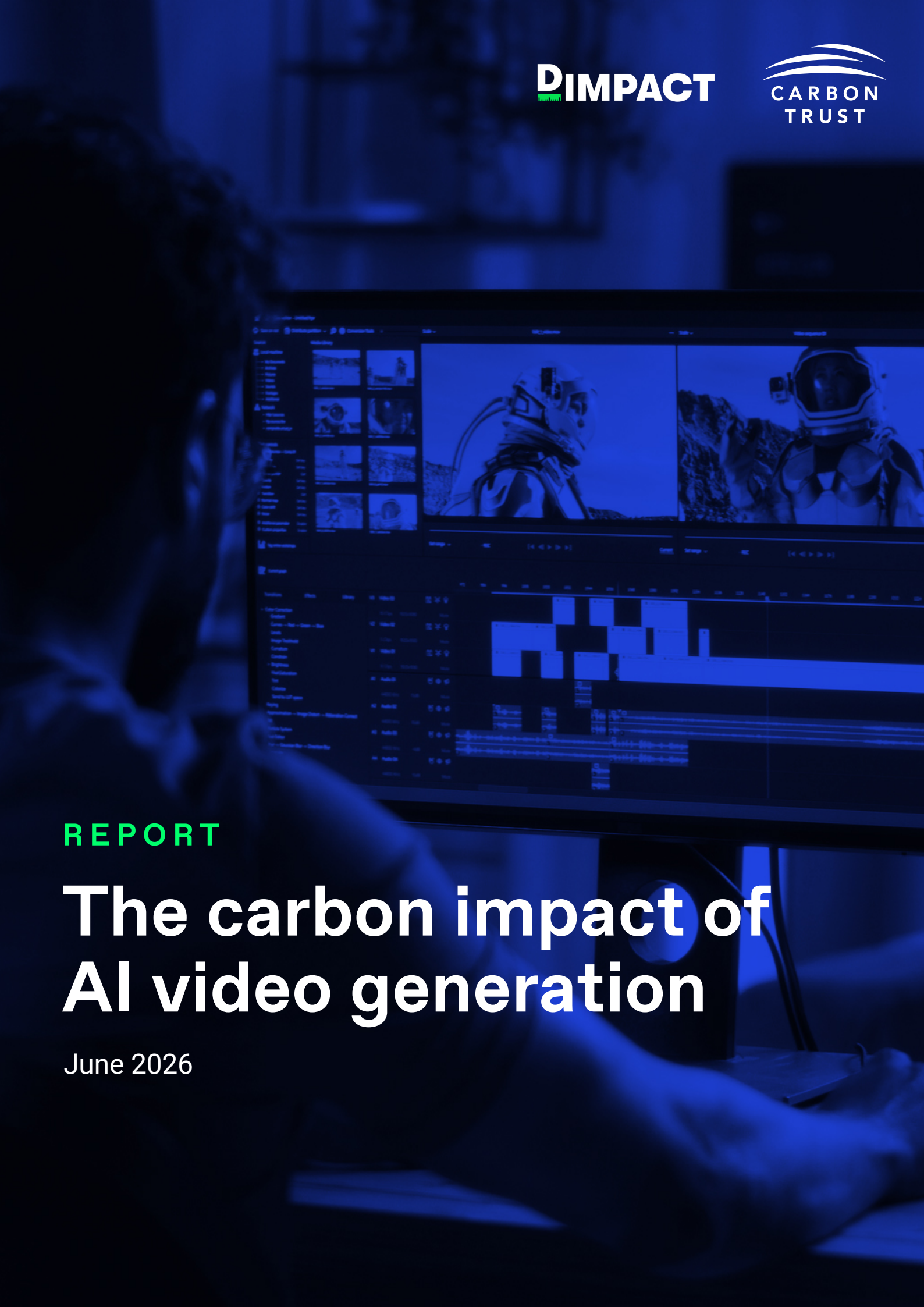


REPORT

The carbon impact of AI video generation

June 2026

About the report

This report provides our current best view of the carbon impacts of AI video generation. We recognise that this is a fast-moving and rapidly changing space and our intention is to bring clarity to how AI video generation is being used in media today and how its carbon impact is currently measured, estimated and understood.

Acknowledgments

The Carbon Trust wrote this report based on an impartial analysis of primary and secondary sources, including direct engagement with stakeholders across the digital media and broader ICT industries.

The Carbon Trust would like to thank everyone that has contributed their time and expertise during the preparation and completion of this report. Special thanks goes to:

- **The Expert Working Group**, which provided valuable feedback on the content and analysis through workshop sessions and draft reviews, comprised:
    - **Mark Butcher, Posetiv Cloud**
    - **Boris Gamazaychikov, Sustainable AI Group**
    - **Simon Hinterholzer, Borderstep Institut**
    - **Dr. Sasha Luccioni, Sustainable AI Group**
    - **Sam Read, Sustainable Entertainment Alliance**
    - **Dr. Daniel Schien, University of Bristol**
    - **Dr. Iuna Tsyrulneva, Earth Observatory of Singapore**
- We thank in particular **Dr Daniel Schien** who kindly authored our 'explainer' on how generative AI works (and further detail in Appendix 4).
- **Members of the DIMPACT group** who provided input on current generative AI use cases and reviews of the report content. Specific acknowledgement goes to the **BBC, ITV, Netflix, Sky and Spotify** for their closer involvement in providing feedback.
- Other stakeholders consulted during the process, including:
    - **BAFTA Albert**: Provided insight on current industry approaches to impact measurement, including relevant calculation methodologies and case studies.
    - **The Green Software Foundation Standards Working Group**: Reviewed the technical methodology, aligning on the high-level approach and providing feedback for refinement.
    - **Professor Ivana Drobnjak, University College London and UNESCO Chair in AI member,** provided helpful feedback on the technical draft

This report was funded by DIMPACT members (BBC, Spotify, Sky and Netflix). For the avoidance of doubt, this report expresses the independent views of the authors.

About the Carbon Trust

Our mission is to accelerate the transition to a decarbonised future.

We have been climate pioneers for more than 20 years, partnering with leading businesses, governments and financial institutions globally. From strategic planning and target setting to activation and communication, we support organisations in turning climate ambition into impact.

We are a global network of over 350 experts with offices in the UK, the Netherlands, Germany, South Africa, Singapore and Mexico. To date, we have helped set more than 200 science-based targets and guided over 3,000 organisations in 70 countries on their route to Net Zero.

About DIMPACT

DIMPACT is a ‘think and do’ coalition working to align industry changemakers and policymakers around meaningful, science-based solutions that reduce the environmental impacts of serving digital media products. DIMPACT convenes and unites industry leaders and changemakers, sharing research, resources, and best practices to catalyse collaboration and accelerate action.

The Carbon Trust project team:

Matt Anderson

Technical Advisor and Lead Author

Matt.Anderson@carbontrust.com

Bob Burgoyne

Project Director

Bob.Burgoyne@carbontrust.com

Mike Hopkins

Project Manager

Mike.Hopkins@carbontrust.com

Wei Yang Lee

Senior Associate

WeiYang.Lee@carbontrust.com

Fedra Bartfai

Senior Analyst

Fedra.Bartfai@carbontrust.com

Iarina Corniciuc

Project Coordinator

Iarina.Corniciuc@carbontrust.com

Martin Barrow

Quality Reviewer - Case Study Analysis

Martin.Barrow@carbontrust.com

Contents

Abbreviations

AWS	Amazon Web Services
AI	Artificial intelligence
API	Application programming interface
CO_2	Carbon dioxide
CO_2e	Carbon dioxide–equivalent emissions
EU	European Union
GPU	Graphics processing unit
GSF	Green Software Foundation
GHG	Greenhouse gas
IEA	International Energy Agency
ISO	International Organisation for Standardisation
ITU	International Telecommunications Union
LCA	Lifecycle analysis
LLMs	Large Language Models
MEP	Mechanical, electrical and plumbing
ML	Machine learning
MW	Megawatt
MWh	Megawatt hour
MP	Megapixel
PCF	Product carbon footprint
PCR	Product category rules
PUE	Power usage effectiveness
SCI	Software carbon intensity
TW	Terawatt
TWh	Terawatt-hour
VFX	Visual effects

Select glossary

Model developer	Organisation which produces a generative AI model – for example OpenAI
Data centre operator	Organisation that operates the data centres used for inference – for example Google
Model provider	Organisation which provides model access to the model user. This can be either be the model developer - e.g. directly accessing chatgpt.com, or a third-party intermediary - e.g. Leonardo, which uses APIs to give access to models from other developers such as Google Veo.
Model user	An end-user of a generative AI service – for example Netflix
Training	The process whereby a generative AI model is built by a model developer
Inference	The use of a generative AI model by a model user

1. Executive summary

The scale and speed at which artificial intelligence (AI) technologies are evolving, being deployed and adopted are considerable, with wide-ranging impacts across the data centre sector, the global economy and the environment. Current estimates by the International Energy Agency predict that by 2030, global data centre electricity consumption will double to around 945 TWh (IEA, 2025a).

However, this speed of deployment is now hitting real-world constraints, primarily around access to power, but also availability of IT equipment, backup power technology and cooling systems. Some prominent AI technology companies are side-stepping the grid connection queue bottleneck through the build-out of new combined cycle gas turbine generation capacity, while others are signing long-term power supply agreements with nuclear power stations – or through a combination of renewable sources and energy storage.

Given these dynamics, concern has been raised among the public, generative AI users, within the research community and by policymakers about the environmental impacts that will result from AI infrastructure build-out and use

Furthermore, generative AI technologies are becoming increasingly accessible and capable, leading to innovations like AI video generation which uses specifically developed AI models to generate video from text, image or video inputs. AI video generation is more computationally intensive than other simpler generative tasks, such as text generation, yet there is a general lack of understanding of the specific energy and carbon impacts as concrete figures are hard to come by. Improving understanding of the carbon impact of AI video generation is of interest to its users, including media companies and production professionals, so it can be used responsibly, and its carbon impact better managed.

The transparency gap leaves users with outdated, incomplete and inconsistent information about the carbon impact of generative AI technologies

Our research has found that transparency around the carbon impact of generative AI technologies from model developers and providers is not sufficient to help its users make informed decisions. Given the speed that models are developed and published, it's important that users have access to up-to-date information, which reflects the latest models and capabilities.

In part, the reluctance by model developers and providers to share this data may come from the risk of data being misused or misinterpreted, particularly as there is no established consensus on how to measure the lifecycle carbon impact of AI for specific uses, such as AI video generation.

In its absence, the scientific community is attempting to fill the gap, and recent studies have explored the energy usage and carbon impacts of different generative AI tasks, including video generation. These reveal, for example, that AI video generation likely requires two orders of magnitude more energy to generate a short five-second video than to respond to a simple text query.

The speed of innovation and complexity of video generation demonstrate the need for a consistent measurement approach

Our review of existing literature concludes that measurement approaches are not consistent when measuring the energy and carbon impact of generative AI, in particular the treatment of impacts from model training and embodied carbon that are attributable to the generated video. We also find that for AI video generation, the resulting carbon impact is highly dependent on a number of key factors:

- training emissions and how they are allocated
- selected video parameters such as resolution, duration and number of denoising steps
- the carbon intensity of electricity used
- the underlying model architecture and infrastructure

To demonstrate the sensitivity to these various parameters and assumptions, we performed analysis building on a recent study that evaluated the energy used by the WAN2.1-T2V-1.3B video generation model (Delavande, Pierrard, & Luccioni, 2025).

We estimate that a 5.4-second video at 15 frames per second and 50 denoising steps has a lifecycle carbon impact per video of around 50 to 100 gCO_2e at 720p resolution, when generated on the US electrical grid. The variation in this range can be mostly attributed to size and allocation of emissions from model training, further emphasising the need for a consistent approach.

