory (RAM): Type of memory used by a computer to store temporary data and instructions needed to run programs, playing a key role in the performance of AI models in terms of fast information processing.

Model Parameter Optimization: Consists in adjusting internal parameter values to maximize model performance, often using optimization algorithms such as gradient descent.

Power Usage Effectiveness (PUE): A measure of a data center's energy efficiency, calculated as the ratio of total energy consumption (including cooling and lighting) to the energy used by IT equipment (servers, storage). A PUE closer to 1 indicates optimal efficiency.

Output Type (Text, Img, Classif): The format or nature of the results generated by an artificial intelligence model, which may be in the form of text, image or classification, depending on the task in hand.

Token Count: The number of text elements (such as words, characters or sub-words) processed by a model during training or inference.



Contributors
& Authors

As a part of The Brandtech Group, **fifty-five** is a data company that helps brands collect, analyze and activate their data to improve their marketing and sustainability strategies. Headquartered in Paris, fifty-five operates across 3 time zones from our 10 offices. The data company was named by Forbes as one of the world's best management consulting firms in 2022, owing to its unique approach that blends consulting, ESG and technology expertise.

Individual authors & contributors:



Ludovic Moulard,
Sustainability Director



Romain Warlop
Data-science Director



Tiyab Konlambigue
Cloud Practices Director



Agathe Gourhannic
Consulting Manager



Sharon Kalifa
Project Lead



Alexandre Helme
Consultant



Domitille Choutet
Consultant



Nicolas Vaudran
Managing Director

Brandtech Group contributors:



Rebecca Sykes
Partner



Hugo Loriot
Partner



Sharanya Sathyaprakash
Gen AI Operations Manager



Amy Arnell
Gen AI Business Manager

Scope3 is a technology company transforming how organizations optimize their AI operations. Through real-time monitoring and optimization, Scope3 helps enterprises reduce unnecessary energy consumption while maintaining peak AI performance.

Led by pioneers who scaled algorithmic systems in digital advertising to create a $100B industry, Scope3's leadership team brings proven expertise in optimizing complex machine learning operations for both economic and environmental outcomes. The company enables organizations to harness AI's transformative power while minimizing their environmental impact, demonstrating that business performance and sustainability can work together. The Brandtech Group is among Scope3's minority investors, reflecting our shared vision for sustainable advertising technology.

Individual contributors:



Brian O'Kelley
Co-founder & CEO



Lillie Ratliff
Head of Strategic Initiatives



Dave Freifeld
Data Science Lead, AI



Detailed Calculation Methodology by **Emission Source**

Environmental Impact **Calculation Methodology for GenAI**

The aim of the calculation methodology is to measure the environmental impact of the GenAI ecosystem as a whole, as well as for specific use cases such as Pencil AI for the creative generation.

Our methodology primarily focuses on the emissions generated by data centers, as they are the most significant contributors to the environmental impact of GenAI. Data centers are necessary for training, storage and inferences. That's why they are identified as the main contributors to GenAI emissions. While networks and user devices also contribute to emissions, their impact is relatively negligible compared to data centers and is therefore excluded from this analysis. This study is particularly aimed at marketers and focuses on the emissions relevant to their use of GenAI.



Example of AI data center provided by Nvidia dedicated for model training and inference.

Two types of emissions are taken into account:

01.

Electricity emissions

Energy used to run the equipment essential to the operation of GenAI's models during the various phases:

GPU (Graphics processing unit):
GPUs are processors specialized in parallel computing, used primarily for intensive machine learning calculations, in particular for training AI models.

Source:
https://www.pny.com/productimages/C38B2FBB-2DA8-4454-A901-AEE4008578E0/images/A100PCIe_3QTR-TopLeft-Dual_NVLinks_Black_Alt.png



GPU NVIDIA A100 80GB, commonly used for these tasks due to their high performance and large memory capacities.

CPU (Central Processing Unit):
CPUs oversee the training process, managing tasks like data loading and synchronization across multiple GPUs.

Source:
https://i.ebayimg.com/images/g/0YwAAOSweyhiV-7o/s-l1600.webp



Choosing the right CPU is essential for optimal GPU performance. For example, the NVIDIA DGX A100 pairs a 64-core AMD EPYC 7742 CPU, showcasing the benefits of high-core-count processors.