Real-world applications are still experimental and varied, with some encouraging indications of use cases which could plausibly support energy efficiency improvements, but further exploration is needed

In real-world applications, AI video generation is being tested and used by production teams in various ways, including for visual effects (VFX) work, without clarity on the associated carbon impact. To provide some context for where and how video generation is being used, we assessed the lifecycle carbon impact of a real-world case study which used video generation to adapt background imagery in a scene for a streaming series.

In this specific case, it was found that by using AI video generation instead of traditional VFX rendering, the amount of electricity required by locally operated VFX workstations could be reduced, resulting in a plausibly lower VFX-related carbon impact for the final scene, with a comparable visual result. However, this should be not interpreted as a general comparison between approaches. This case study was for a specific, bounded production example and the resulting impact depends on several factors including the process boundary, data quality, allocation of model training emissions and resulting uncertainty.

We provide recommendations for how digital media companies can begin to monitor and manage these carbon impacts but, given the rapid evolution in the space and the wide range of uncertainty around some of the key input parameters, we recommend further exploration of this topic. This would support the establishment of best practice guidelines for the sustainable use of AI video generation in production and creative processes, which considers impacts from a holistic perspective.

The call to action: Bridging the transparency gap through alignment on clear disclosure guidance

We recognise that the disclosure of emissions is a sensitive issue and companies may hesitate to share emissions figures without a consistent measurement approach. This report makes a start on addressing this challenge. It provides a methodological foundation to build on and test with stakeholders and we invite continued engagement in developing a methodology to improve the transparency around the carbon impacts of generative AI.

We strongly recommend and welcome constructive dialogue between model developers, data centre operators, model providers, LCA practitioners, the scientific community and users to develop consistent lifecycle assessment methods in the form of product category rules (PCR) for AI video generation.

With consistent rules in place:

- model providers can confidently disclose emissions associated with using their services
- all parties can identify and implement opportunities to reduce these emissions
- users will be able to make more informed decisions by considering the carbon impact of use

Taken together, these actions will help build a more transparent and consistent understanding of the carbon impacts of AI video generation, enabling organisations to manage them more effectively.

2. Introduction

What is the carbon impact of AI video generation?

This question is increasingly being asked by the public, as well as professionals in media production and academia who are seeking to better understand the environmental consequences of using generative AI. At present the question is difficult to answer.

To illustrate this, take Coca-Cola's recent one-minute 2025 holiday ad which, according to its creators, required 70,000 AI-generated video clips. Based on this figure, one recent estimate put the electricity and carbon impact of this one ad at 70 MWh of electricity and 27 tonnes of carbon dioxide equivalent (CO_2e) (Ketan Joshi, 2026), approximately the annual electricity consumption of seven American households (IEA, 2025c). With a few tweaks to assumptions related to the electricity used per prompt, as well as the duration and resolution of each generated video, we could reasonably estimate an impact of 2.6 tonnes of CO_2e.[2] This range spans an order of magnitude, demonstrating the lack of consistent measurement methods and resulting uncertainty surrounding this topic.

Leading AI video generation developers don't currently disclose the energy or emissions required to train their video generation models or report these metrics related to inference and video generation[3]. We understand that disclosing this data is not trivial. From a technical perspective, this is in part due to a lack of a consistent approach to measuring these emissions. For example, what method should a video generation model provider use to report carbon figures of their models if they lack access to energy data from their use within data centres? And for model developers, providers and data centre operators, disclosing emissions figures comes with some risk, such as the potential for unfair comparisons to competitors who may be using different emissions estimation methodologies, or the potential for figures to be taken out of context and sensationalised in the media.

Our objective is to bring clarity to the discussion and to invite collaboration on a path forward

This report intends to promote balanced discussion and consistency of methodology to help address these concerns. It also provides some practical guidance for those using generative AI for video generation; while we recognise that we don't hold all the answers and that this technology is rapidly evolving, we think we know enough now to help inform how we use generative AI video in an environmentally responsible way. To our knowledge, there has not been a comprehensive study which evaluates the lifecycle carbon impact for video generation which creates a production-ready video asset (i.e. a video asset used in a final piece of content such as a series, film or advertisement). This report represents a first effort to fill this gap.

The key objectives of this report are to:

- Demystify existing research into the emissions associated with AI models and systems, with particular focus on generative video (Sections 4 and 5).

[2] Estimated at 90 Wh per 5.4 second video from Delavande et al. (2025) and a carbon intensity of US electricity of 0.410 kg CO_2e per kWh.

[3] We asked all the leading model developers to share data / methodology and received no substantive responses.

- Propose best practice methodologies for how AI-related emissions for video generation should be estimated and reported (Section 6).
- Contextualise AI emissions by comparing AI solutions to counterfactual non-AI solutions, with specific relevance for the digital media industry (Section 7).
- Encourage more open, transparent, and consistent AI-related emissions data sharing across the value chain (throughout and Section 8).

3. Background

3.1. Generative AI is already changing the way we think about media production and content creation

While AI is being broadly adopted across virtually all sectors, the application of generative AI in media production and content creation is on the rise. We already have examples of the speed with which generative AI is changing production processes, such as the production of video ads entirely with generated video.

The reality of how generative AI is used in production environments is more nuanced, however. Rather than always simply replacing existing processes wholesale, production teams are exploring and experimenting with how the technology can support and enhance their existing work throughout the production process (i.e. from pre-production to post-production and through to marketing and distribution), while grappling with the carbon (and other) impacts of its use.

Table 1. Example applications of generative AI-assisted technologies in media production processes

Production phase	Applications of generative AI in media production	Description
Pre-production	Scripts	Use of generative text-to-text or text-to-audio to test ideas and analyse scripts
	Storyboarding	Use of generative text-to-image to generate storyboards and reference imagery
	Shot lists	Use of generative text-to-text to suggest shot lists
Production	Backgrounds and elements	Use of generative text-to-image and text-to-video to generate backgrounds and elements that appear on screen

	AI-assisted photography	Use of AI-driven cameras to better track actors and make adjustments to focus and framing
Post-production	Reshoots	Use of generative AI to edit video in place of reshoots
	Visual effects	Use of generative AI to enhance visual effects

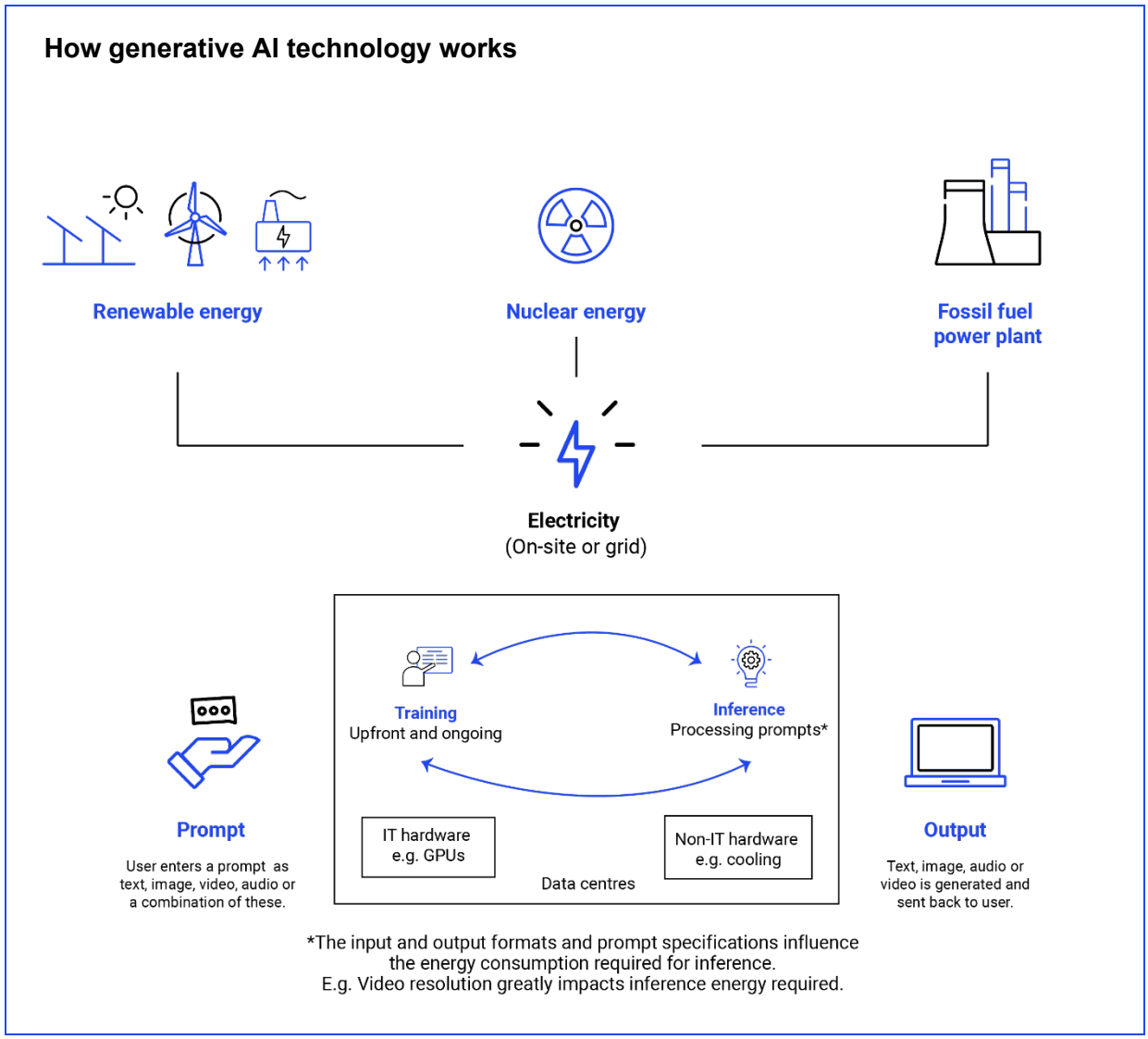
Depending on which set of headlines you read, you might interpret the increasing adoption of AI as either a solution (Nicholas Stern, 2025) or accelerant (Joshi, 2026) to the climate crisis. Part of this comes down to an unhelpful conflation of terminology in the popular discourse with 'AI' used to describe multiple different underlying technologies, from long-standing machine learning approaches, and predictive modelling, as well as the newer 'generative' applications.

Our focus for this report is generative AI, specifically for video and understanding the climate impact of these technologies. We acknowledge, but don't attempt to address, other potential effects such as impact on jobs, electricity costs, resource consumption or air and noise pollution.

And when thinking about the climate impacts of generative AI for video, there is a difference between use cases which would displace an alternative method of making videos, and the use of generative AI to create content that *otherwise would never have existed*. The carbon impact of the former should be considered in a comparative sense, whereas the latter will be additive. We acknowledge both aspects in this paper, but with a focus on understanding the comparative case, and considering how to minimise emissions where generative AI for video is used.

3.2. Implications of the data centre sector's environmental footprint on generative AI emissions

While we may be limited in our understanding of the environmental impacts of specific AI models, we have a clearer view of the impacts of the data centre industry, and therefore generative AI, albeit at an aggregate and global scale.



In 2024, the International Energy Agency (IEA) estimates that data centres consumed about 1.5% of global electricity use, equivalent to about 415 terawatt-hours (TWh) of electricity

The IEA estimates that this figure is expected to more than double to around 945 TWh by 2030, with AI being the most prominent factor of this growth as training and use of generative AI models continue to increase. Exact projections are difficult to predict so the IEA has modelled a range of scenarios which peg a range between about 700 TWh and 1250 TWh by 2030 (IEA, 2025c).

A separate report from the IEA-4E programme assessed estimated ranges for global electricity use in data centres attributable specifically to AI and found a likely range between 200 and 400 TWh by 2030 as shown in Figure 1 (Kamiya, 2025), which equates to roughly 30% of the IEA's projected global data centre electricity consumption.

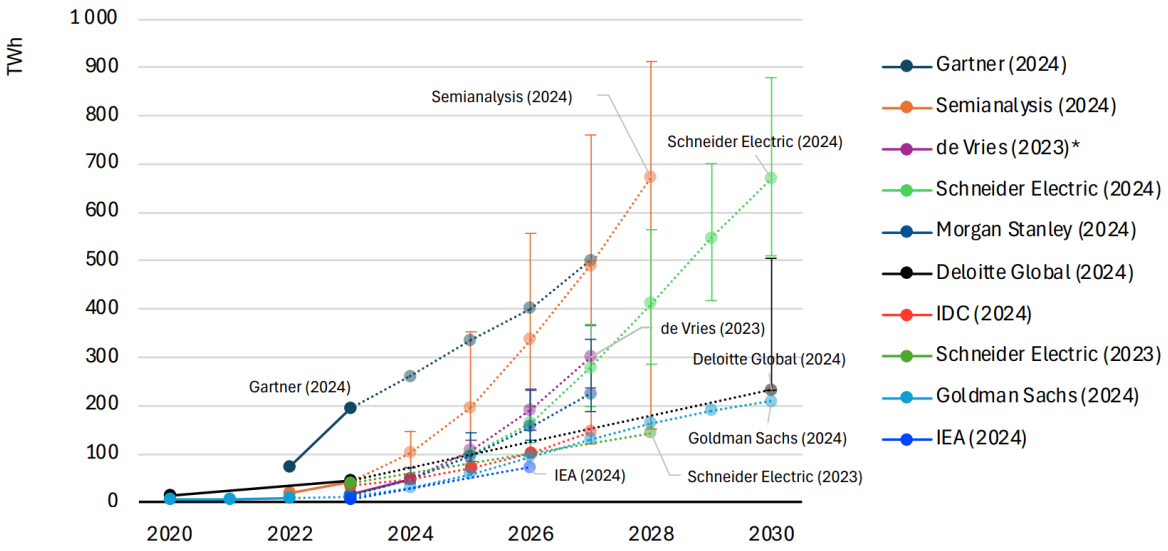


Figure 1. Selected global AI energy use projections, 2020-2030, (Kamiya, 2025)

Notes: Darker circles indicate historical estimates; lighter circles and dotted lines indicate projections. Error bars indicate ranges included in scenarios. de Vries (2023) totals based on linear interpolation between the study's assumed 2023 and 2027 production of AI accelerators, to calculate cumulative energy consumption (i.e., each year's newly built AI accelerators added to the consumption of existing stock).

Powering the use of generative AI data centres

Data centres, which house the powerful servers used to train and operate generative AI models, are the main contributor to AI's direct environmental footprint. These facilities are designed to accommodate servers, storage systems, networking equipment, and associated components, all installed in racks and organised into rows.

In terms of size, measured in megawatts (MW), a conventional data centre typically ranges from 10 to 25 MW in capacity. In contrast, a hyperscale, AI-focused facility can exceed 100 MW, drawing electricity at levels comparable to 100,000 households (IEA, 2025a). The largest data centres currently under construction can reach approximately 2,000 MW (equivalent to 2 million households), while the largest planned facilities considered are projected to have capacities of up to 5,000 MW.

Meeting this rapidly growing electricity demand requires a diverse energy mix. Renewables are expected to meet around half of the global growth in data centre electricity demand, supported by energy storage and the broader power grid (IEA, 2025c). Meanwhile, natural gas plants are being deployed on-site to support new data centres, especially in the US, where the gas power capacity in development tripled from 2024 to 2025 (Global Energy Monitor, 2026). Other dispatchable sources – those which can be adjusted or turned on or off – will continue to play a key role, alongside efforts from the tech sector to advance nuclear and geothermal solutions.

The largest regional hotspot is in the United States, which is estimated to consume nearly 100 TWh of electricity from GPU-accelerated AI servers in 2023

In the United States, which accounts for the largest share of global electricity consumption from data centres, a 2024 report from Berkeley Lab estimates that GPU-accelerated AI servers used nearly 100 TWh of electricity in 2023 in the US alone (Shehabi, 2024).

This regional clustering is shown below in Figure 2. In the United States, data centres already account for more than 10% of electricity consumption in six states and in Virginia this figure reaches roughly 25%. (IEA, 2025a)

Figure 2. Global map of data centre clusters, (IEA, 2025a)

Globally, data centres are projected to be one of only a few sectors that see an emissions increase between now and 2030

When viewed through a carbon lens, the IEA estimates that global emissions from the electricity consumption of data centres account for around 180 million tonnes of CO_2 in 2024, which represents roughly 0.5% of global fuel combustion emissions.

By 2030, data centre emissions are expected to grow to about 1% of global combustion emissions in their base case scenario, with the potential to rise as high as 1.4% in a high growth 'lift-off' scenario.[4]

[4] The IEA base case projections are based on the latest industry forecasts for server shipments, while the lift-off case assumes stronger AI uptake and fewer local constraints on data centre buildout.

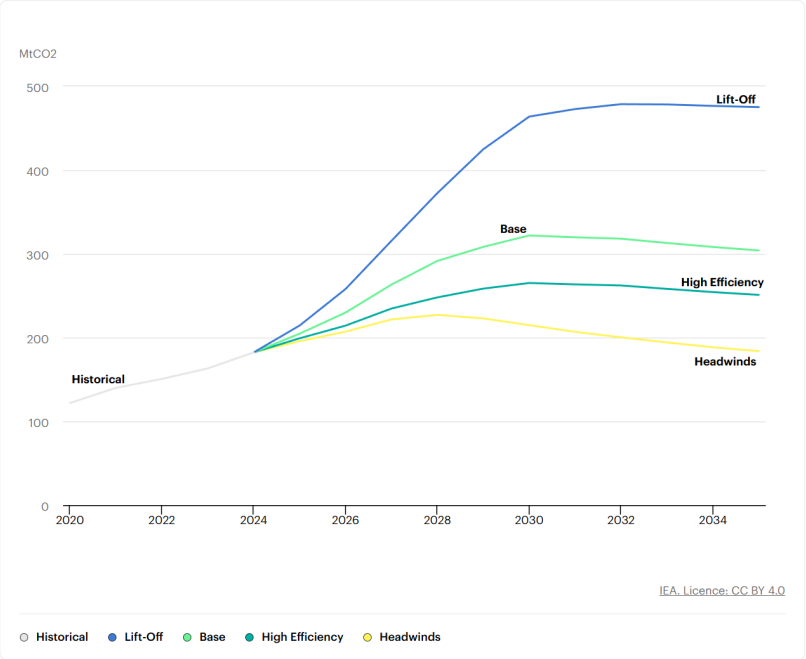


Figure 3. Estimated global carbon emissions of data centres, (IEA, 2025b)

How is the growth of generative AI affecting emissions?

We're already seeing evidence of increased emissions from the ongoing data centre buildout, where in the United States emissions jumped in 2025 after two consecutive years of emissions reduction (Rhodium Group, 2025).

This growth in emissions is attributed primarily to unusually fast growth in electricity demand, which can be traced back to the rapid deployment of new data centres, often built to support AI use cases. The clustering effect is saturating existing electricity supply in data centre hotspots, which in turn is driving local utilities to develop more electricity generation capacity and placing strain on grids.

In the midst of long connection queues to the grid in most markets, some data centre companies are also looking to 'skip the queue' through behind the meter (off-grid) solutions. In some instances, this demand can accelerate the build-out of renewables (Carbon Trust, 2025), but in others it can stimulate additional fossil fuel generation capacity (Global Energy Monitor, 2026).

Yet the path ahead is uncertain, demonstrating the need for a collaborative approach to measuring, monitoring and managing the carbon impacts of generative AI

We can see how the near-term implications of the adoption of generative AI are starting to effect emissions in both the data centre sector and in the wider technology sector, where digital companies that are heavily investing in AI saw operational emissions rise between 2020 and 2023 (ITU and WBA, 2025).

The longer-term outcomes are uncertain, and AI technology is rapidly evolving with many different factors at play, such as service demand, efficiency, power delivery techniques, deployment constraints

and policies, which will all affect future outcomes (Koomey & Masanet, 2026). To better plan for the path ahead, we need constructive dialogue among stakeholders to develop consistent measurement approaches, improve transparency and monitor carbon impacts as the technology continues to evolve.

3.3. The landscape of transparency for the carbon impact of generative AI technologies

Generative AI technologies are rapidly evolving and, with the growing role of these technologies in business and everyday life, public and customer expectations around accountability are evolving as well

Public stakeholders are increasingly demanding accountability, and customers need emissions data to inform their own decision making and carbon reduction planning, as well as to meet their reporting obligations. The data centre sector has demonstrated before that it is receptive to these needs where, influenced by requests from customers, major hyperscalers now voluntarily offer customers reports that show the carbon footprint of their cloud service use.

As climate and sustainability reporting requirements evolve globally, disclosure of energy and climate metrics can help to meet these customer needs, while also proactively and transparently adding to the discourse around generative AI's development.

Greater transparency would help model providers, customers, grid and utility planners, and the general public collectively make more informed decisions. For digital media companies, understanding the environmental impact of producing a piece of content is crucial to inform decarbonisation strategies and address questions and concerns from their wider employee base.

Environmental reporting across AI models remains inconsistent, with significant variation in how and whether developers disclose their impact

OpenRouter, a major Application Programming Interface (API) platform for large language models, publishes traffic data and token usage for the top 20 models. An analysis of May 2025 activity shows a disclosure gap where 84% of usage comes from models with no environmental disclosure, 14% through models with indirect disclosure, and just 2% from models with direct disclosure (Luccioni, Gamazaychikov, Alves de Costa, & Strubell, 2025).

Within the top 20 most-used models, only one (Meta Llama 3.3 70B) directly releases environmental data, while three (DeepSeek R1, DeepSeek V3, Mistral Nemo) provide indirect disclosures by sharing compute details (e.g. GPU type, training duration). In a landscape where environmental disclosures remain limited even among major AI model developers,[5] the transparency gap is even wider for smaller developers such as startups and academic labs, where measurement and reporting practices are less established and financial resources more limited.

[5] We directly asked the major developers to share their emissions calculations but didn't receive any substantive responses.

There is some progress being made on transparency: Some organisations have started voluntarily disclosing estimates of the environmental footprint of generative AI, while third parties are starting to consolidate and benchmark available data for AI users

Mistral AI has initiated a lifecycle analysis (LCA) of its AI model, assessing the environmental footprint from the model conception to the end-user equipment, focusing on greenhouse gas emissions, water use, and resource depletion. The study was peer-reviewed by two consultancies specialised in environmental audits in the digital industry (Mistral AI, 2025).

Google has developed a methodology for measuring the carbon emissions, energy consumption, and water consumption of AI inference (Elsworth, et al., 2025). While these initiatives are promising, the absence of consistent methodologies and assumptions makes comparisons of AI models across providers challenging.

When taken together, the theme is one of inconsistency, which can partially be traced back to a lack of consistent and agreed measurement methods. We explore these inconsistencies in Sections 4 and 6 to determine where the gaps are and how we can begin to address them.

How does generative AI work and what underpins its energy consumption?

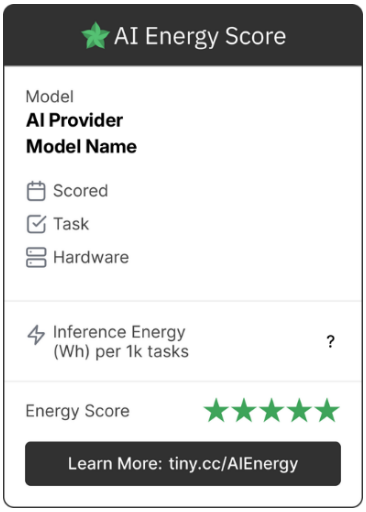
Various benchmarking frameworks and tools are available enabling comparison between models on energy use for inference.

The ML.ENERGY leaderboard measures and showcases inference time and energy consumption across text-to-text, text-to-image and text-to-video models (ML.ENERGY, 2026).

The AI Energy Score label assigns models a relative energy efficiency rating based on their GPU watt-hour consumption for specific tasks, awarding them one to five stars and publishing the results on a leaderboard (AI Energy Score, 2026).