RAM (Random Access Memory):
RAM that temporarily stores data being processed for rapid access.

Ample RAM is crucial for data-intensive tasks and GPU performance. The DGX A100 features 1 TB of system memory, emphasizing the need for sufficient RAM.

Source:
https://upload.wikimedia.org/wikipedia/commons/9/9a/NetApp_ONTAP_AI.jpg



Example of a data center unit of DGX A100, embodying the NVIDIA DGX A100, a AMD EPYC 7742 CPU and 1 TB of RAM.

02.

Embodied Emissions

Embodied emissions refer to the greenhouse gases produced throughout the lifecycle of data center equipment, including production, transportation, and disposal. These emissions represent a significant upfront environmental cost spread over the equipment's operational life.

For instance, NVIDIA, an USA based company that's the key supplier of GPUs data center, outsources chip manufacturing primarily to Taiwan Semiconductor Manufacturing Company (TSMC) in Taiwan, which produces most of NVIDIA's advanced chips using cutting-edge processes. NVIDIA also partners with Samsung Electronics in South Korea for specific GPU models.

The global distribution of these facilities adds to embodied emissions, as semiconductor fabrication is energy-intensive, and transporting components across continents further increases the environmental footprint of data center operations.

The methodology is based on two key elements of the GenAI operation:

1. **Model training:** The initial phase where the model learns from large datasets to develop its capabilities. This phase relies on computational resources like GPUs and CPUs for processing, while data storage is crucial for maintaining datasets and saving model artifacts, such as checkpoints and learned parameters.

2. **Inference:** The phase where the trained model is used to generate predictions or responses based on user inputs. This process involves utilizing the model's learned parameters to produce outputs, often relying on computational resources like GPUs or CPUs for efficient execution.

In a later phase, the methodology will also integrate the water consumption used to cool servers in data centers and abiotic resources which refers to the use of non-renewable metals and minerals for hardware infrastructure.

Prerequisites for Evaluating **Environmental Impact**

To estimate the environmental impact of GenAI, the three following elements are necessary :

1. Runtime measurement:

- The runtime, expressed in GPU hours, measures the total processing time of the computing resources (GPU, CPU and RAM) required to run the models.

- Focusing on GPUs is appropriate because they are the primary contributors to energy consumption during AI model training and inference. For instance, the GPUs that trained GPT-3 (the precursor to ChatGPT 3.5) consumed approximately 1,300 megawatt-hours of electricity, equivalent to the monthly usage of about 1,450 average U.S. households.

  Source: https://news.mit.edu/2023/new-tools-available-reduce-energy-that-ai-models devour-1005

- *Formula:*
  *Runtime (in GPU hours) = time spent during which computing resources are processing * number of GPUs*

2. Key variables (which influence emissions):

- **Model Size:** Larger models with more parameters require more computational resources, leading to increased energy consumption.

- **Output Type:** Different tasks (e.g., text generation, image synthesis) have varying computational demands. For example, generative AI models like Stable Diffusion, used for text-to-image generation, have specific energy consumption profiles during training and inference.

- **Output Size:** The complexity and size of the generated outputs (e.g., number of tokens in text, resolution of images) affect the computational workload and, consequently, energy usage.

3. Environmental impact of previous generations of models or Emissions across model generations

Evaluating the environmental footprint of AI models requires considering the entire development lifecycle, including all versions and iterations. While improvements in newer versions may enhance efficiency, the cumulative impact of training multiple versions can be substantial. For instance, the BLOOM model, a large multilingual AI model, was estimated to generate over 50 metric tons of CO_2 over its lifecycle, comparable to 80 round-trip flights between London and New York in a study made by Sasha Luccioni AI & Climate Lead at Hugging Face.

Source: https://time.com/7012753/sasha-luccioni

Environmental Impact of **Model Training**

1. Calculating Electricity Emissions:

Formula:
Electricity emissions = RunTime × Average GPU power × Average data center carbon intensity.