On the user side, the Carbon AI Extension for web browsers developed by Hiili tracks energy consumption and emissions from various LLM platforms by counting input and output tokens. (Caravaca, Cuevas, & Cuevas, 2025).

AI Energy Score label



4. Estimates of generative AI's carbon impact

4.1. A lack of data means emissions are hard to quantify

For the average user trying to better understand the emissions impact of their generative AI use, it's very difficult to come up with a reliable estimate. And for companies seeking to better estimate, measure and reduce the emissions impact, this is problematic as they are largely left in the dark.[6]

However, through academic literature and studies from technology companies themselves, we can glean useful information to inform a best guess of emissions impact to train and use generative AI models today and determine where we need to improve in terms of measurement methodology and reporting transparency in the future.

A word of caution: generative AI technology is rapidly evolving and any energy and emissions estimates from even just a few years ago are quickly outdated. Thus, we seek to establish a reasonable view of what emissions are like today (or in some cases, within the past few years) but acknowledge that these are just a snapshot in time.

How does generative AI work and what underpins its energy consumption?

Generative AI is an energy intense software application because it involves a particularly large amounts of mathematical operations, carried out on specialised GPUs which are currently only feasibly operated in data centres. The architecture of a generative AI model determines the 'logic' of how a model performs tasks and which applications they are suited for. The structure of a model also heavily influences energy consumption, therefore carbon emissions. As generative AI technology is advancing rapidly, development and selection of model architectures for specific applications is also changing fast. At present, the two most prevalent architectures are large language transformer models (for text) and diffusion models (for images).

Large Language Models (LLMs) generate text by 'predicting' words based on context. They are trained on large amounts of text data to learn the structure of language. For LLMs, the model architecture is typically a transformer model. Transformer models can focus on different parts of the input text simultaneously while making predictions through a 'self-attention' mechanism.

- For energy consumption, the length of the response matters more than the input text (the query, chat history and any attachments).
- Users can therefore constrain this energy use in their prompt by asking for a succinct answer or specifying its length.

Diffusion models are used in applications such as image and video generation. A diffusion model starts from just an image of noise (a random distribution of coloured pixels in the output image) and repeatedly removes noise until an image emerges.

[6] For example see (Luccioni, Gamazaychikov, Alves de Costa, & Strubell, 2025)

During training a diffusion model is taught what a normal image looks like as it is progressively diffused (made more noisy). During inference (image generation) the process is reversed. The training process also associates a text condition (a description of the image) with the image. When an image is generated, the user's prompt acts as the text condition.

Video generation is more energy-intensive than image generation because it extends image generation across time. Not only is a video composed on many images in each second (the number of frames) but a video model must also represent space and time: what appears in each frame; how objects and camera position change across frames.

A slowly changing scene is often easier because adjacent frames are similar. Fast or complex motion is harder. Models that are capable of generating complex scenes thus tend to be more complex and thus energy consuming. For a given model however the complexity of a specific scene does not affect the energy consumption.

For energy consumption, as we discuss in the sensitivity analysis in Section 5, the main influences are the number of denoising steps, the resolution of the images and the length of the video. Reducing these parameters reduces the energy consumption associated with generative video production.

4.2. Estimated emissions to train generative AI large language models

There is a broad understanding that training generative AI models, such as large language models (LLMs), uses a lot of electricity, but the big question is just how much? And what is the resulting carbon impact?

We acknowledge that, while this paper aims to bring clarity to AI video generation, existing studies don't explore training emissions related specifically to these types of models. Instead, our best understanding of how training emissions for generative models are estimated is related primarily to text-to-text LLMs.

In reviewing the available estimates of training emissions of LLMs, see Figure 4 below, a key point to bear in mind is that these figures do not follow consistent methodologies nor have identical boundaries, e.g., some only measure GPU emissions, while others include an estimate of emissions from all IT hardware in the data centre used for training and include data centre overheads. So, they are useful to get the lay of the land, but difficult to make direct, 1:1 comparisons.

A 2021 study by Google and UC Berkeley explored this and found that OpenAI's GPT-3 model (arguably the leading text-to-text LLM at the time), was trained using 1,287 MWh of electricity over a total of nearly 15 training days (Patterson, et al., 2021). In carbon terms, this was estimated to equate to 552 tonnes of CO_2-equivalent emissions (tCO_2e), roughly the amount of carbon emissions from construction of a 1,000 m^2 industrial building.

Since then, a number of studies have researched this further. Findings indicate that there is no one size fits all carbon impact value that we can assign to model training, as it depends on many variables, from the architecture of the model, to the model size (generally indicated by the number of parameters, in billions), the GPU and IT hardware the model is trained on, where the data centre is located, the generation source of electricity used, and so on.

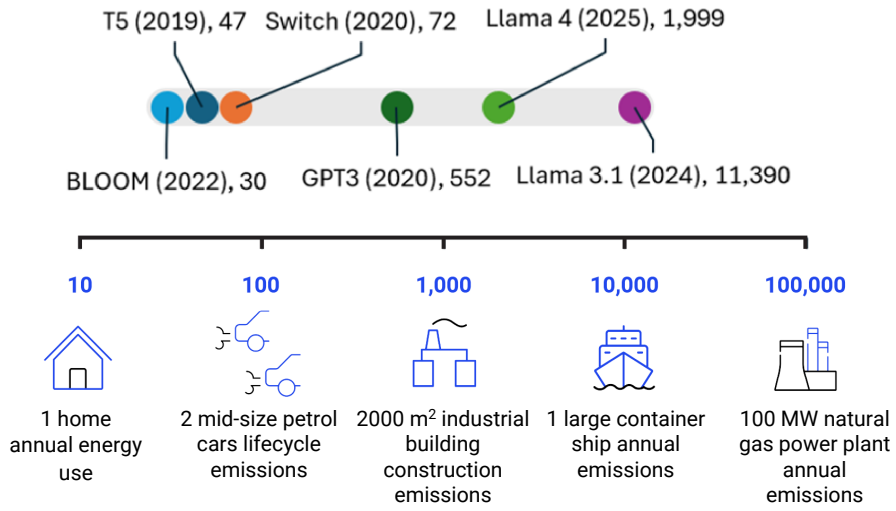


Figure 4. Carbon emissions from LLM training of six selected models, including reference points

Figure 4 was developed based on analysis from (Patterson, et al., 2021) *and* (Luccioni, Viguier, & Ligozat, 2022) *in addition to Llama model cards* (Meta, 2025a), (Meta, 2025b).

What's striking is the breadth of the range, where on the low end, models can be trained with a carbon impact around 50 tCO_2e, stretching up to around 10,000 tCO_2e on the high end.

For a sense of scale, that's like comparing the annual emissions of one mid-size gasoline-powered car to the annual emissions of a large container ship.

We can also see that more recent foundation models show an order of magnitude increase in training emissions relative to foundation models from two generations ago, such as GPT-3. It's difficult to say if this is a trend, i.e., training newer models is generating even more carbon emissions. But analysis from Epoch AI in June 2025 does indicate that since GPT-4 was trained in 2023, over 30 publicly announced AI models (including Veo 2, Veo 3, Claude 4, Grok 3 and Mistral Large) have been developed with similar amounts of compute (Epoch AI, 2025). The cumulative emissions across all of these are likely to be significant.

From a transparency perspective, Meta's Llama 3 and Llama 4 series of models are one of few foundation models with documentation that explicitly states the carbon emissions of training.

Mistral AI has measured and reported the lifecycle emissions of their Large 2 Model following the Frugal AI methodology developed by AFNOR (AFNOR, 2024), however they do not explicitly separate training and inference emissions. Meta's model card documentation (Meta, 2025a), for example, only captures emissions from electricity consumed by the GPUs used to train the models (using a location-based accounting approach), and doesn't capture the full impact from other IT hardware, nor the embodied emissions of the hardware and building infrastructure.

While there has been limited voluntary disclosure of training emissions to date, the EU AI Act (article 53 (EU Artificial Intelligence Act, 2024)) requires that developers of general-purpose AI models provide information on the energy consumption of the models, so at least we should see disclosure of this part of the puzzle in time (though as we've seen, a lot more detail would be required to have a full picture of these emissions).

An area warranting further study is how training emissions compare across different model architectures, such as diffusion transformer models, which are used more prominently in image and video generation.

4.3. Estimated lifecycle emissions of LLMs

Training is only one piece of the emissions puzzle for generative AI models. Much like other products, such as a car or a computer, generative AI models have a lifecycle, from beginning to end-of-life. This lifecycle starts with inception and design, includes training, inference, and eventually decommissioning (the full lifecycle is explored in greater depth in Section 6, but this gives you a rough idea).

As it relates to generative AI applications, such as generative video, we're interested in understanding how the various stages of the lifecycle compare from a carbon perspective. This helps give us a sense of where in the lifecycle and use of the model the largest impacts are that we should be concerned about and seeking to improve, as well as provides insights into how to appropriately account for emissions across the lifecycle of the model.

Figure 5 below shows the breakdown of emissions by lifecycle stage, represented by training, inference, embodied carbon and other, as a percentage of total lifecycle emissions across four models.

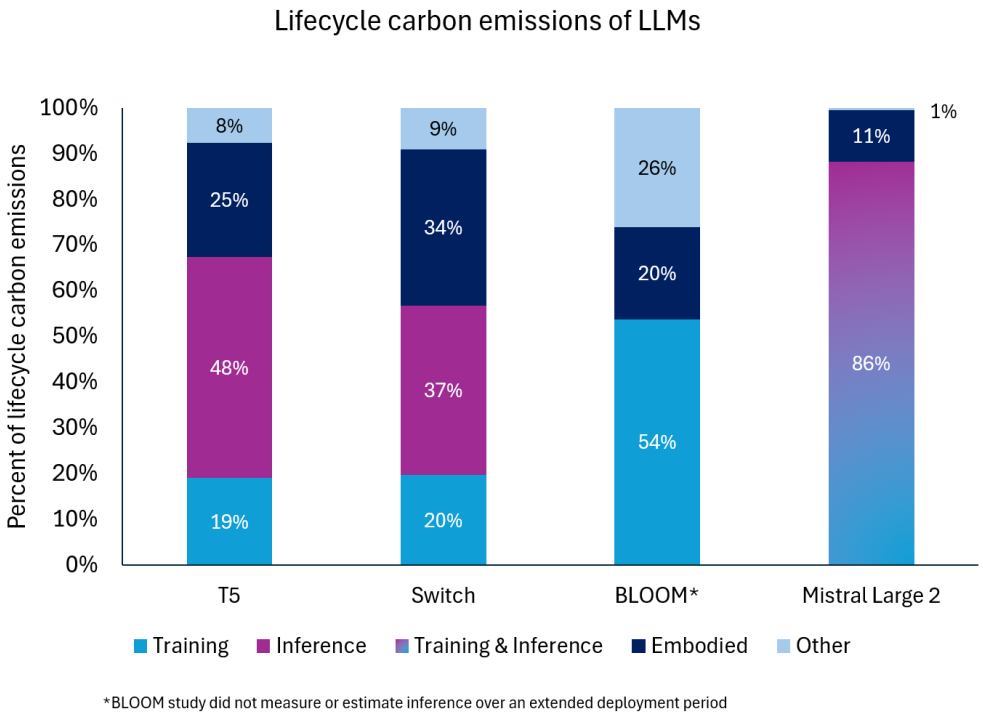


Figure 5. Lifecycle carbon emissions of LLMs of four selected models

Figure 5 was developed based on analysis from (Patterson, et al., 2021)*,* (Faiz, et al., 2024)*, and* (Luccioni, Viguier, & Ligozat, 2022) *in addition to reported figures from Mistral AI* (Mistral AI, 2025)*. For T5 and Switch models, other includes emissions from experiment and storage. For BLOOM, other includes emissions from idle GPUs. For Mistral Large 2 other includes emissions from model conception, network traffic and end-user equipment. Mistral AI reported that Mistral Large 2 produced 20,400 tonnes CO_2e over an 18-month operating period (roughly twice the Llama 3.1 training emissions reported by Meta).*

The variation in boundary, the electrical grid and methodology used to evaluate the lifecycle carbon emissions is immediately evident, making drawing conclusions difficult. We can, however, draw a few meaningful conclusions, namely that embodied carbon is significant in all studies, ranging from 11% to 34% across the four models evaluated.

From T5 and Switch, we also see the relative size between training and inference, where inference is roughly 1.5 to 2.5 times the emissions of training[7]. However, we expect that aggregate inference emissions over the lifecycle of a model, and therefore its relative size compared to training, to be highly dependent on the popularity of the model.

[7] Meta observed a 'rough power capacity breakdown of 10:20:70 for AI infrastructures devoted to the three key phases — Experimentation, Training, and Inference' (Wu, et al., 2022), further evidence of the variation in this ratio.

4.4. Estimated energy consumption and emissions of generative AI inference

Turning to the emissions impact of inference, we have reviewed recent analysis to get a sense of how different applications (text-to-text, text-to-audio, text-to-image and text-to-video) compare.

Importantly, the analysis presented in this section is not an exhaustive review of energy and emissions metrics of generative AI inference. We also understand that there are many variables which influence the energy consumption of a single inference output, from the length of the input query to the configuration of output parameters, such as image resolution. The effect of these variables is not intentionally represented here, but the ranges implicitly capture this variation in some cases.

Furthermore, as we have touched on previously and as is discussed in greater depth in Section 6, these studies are not completely consistent from a methodological perspective, so the precision by which we can draw conclusions from these figures is limited.

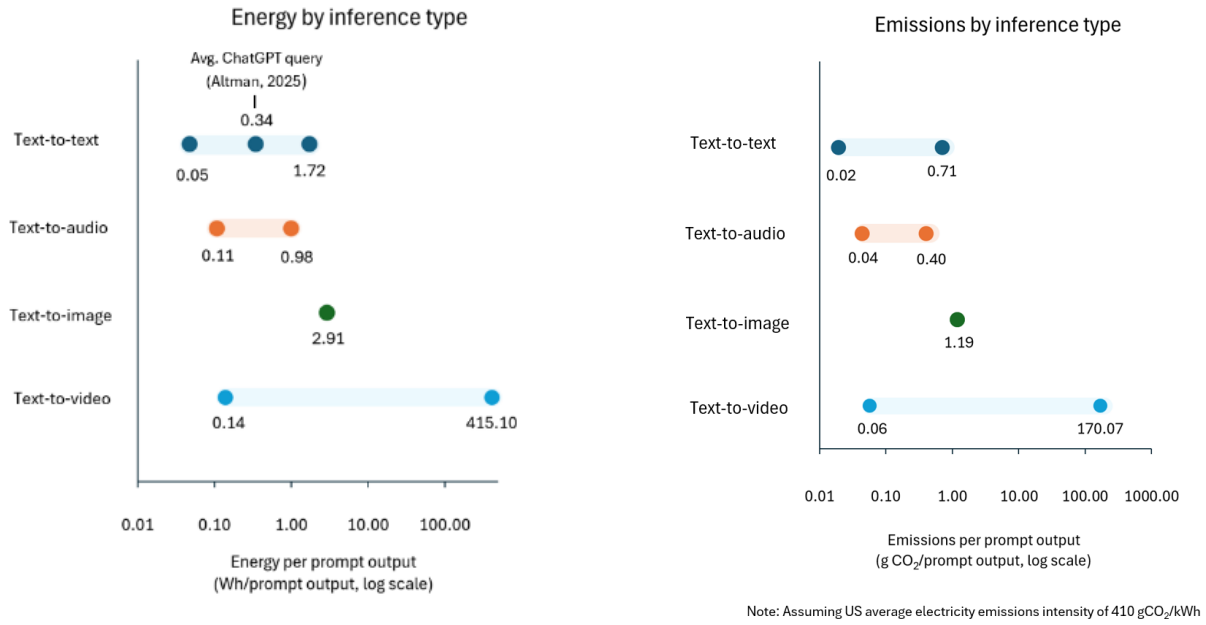


Figure 6. Energy and emissions by inference type from various studies

Energy values in Figure 6 *were developed based on analysis conducted in various studies and is not exhaustive. Text-to-text figures (min, max) are based on analysis from* (Luccioni, Jernite, & Strubell, 2024) *and* (Elsworth, et al., 2025)*, respectively. Text-to-audio figures are based on analysis from* (Passoni, Ronchini, Comanducci, Serizel, & Antonacci, 2025)*. Text-to-image is represented by a single value based on analysis from* (Luccioni, Jernite, & Strubell, 2024)*. Text-to-video figures are based on analysis from* (Delavande, Pierrard, & Luccioni, 2025)*. Carbon values were developed by applying an electricity emissions intensity of 410 gCO_2/kWh, representative of the US average.*

On a relative basis, we find that there may be roughly an order of magnitude between the carbon impact of text-to-text inference and text-to-image. We also find that there is likely another order of magnitude between the carbon impact of text-to-image and text-to-video.

Using Sam Altman's stated figure[8] of 0.34 Wh for how much energy a ChatGPT query uses (Altman, 2025), we see that image generation is roughly 9 times greater. And comparing image generation at 2.91 Wh/image to video generation (based on 90 Wh for a single short video from (Delavande, Pierrard, & Luccioni, 2025)), we see that video generation is roughly 30 times more costly, though this heavily depends on the models used and parameters selected for image and video generation, as explored further in Section 5.

5. Sensitivity analysis of the carbon impact of AI text-to-video generation

To better understand the carbon impact of AI video generation, we analyse how the lifecycle carbon impact may vary based on the size of training emissions, inference parameters which users may select such as resolution, number of frames and denoising steps, as well as the carbon intensity of the electrical grid. This is a useful exercise to determine a plausible range of carbon impact based on existing knowledge in this area, as well as to demonstrate the wide range of uncertainty around such estimates which underscores the need for a consistent measurement approach.

Our analysis includes an estimation of the emissions from model creation (i.e. training) and inference, both operational and embodied (this boundary is explored further in Section 6.1.2). To simplify the analysis, we assume that on-going retraining and testing emissions are aggregated with training emissions and that end-of-life emissions are immaterial.

A word of caution: This analysis is built on the theoretical prediction model from (Delavande, Pierrard, & Luccioni, 2025) to represent the energy consumption of *text-to-video* diffusion transformer models (specifically WAN2.1-T2V-1.3B in this case). The existing literature does not similarly explore image-to-video nor video-to-video models, so we cannot draw any meaningful conclusions for other types of models. Furthermore, our analysis is heavily reliant on a single study characteristic of a specific text-to-video model, and relies on many assumptions, so we expect a relatively wide range of uncertainty if applied to other video generation models.

5.1. Summarised results of sensitivity analysis

Figure 7 and Figure 8 present a summary of the sensitivity analysis results across two resolutions: 1280x720 (i.e. 720p) and 1920x1080 (i.e. 1080p).

Nominal values[9] for the carbon impact per 5.4 second video at 15 frames per second range from 49 to 88 gCO_2e at 720p.

[8] This is a widely quoted figure in the media, but we don't know what it is representative of or the methodology used to compute it. See (Sasha Luccioni, 2025) for discussion.

[9] 5.4 seconds and 15 frames per second were used as nominal values for this analysis because they relate most closely to the parameters available in the WAN2.1-T2V-1.3B model and to the analysis performed by Delavande et al.

Furthermore, we clearly see that the carbon impact is highly sensitive to resolution, as expected based on the prediction model, where carbon impact increases with resolution.

We also see the sensitivity of carbon impact relative to the size of allocated training-related emissions (this sensitivity is shown by the green range in Figure 7 and Figure 8, and can be seen in further detail in the Appendix) and that the electrical grid also affects the resulting carbon impact, roughly proportional to the carbon intensity of the local grid (the effect of which is seen Figure 9).

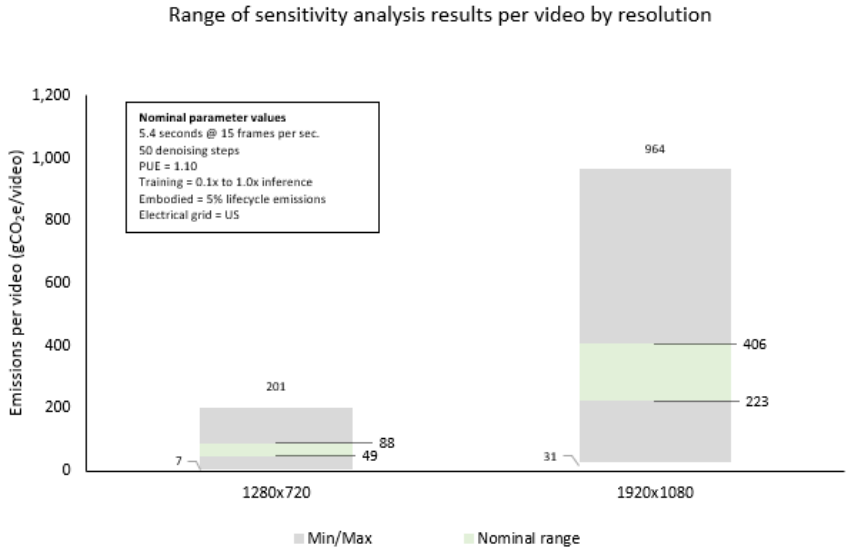
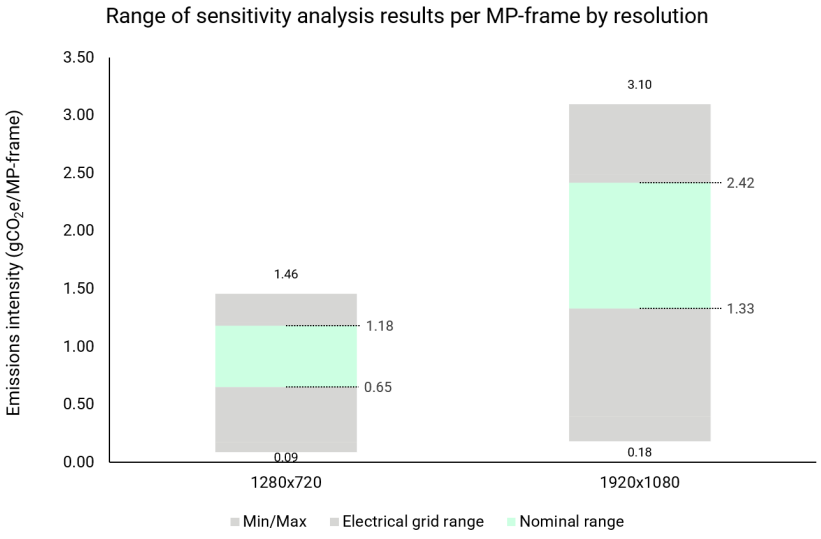


Figure 7. Sensitivity analysis results per video by resolution

In Figure 8, we evaluate an alternative metric, carbon impact per MP-frame, in part so that we can assess whether this metric would serve as a useful functional unit and allocation key (as discussed later in Section 6.2.4). We see that the increase in emissions is less pronounced across resolutions, but significant, nonetheless. We also see that while this metric does a better job of normalising impact across different resolutions, it does still not strongly correlate to the underlying energy consumption during inference, so we rely on emissions per video at specified video settings going forward.



Nominal range represents nominal parameter values and the US grid. Min represents the minimum value from the sensitivity analysis (5 denoising steps, 81 frames, US grid) and max represents the maximum value from the sensivitity analysis (50 denoising steps, 150 frames, US grid).