Parameters:

- RunTime: Total number of GPU hours required for training.
- Average GPU power: Average consumption in kW of a GPU in use.
- Average Data Center Carbon Intensity: Average amount of CO_2 emitted per kWh, variable according to the energy source used (coal, nuclear, etc.).

2. Calculation of Embodied Emissions

We use Boavizta's impact factors, which detail the emissions linked to the manufacture and life cycle of GPUs.

Formula:
Embodied Emissions = (RunTime ÷ GPU lifetime) × GPU emission factor

Parameters:

- RunTime: Total number of GPU hours required for training.
- GPU lifetime: estimated at 3 years, with a utilization rate of over 60%.
- GPU Emission Factor*: Value provided by Boavizta, which includes manufacturing and transport emissions for components such as the die, memory and PCB. It is important to note that the GPU Emission Factor also takes into account the usage of CPUs and RAM, not just GPUs.

***Boavizta Impact Factors:** Details of emissions (manufacturing and/or transport) for GPU components:

→ **GPU die:** The heart of the GPU, the main chip.
→ **GPU memory:** GPU-specific memory modules.
→ **GPU PCB (Printed Circuit Board):** Base on which the die, memory and other components are mounted.

3. Dimensions and Evaluation Parameters:

To assess the impact of AI models on resources and energy, several dimensions and parameters need to be taken into account, including task type, model size and popularity.

- **Task type:** Different tasks (text, image) may require different computing power.
- **Model size:** Large models (e.g. 80-90 billion parameters) consume more energy.
- **Model popularity:** Models with high usage (>100k downloads) have a higher impact due to the high demand on resources.

4. Evaluation of Data Storage for Training Data:

In this section, we look at the environmental impact of storing the datasets used to train models. Data storage generally takes less energy than training, but can still have a significant impact over time, if data volumes are large and kept for long periods. We examine this impact by comparing different models.

A. Data and Datasets

For each model, the data preparation process follows several stages before arriving at the final dataset used for training. These steps include :

- **Raw Dataset:** Initial data collected, often in very large volumes.
- **Prepared Dataset:** Raw data is then filtered and processed to eliminate irrelevant information.
- **Filtered Training Dataset:** Final version of the dataset, ready for training. After additional filters, the final volume can be significantly reduced.

For each model, the total volume of data stored is the result of the sum of these different steps, which give an idea of the data used and stored in the data centers.

B. Total Stored Data Calculation

Formula: Total Stored Data (GB) = (Raw Dataset (TB) + Quantized Dataset (TB) + Training Data (TB)) × 2^10

This calculation takes into account the various stages of data processing to capture the total footprint of stored data over time.

C. Coefficient NegaOctet (EF - Emission Factor)

The NegaOctet EF factor is an emission coefficient in gCO_2/GB for 10 years of storage. It represents the amount of CO_2 emitted per gigabyte of data stored for a decade in a data center. This factor depends on the data center's infrastructure and energy efficiency standards.

For example, for GPT-3, the NegaOctet EF is 25.3 gCO_2/GB for 10 years' storage. This means that one gigabyte of data stored for 10 years will generate around 25.3 grams of CO_2.

D. Calculation of Total Emissions of Data Storage over 10 years

Total CO_2 emissions of data storage **over 10 years** are calculated with the parameters explained above and using the following formula:

Formula: Total Emissions (tCO_2) = (Total Stored Data × Coefficient NegaOctet EF) / 10^6

This calculation converts gCO_2 emissions into tonnes of CO_2 equivalent (tCO_2) for 10 years of storage.

Environmental Impact of **Inferences**

Inference corresponds to the operations performed by the model to generate results based on user requests.

1. Limitations of the Luccioni and Strubell Experiment

Luccioni and Strubell's work provides an interesting basis on the environmental impact of inferences, but has important limitations for our study:

- **Model size**: Their analysis is limited to small models of less than 10 billion parameters, which is insufficient for the more complex models we wish to study (such as those exceeding 100 billion parameters).

- **Prompt complexity**: The prompts used in their study are simple and small. We aim to explore more complex and varied queries, to better understand the impact of task complexity on energy consumption.

- **Significant variance:** The standard deviations observed in their results, particularly for image generation, are high and limit the reliability of the data. This high standard deviation indicates that energy consumption is highly dependent on the model, sample size and tasks performed.