Figure 8. Sensitivity analysis results per MP-frame by resolution

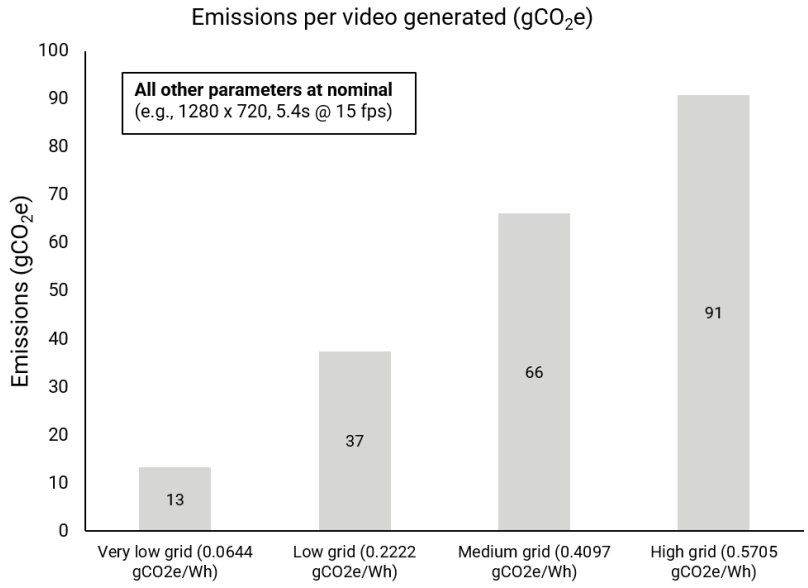


Figure 9. Sensitivity of emissions per video to the emissions intensity of electricity

In addition to resolution, training and the electrical grid, both the quantity of denoising steps and number of frames significantly affect the resulting carbon impact of video generation. As expected by the prediction model, the carbon impact of video generation varies linearly with the number of denoising steps (Figure 10).

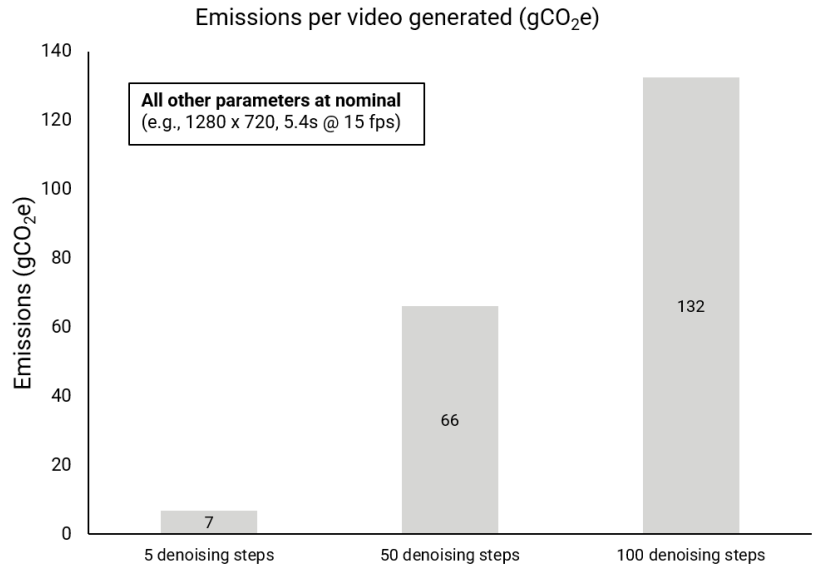


Figure 10. Sensitivity of emissions per video to number of denoising steps

Figure 11 shows the sensitivity to number of frames, where we see a quadratic increase in emissions per video. Importantly, this is expected behaviour based on the prediction model, where energy consumption relates quadratically to the number of frames.

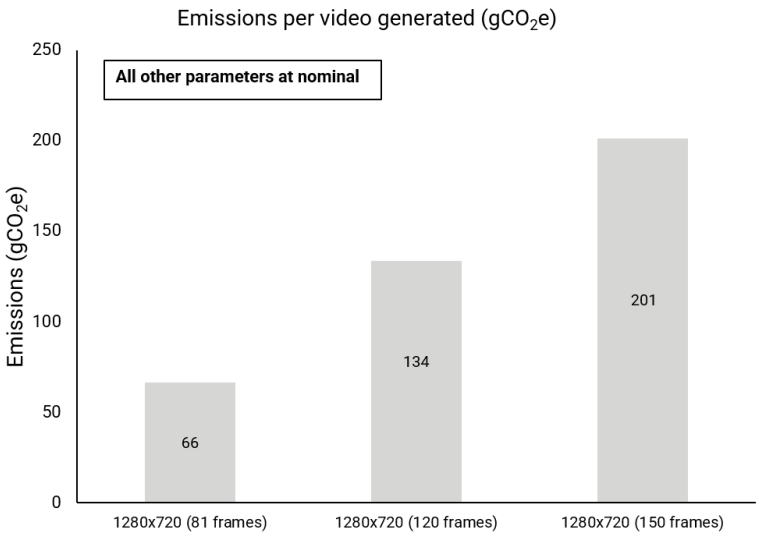


Figure 11. Sensitivity of emissions per video to number of frames

5.2. Sensitivity analysis conclusions

Clearly, we see that the carbon impact of AI text-to-video generation is highly variable and dependent on a number of factors. We have identified that the carbon impact is highly sensitive to both resolution and number of frames. To a lesser, but still meaningful extent, the carbon impact is also sensitive to denoising steps and the carbon intensity of electricity used in both training and inference data centres.

- The sensitivity analysis shows a large range of emissions, demonstrating the need for a consistent measurement and reporting methodology, which could be achieved through Product Category Rules.
- Nominally, we estimate that a 5.4-second video at 15 frames per second and 50 denoising steps (generated by WAN2.1-T2V-1.3B text-to-video diffusion transformer model) has a lifecycle carbon impact per video of around 50 to 100 gCO_2e at 720p and with the US electrical grid.
- The assessed functional unit of gCO_2e/MP-frame does a reasonably good job of normalising results, but due to the quadratic nature of the relationship between GPU energy consumption and resolution and frames, it is not perfect, so we instead select a functional unit of gCO_2e/video at specified video settings (e.g., in this case, a 5.4s video at 720p, 15 frames per second and 50 denoising steps).
- Carbon impact increases quickly as higher resolutions are selected.
- Model providers are encouraged to set default denoising step values that balance video quality with energy consumption and to provide users with guidance on striking this balance.
- For users, we suggest using the lowest resolution and video duration possible to meet their needs (in Section 7.6, we discuss the potential for upscaling to be used as a way to enable generation of videos at lower resolutions).

6. Measuring the carbon impact of generative AI

A key objective of this report is to propose a best practice methodology for how generative AI-related emissions should be estimated and reported for video generation. In doing so, we aim to improve how we measure and understand the carbon impact of generative AI from a lifecycle perspective. We also aim to encourage transparency and consistency in how the carbon impact of generative AI is measured and reported so that we can make more informed decisions about how we use this technology.

In this section, we review existing studies and methodologies[10,11] to assess completeness and consistency with carbon accounting standards (such as the GHG Protocol Product Standard and ISO 14067) and principles and propose a view on best practice in this area.

This section is a technical deep-dive and is intended for an audience of product carbon footprinting practitioners, however the findings and learnings of this section are applied in a case study which follows in Section 7. Furthermore, the boundary diagrams (Figure 14 and Figure 15) and functional units (Section 6.2) explored in this section may be of interest to a broader audience.

6.1. AI system lifecycle boundary

6.1.1. Boundaries established in literature and existing methodologies

The AI system[12] lifecycle may be defined as below in Figure 12 which follows the AI system from inception (where the idea for the model is initially conceived and explored), through design, verification, deployment, operation and ultimately retirement[13]. Green Software Foundation's Software Carbon Intensity for AI (SCI for AI) defines a similar lifecycle to include inception, design and development, deployment, operation and monitoring and end-of-life.

[10] The two principal methodologies reviewed being AFNOR's 'Frugal AI' Spec 2314 (a French Ministry for Ecological Transition and Territorial Cohesion initiative) and Green Software Foundation's Software Carbon Intensity for AI Specification (SCI for AI).

[11] At the time of writing this draft, the ITU-T L.1801 guidelines for measuring the environmental impact of AI have recently been released, these haven't been reviewed in detail alongside the existing methodologies.

[12] The word 'system' is used in contrast to 'model' to indicate that the model sits within a wider structure which supports and enables the function of the model.

[13] From this perspective, training would fall within the design and development stage and inference would fall within operation.

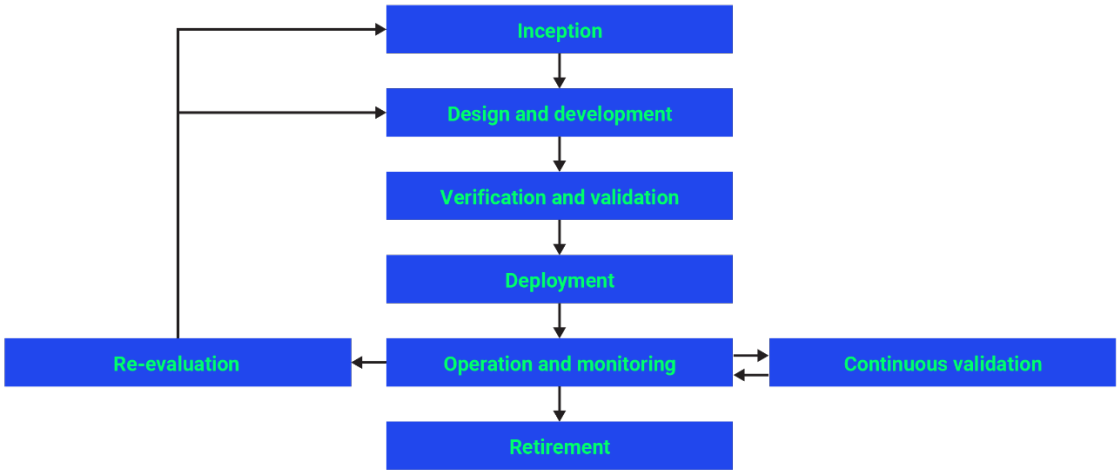


Figure 12. AI system lifecycle as defined by AFNOR Spec 2314 (AFNOR, 2024).

Table 2. Comparison of boundaries used in literature and current methodologies to assess energy and carbon impact of generative AI systems

Study/ Source	IT hardware boundary						Operational boundary		Embodied boundary			End-of-life (decomm-issioning)
	GPU	CPU	Memory	Network-ing	Storage	Idle capacity	IT hardware	Support services (e.g., MEP/PUE)	IT hardware	Building infrastructure		
										Building structure	Support services	
AFNOR Spec 2314	Y	Y	Y	Y	Y	U	Y	Y	Y	O	O	Y
GSF SCI for AI	Y	Y	Y	Y	Y	Y[1]	Y	Y	Y	O	U	Y
ICT Sector Guidance*	Y	Y	Y	Y	Y	Y[2]	Y	Y	Y	O	O	U
Patterson et al (2021)	Y	Y	Y	Y	U	U	Y	Y	N	N	N	N
Wu et al (2022)	Y	Y	Y	Y	Y	U	Y	Y	Y	N	N	N
Luccioni et al (2022)	Y	N	U	P	P	U	Y	Y	P[3]	N	N	N
Faiz et al (2024)	Y	Y	Y	Y	Y	U	Y	Y	Y	N	N	N
Legend	**Y = Yes (in boundary)**			**N = No (outside boundary)**		**P = Partially in boundary**		**O = Optional/recommended**		**U = Unspecified**		

*Notes for Table 2: *The ICT Sector Guidance Built on the GHG Protocol Chapter 4 covers Data Centre Services but does not explicitly address AI systems*
1. Determined via requirements of GSF Software Carbon Intensity (SCI) specification
2. If provisioned for the service
3. GPU and servers

The AI system lifecycle follows an iterative approach, where the model is continuously validated and periodically re-evaluated and potentially retrained during its lifecycle. This lifecycle view ensures that these processes are captured within the lifecycle boundary

Leading product carbon footprinting (PCF) standards (such as the GHG Protocol Product Standard), define the lifecycle stages as shown in Figure 13, which include material acquisition and processing, as well as production, distribution & storage, use and end-of-life. This example also demonstrates how, for a company that manufactures a product such as a car, the product lifecycle stages map to Scope 1, 2 and 3 as used in organisational reporting.

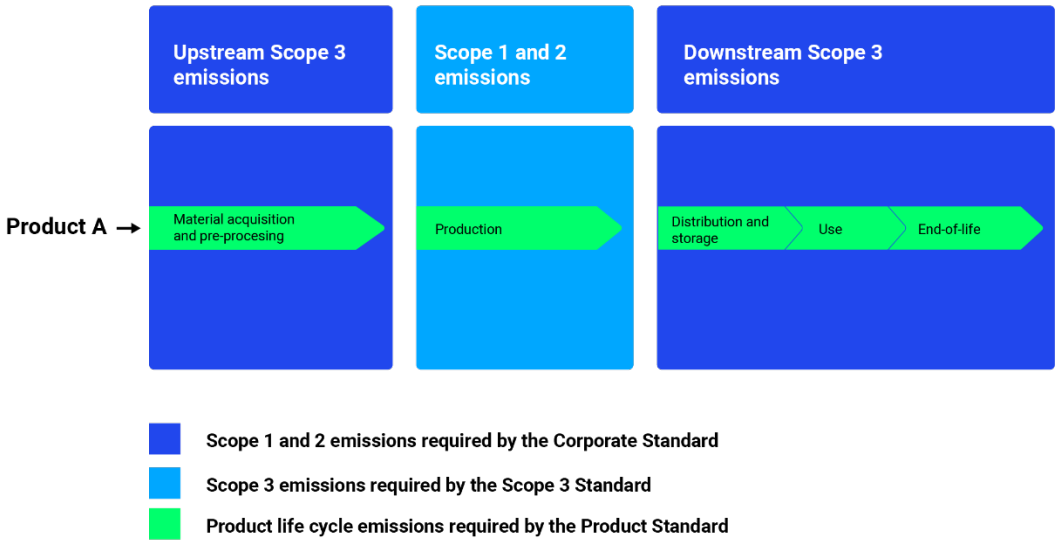


Figure 13. The GHG Protocol Product Standard lifecycle stages and their relationship with the Scope 1, 2 and 3 emissions for a company that produces 'product A' (Greenhouse Gas Protocol, 2011)

Building on the AI system lifecycle defined by existing methodologies and applying the approach from PCF standards, we assess in Table 2 how these approaches treat the various elements of the AI system. Specifically, we are interested in reviewing which elements are included in the operational boundary (such as the energy used by IT hardware and support services in a data centre), which elements are included in the embodied boundary (such as the production and end-of-life emissions of IT hardware and building infrastructure) and how the end-of-life of the AI system is treated.

Among the three established methodologies (AFNOR, GSF and GHGP ICT Sector Guidance), we see a good degree of consistency.

All of the key IT hardware components fall within the footprinting boundary, both from an operational and embodied carbon perspective. Support services (such as cooling, lighting and other overhead) are

included as well across all three. Typically, idle capacity and end-of-life of the AI system (i.e. decommissioning) are included in the lifecycle boundary, but in some cases these are not explicitly addressed.

We also note that while embodied carbon of IT hardware is included in the lifecycle boundary, the embodied carbon related to the data centre building structure and support services are either left as an optional inclusion or not addressed explicitly.

When reviewing studies from literature that evaluate the environmental impacts of AI models, we note a higher degree of inconsistency and an inclination to focus primarily on the operational emissions.

6.1.2. Proposed generative AI system lifecycle boundary

In Figure 14, we propose our view of an appropriate boundary to evaluate the lifecycle carbon impact of an AI system. This is broadly consistent with the methodologies established by AFNOR Spec 2314 and GSF's SCI for AI. For completeness and clarity, we explicitly identify idle capacity, embodied carbon (of IT hardware, building structure and support services) and retirement/decommissioning as within boundary.

Note that ultimately the boundary is intrinsically linked to the function of the product or service being evaluated. For more on this subject see Section 6.2.

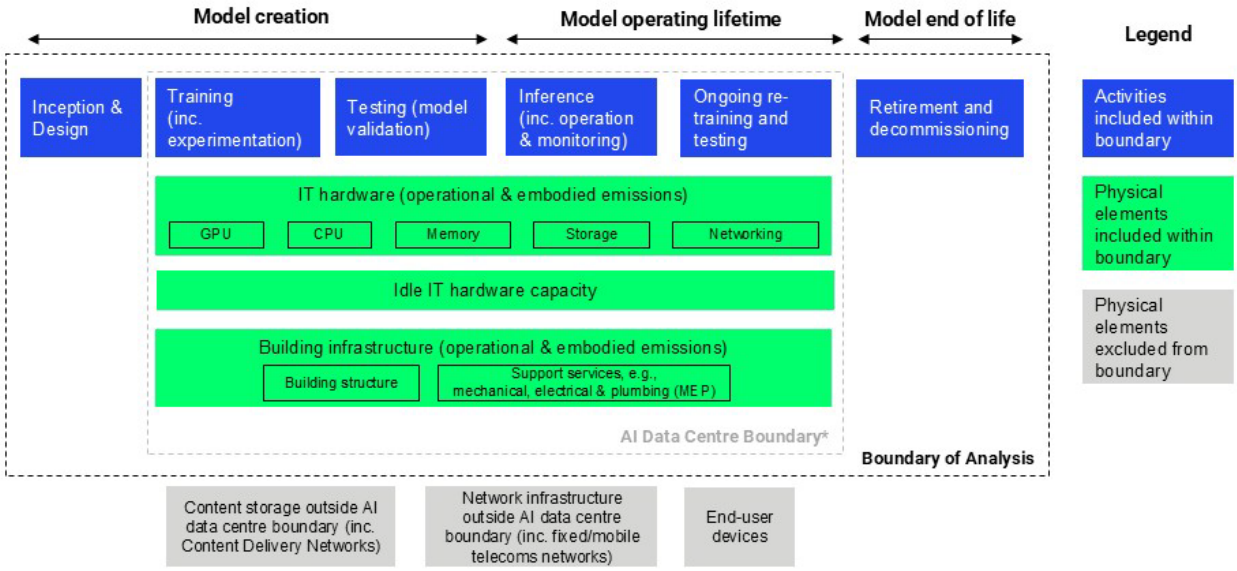


Figure 14. Proposed AI system lifecycle boundary

**The AI data centre boundary will comprise multiple data centres in different locations and with different purposes e.g. model training will often be carried out in data centres with tailored IT hardware for this task.*

Treatment of idle capacity in data centre IT hardware

Regarding idle capacity, we have observed from Google's recent study on the environmental impact of text generation that both overhead from support services and idle machines are a significant contributor to the total operational energy consumption during inference (8% and 10%, respectively) (Elsworth, et al.,

2025).[14] We also observe that in the Amazon Web Services (AWS) Customer Carbon Footprint Methodology v3.0, the approach inherently captures idle capacity through IT hardware assigned, but not necessarily fully utilised, to the service (Amazon Sustainability, 2025).

We therefore expect that emissions from idle hardware capacity to be significant and should be included in the boundary.

Assessment of materiality of embodied carbon of building infrastructure

The embodied carbon of IT hardware is generally well understood as being material (see Figure 5) and therefore recommended for inclusion in the AI system lifecycle boundary. However, the materiality of embodied carbon from the building infrastructure (e.g., building structure and support services) is less understood.

We evaluate this subject and determine that in many cases, but not all, we expect embodied carbon from the data centre building structure and support services to be relatively immaterial. We perform this assessment by comparing the embodied carbon emissions per m^2 of data centre floor space, amortised over the life of the building, to an annual operational carbon impact of IT hardware per m^2. We establish a materiality threshold of 1% relative to the emissions of IT hardware.

In our analysis, the embodied carbon of the building structure is amortised over a 60-year period (and assumes that MEP is replaced 2-3 times over that period)[15] and rack power densities are evaluated from 5 kW per m^2 to 22 kW per m^2. We develop a low, medium and high scenario for the carbon impact of IT hardware by further varying the emissions intensity of electricity.

Table 3. Materiality assessment of embodied carbon of data centre infrastructure

Emission source	Value	Unit	Ratio embodied/IT (>1% = significant)	Note
Embodied	117	$kgCO_2e/m^2$/year		Assume 7000 $kgCO_2e/m^2$ for building and MEP over 60 yr life
IT (low), operational	1,518	$kgCO_2e/m^2$/year	8%	Assume 5 kW/m^2 (50 kW/rack), 50% utilisation, located in France

[14] Note also this recent paper on the topic (Yiran Lei, 2026).

[15] We further evaluated materiality with a more conservative estimate of building structure lifetime of 12 years and determined that the findings hold, as in the medium scenario, the embodied to IT emissions ratio only increased to 0.75%.

IT (med.), operational	22,944	$kgCO_2e/m^2/year$	0.51%	Assume 12 kW/m^2, 65% utilisation, located in Virginia, USA
IT (high), operational	86,070	$kgCO_2e/m^2/year$	0.14%	Assume 22 kW/m^2 (130 kW/rack), 80% utilisation, Global avg. elec.

Sources: Embodied carbon estimates based on (ARUP, 2025). Estimates for rack power (kW/rack) derived from (Uptime Institute, 2025), rack dimensions from (Schneider Electric, 2015) and (Profile IT Solutions, n.d.). and utilisation rates from (Shehabi, 2024).

Our findings indicate that embodied emissions of building structure and MEP support services should normally be included as best practice but may be justifiably excluded based on materiality if rationale is provided through an analytical assessment. If excluded, this exclusion should be explicitly stated and justified.

In some cases, particularly where the emissions intensity of electricity is low and IT rack power density is also low, then embodied carbon can become significant, as shown in the example above. This is a particularly important point as grids continue to decarbonise that the embodied emissions of building infrastructure may become more significant.

Justifications for excluded downstream infrastructure

Some downstream infrastructure is excluded from the boundary of the methodology proposed here, while acknowledging that it is necessary for the full end-to-end process of AI video generation. These excluded elements - data storage,[16] network infrastructure,[17] and end-user devices – are indicated in grey boxes in the boundary diagram in Figure 14.

In other digital applications such as video streaming, content is produced once and delivered to multiple end-users. This means the emissions from storage and delivery of content are more significant per piece of content. In the case of AI video generation, we have seen in Section 4.4 that each prompt drives significant energy consumption during generation of the video, which we expect makes the storage and delivery of the resulting content less material per prompt. The intended focus of this paper is on measuring the emissions of AI video generation itself, so we propose the exclusion of downstream data storage and network infrastructure.

Similarly, we exclude end-user devices from the AI system lifecycle boundary as the intended focus of this methodology is specific to the AI video generation process which typically occurs within a data

[16] Data storage considers servers, outside of the AI data centres performing training or inference, storing information which is either used by models or created by models and delivered to end-users.

[17] Network infrastructure considers telecommunications infrastructure used to transport data to and from the data centres performing training or inference. This includes customer premise equipment, though any such equipment within the data centres used for training and inference is considered included, as indicated by the blue 'Networking' box in Figure 10.

centre. This proposed methodology is data-centre focused, but as end-user devices' capabilities to perform inference locally evolves and improves, an adapted boundary should be applied to capture these emissions. However, as explored in Section 6.1.3 when evaluating the emissions of a final video asset, end-user devices used during production of the asset should be included in the boundary for completeness.

6.1.3. Conceptual lifecycle boundary for generated video asset

The AI system lifecycle boundary proposed in Section 6.1.2 is built on existing methodologies and in reference to prior studies. However, as noted, there appears to be a gap in the literature outlining the lifecycle carbon impact of a production ready video asset.

We therefore propose a conceptual schematic of the lifecycle boundary for production-ready generated video assets, as show in Figure 15. As mentioned in Section 6.1.2, this boundary is heavily influenced by the function of the product or service being studied and the selected functional unit. For now, we present this boundary relative to a production-ready video asset, and we explore functional units in greater depth in Section 6.2.