For these reasons, we set out to extend their work with more advanced experiments, incorporating larger models and more diverse task types.

2. Methodology of Our Experiment

Our experiment aims to obtain more precise data on the environmental impact of inferences for larger AI models and varied tasks. Here is our experimental protocol:

- **Deployment of Open Source Models**: We download and deploy open source models on our own cloud infrastructure. We selected models with a high number of parameters to be as close as possible to real life use cases.

- **Use of CodeCarbon:** We use the CodeCarbon library to monitor resource energy consumption (CPU, GPU, RAM) and inference runtime in real time.

- **Multiple Tests on Different Models:** We perform text-to-text and text-to-image(including inpainting) inferences with different prompts, to observe various lengths of output and size of the generated image on runtime and energy consumption. See table below for the list of used models.

- **Measuring Energy per 1,000 Requests:** At the end of each test run, we record energy consumption for 1,000 queries, allowing us to standardize results by model size.

For each experiment (one task, one prompt, one image size) we repeated the generation multiple times to limit bias.

TASK	MODEL	NUMBER OF PARAMETERS (B)
text-to-text	Llama-3-70B	70
text-to-image	Stable diffusion 3.5 large	8
text-to-image	Stable diffusion XL	6.6
text-to-image	Stable diffusion large turbo	8
inpainting	Stable diffusion 2 inpainting	0.86
inpainting	Stable diffusion XL inpainting	6.6
inpainting	Stable diffusion inpainting	0.86
inpainting	Kandinsky 2.2 Inpainting	3.6

List of models used for the experiments

3. Formula for Estimating Usage Emissions

Each type of use (text-to-text, text-to-image, text classification) requires a separate model and produces different outputs. We use a common formula to estimate emissions, while adapting certain parameters to suit each use case:

Formula: Usage emissions = RunTime × Data Center PUE (Power Usage Effectiveness) × Average Carbon Intensity

PUE (Power Usage Effectiveness): PUE is an indicator of data center energy efficiency. It represents the ratio between a data center's total energy consumption (including cooling and lighting) and the energy directly used by IT equipment (servers, storage). An ideal PUE is close to 1, meaning that almost all energy is used for server operation, with minimal losses. The higher the PUE, the more energy the data center consumes.

4. Key Variables Influencing Usage Emissions

Model characteristics play a crucial role in the environmental impact of inferences. Specific variables that influence emissions include:

- **Number of parameters (Params):** Models with more parameters consume more resources.
- **Quantization (Quant):** The use or non-use of quantization (reducing numerical precision) can influence energy efficiency.
- **Parameter accuracy (Accuracy):** Numerical precision (e.g. float16, bfloat16) also affects energy consumption.
- **Model architecture (Architecture):** Some models use expert architectures, which can optimize performance for the task in hand.

A more precise formula would be :

Usage emission (output size, input complexity) = f(params, quant, precision, architecture) × Data Center PUE × Average Carbon Intensity

where f is an unknown, but certainly non-linear function. To learn this function, we'll need lots of data to train a model.

5. Experimental Protocol by Use Case

We apply specific experimental protocols depending on the type of task:

Text-to-Text:

- **Parameters Measured:** Number of tokens generated and length of the prompt in input.
- **Protocol:** Testing different models with text queries of increasing complexity to observe energy consumption.

Text-to-Image:

- **Parameters Measured:** Size in pixels of the generated image and complexity of the prompt.
- **Protocol:** Testing of different models with prompts varying from simple descriptions to complex and precise requests.

Image classification:

- **Parameters Measured:** Input image size (always one output unit).
- **Protocol:** The prompt remains constant, but the input images vary to observe the impact of image size on energy consumption.

Data collected for each use case includes: model name, number of parameters, architecture type, quantization, precision, number of tokens (or pixels), energy consumed, and output size.

6. Next Steps

To complete our study, next steps include:

- **Model Optimization:** Test different optimizations (such as quantization) to measure their impact on energy performance.
- **Expanding the Model Sample:** Test more models to obtain more representative results.
- **New Use Cases:** Test other use cases, such as inpainting (image restoration), to enrich our analysis.