As generated video is an output of generative AI systems and many users share the same system, the AI system impacts must be allocated or apportioned to each individual inference output (we cover this subject in more detail in Section 6.3). We also recognise that the energy and carbon impacts of generated video are dependent on parameters such as the model used, video resolution, denoising steps[18] as explored in Section 5 of this report through sensitivity analysis.

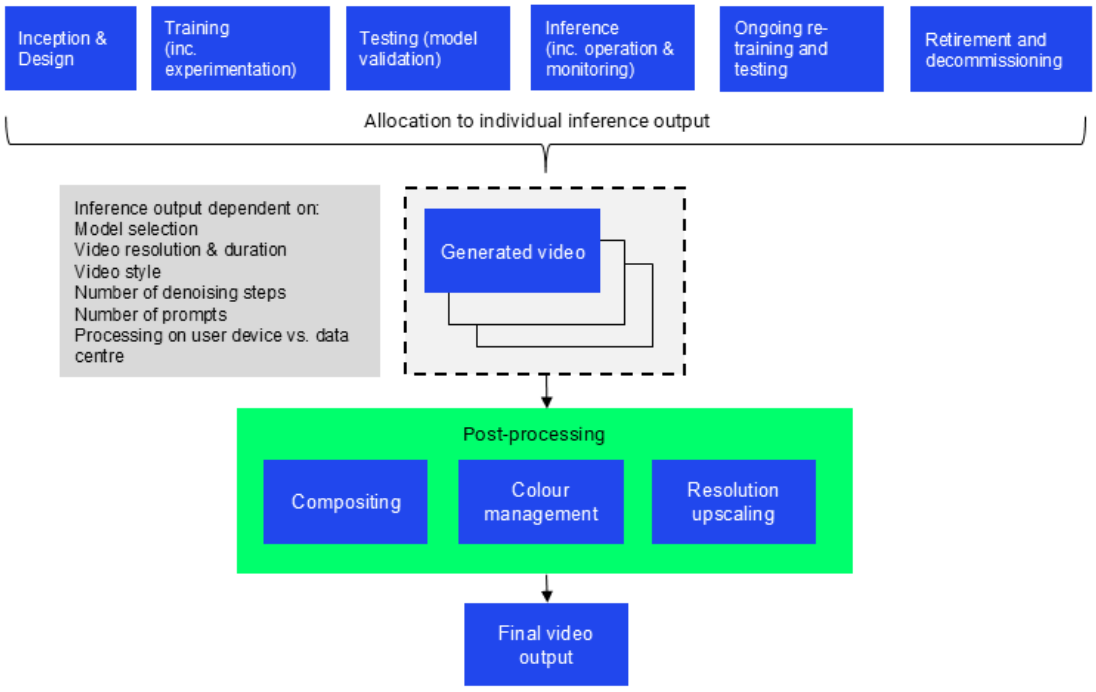


Figure 15. Conceptual schematic of generated video asset lifecycle boundary

[18] The iterative process used by diffusion models to convert noise into coherent video outputs. This process is identified as the primary of carbon emissions in one video generation model (Li, 2024).

Finally, we acknowledge that for a final video asset, the resulting carbon impact is affected by the quantity of videos generated and as well as by an array of post-processing activities, so these have been included in the boundary for completeness but are highly dependent on the specific video asset and production processes used.

6.2. Functional unit

When measuring the carbon impact of a product, the functional unit serves as a unit of analysis for defined characteristics of the product. We seek to explore and define best practice functional units (FUs) which can be used to measure the carbon impact of generative AI, with specific functional units defined for model training, inference and video generation application.

Selecting an appropriate functional unit requires understanding why the product was created, what its intended purpose is and what characteristics define the level of quality that the product offers.

Example functional units

Smartphone: Lifecycle emissions over a smartphone's 5-year life for everyday consumer use as a multi-purpose communications device (in this example, the product was created to meet consumer demand for a communications solution, intended for use as a multi-purpose communications device for 'everyday' uses such as phone calls, web browsing and social media, over a defined service life of 5 years).

Video streaming: Lifecycle emissions per minute of video streamed on a smart TV at 4K UHD and 60 fps.

6.2.1. Evaluation of functional unit with reference to existing methodologies

Two key themes are apparent from reviewing existing methodologies to measure the carbon impact of AI systems and services.

Distinction between AI system evaluation versus AI service evaluation

AFNOR Spec 2314 makes a distinction between the determination of a functional unit for an AI system and an AI service.[19] This way, there is more clarity for the impact of the model system itself as well as measurement of the total impact from a user perspective to access and use the model.

The AI system FU specified by AFNOR SPEC 2314 is 'Providing the system for one year for x queries' and the AI service FU specified is 'Providing the service for one year to all users'.

Similarly, Green Software Foundation's Software Carbon Intensity for AI Specification makes a specific distinction between impact evaluation from the provider perspective and consumer perspective.

[19] The system includes the impact from AI model lifecycle elements for training, retraining and inference over a one-year period, while the AI service additionally includes the impacts of other elements required for the end-to-end service to function, such as web hosting servers and end-user devices.

In this approach, the provider perspective accounts for the impacts from inception, design and development, deployment and retirement, while the consumer accounts for impacts from operation and monitoring (including inference) and retirement, but omits the preceding stages in the lifecycle.

In contrast to the approach taken in AFNOR SPEC 2314, GSF's SCI for AI Specification makes a clear delineation between evaluating impact to develop and train a model compared to using the model.

Absolute vs. intensity measurements[20]

AFNOR Spec 2314 and GSF's SCI for AI Specification diverge in their approach to measuring and reporting impact through adoption of absolute and intensity-based measurement approaches. While AFNOR Spec 2314 is not explicit in this regard, using their proposed functional unit would capture the absolute emissions over a one-year period.

GSF's SCI for AI adopts an intensity-based approach, where emissions from the model provider's perspective are reported per unit of compute, per token or per parameter. From the consumer perspective, emissions are reported similarly, such as per image.

Both approaches have their merits and are acceptable according to PCF standards. Using an absolute measurement approach can prove more transparent by showing the full-view of the emissions profile, while an intensity-based approach allows for better comparison to other studies since results are normalised to a specific unit of measurement. Furthermore, from a user perspective, an intensity-based approach can be helpful to determine the impact per output of the AI service.

6.2.2. Proposed approach

For maximum clarity and usefulness for model developers, model providers and end-users, we propose an approach which closely aligns and builds on these methodologies by delineating the measurement of carbon impact by training, inference and final video asset, as shown below. In doing so, we recognise that carbon impact measurement and reporting serve different needs for different groups.

Key to this approach is that the functional unit encompasses all attributable activities in the lifecycle for the measured outcome (see Figure 14 and Figure 15). So, for example, to measure the emissions impact of inference, we include an apportionment of emissions from training since the model is required to be trained in order to perform the inference and generate an output (see more about allocation in Section 6.3).

6.2.3. Unit of analysis for AI model training

Given the scale and significance of the carbon impact attributable to model training (see Figure 4), we propose the following approach for defining the unit of analysis[21] for AI model training, which clearly expresses the total emissions required to train the model. As here we are particularly focused on the training aspect of the AI system lifecycle, we view the model after initial training as an 'intermediate

[20] Absolute emissions refer to all emissions in a certain boundary over a specific time period (e.g., total lifecycle emissions over the life of a car), while emissions intensity refers to emissions per unit of measurement (e.g., lifecycle emissions per km driven)

[21] A unit of analysis is specified for intermediate products, rather than a functional unit, as the function of the product may be unknown as is the case with a multi-purpose generative AI model.

product' and omit certain lifecycle stages of the AI system (such as deployment, operation & monitoring, and retirement).[22]

Due to the iterative nature of AI systems, we proposed a unit of analysis related to the initial launch of the model, as well as a unit of analysis related to ongoing re-training of the model.

Unit of analysis (at initial launch): total model creation emissions per fully trained model for use at initial rollout, including inception, design, training and testing.

Unit of analysis (ongoing): total emissions to re-train the model, over a specified time-period, including validation activities associated with each re-training cycle.

Additionally, intensity-based units of analysis may supplement the units of analysis above, as proposed by GSF's SCI for AI.

6.2.4. Functional unit for AI inference

In defining a functional unit for AI inference, we expand the boundary to include the full AI system lifecycle per Figure 14 (i.e., cradle-to-grave approach), as we can determine the ultimate function of the model, such as generating a video.

Here we are primarily interested in inference for video generation and propose the functional unit below, based on emissions per video at a specified duration, resolution, frame rate and number of denoising steps. This approach has been chosen for better consistency and comparability across models.

Functional unit (per inference): Lifecycle emissions (inception to retirement) per 5-second video at 720p resolution, 30 frames per second and 50 denoising steps.

We also propose that model providers specify a range of functional units, shown below, so that users can understand the relative impact incurred by changing parameters:

- Lifecycle emissions (inception to retirement) per 5-second video at 720p resolution, 30 frames per second and 50 denoising steps
- Lifecycle emissions (inception to retirement) per 5-second video at 1080p resolution, 30 frames per second and 50 denoising steps
- Lifecycle emissions (inception to retirement) per 5-second video at 2160p resolution, 30 frames per second and 50 denoising steps

We propose this approach, rather than a per token approach, because we expect that the relationship between resolution and token and duration and token to be quadratic, rather than linear (Delavande, Pierrard, & Luccioni, 2025), and for clarity so that users can understand the context of the resulting carbon impact values. Additionally, our evaluation of alternative metrics, such as emissions per megapixel-frame, did not result in the desired outcome in terms of relationship to the underlying energy consumption of the IT hardware and comparability across resolutions and video durations (see Section 5 for sensitivity analysis on this topic).

[22] In the product carbon footprinting world, this is referred to as a cradle-to-gate footprint.

This functional unit represents our view of the best balance between accuracy and feasibility in application. However, we acknowledge that the number of denoising steps, while critical to calculating emissions, may not be easily accessible to the user. This is an area where increased transparency from model vendors can help through publishing the standard number of steps or, better yet, providing flexibility to change this parameter.

Video generation models which also generate audio

Some video generation models also produce audio with the resulting video output. In these cases, we recommend the functional unit above, however it should be made clear in the functional unit that audio is included, and the characteristics of the audio should be defined as well. We expect that video generation is significantly more energy and carbon intensive than audio generation (see Figure 6), hence the selection of functional unit best suited for video generation. When evaluating the carbon impact of audio generation specifically, a suitable functional unit should be selected that is appropriately linked to the audio generation model's architecture

For full transparency of system level impacts, we also propose a functional unit relating to the absolute emissions of the AI system, from inception to retirement, similar to AFNOR Spec 2314. Model developers/providers should concurrently publish this figure at time of launch and update it periodically if impacts from ongoing use of the AI system vary significantly.

Functional unit (AI system): Lifecycle emissions (inception to retirement) to provide the AI system for a specified number of videos representative of expected user patterns over a specified lifetime of the model.

6.2.5. Functional unit for generated video asset

Finally, we evaluate the selection of functional unit for a generated video asset. This will be highly dependent on the specific use of AI video generation and the form of the final output and should be considered carefully on a case-by-case basis. We also note the further expanded boundary, as indicated in Figure 15, which includes the total quantity of videos generated and post-processing required for a production ready asset.

We propose some general examples based on applications from Table 1.

Functional unit (reshoots and visual effects): Lifecycle emissions per scene for a specified scene type (e.g., aerial shot, car chase, etc.) and level of quality (i.e., resolution, framerate, duration, number of shots)

Functional unit (video generated ad): Lifecycle emissions per video advertisement at a specified level of quality (i.e., resolution, framerate and duration)

6.3. Allocation approach

As with any product that relies on shared processes, there must be a way to allocate, or apportion, shared processes to individual products as part of the carbon footprinting approach.[23] This is certainly true for generative AI systems and services, which rely on shared computational hardware and data centre infrastructure.

6.3.1. Allocation approaches established in existing literature and methodologies

Allocation is a key aspect of product carbon footprinting and it is not a trivial task to determine the appropriate allocation approach, as the selected approach can significantly affect the carbon impact of the product being studied.

The fundamental principle for determining an allocation approach is that it should be based on the underlying physical relationships between the product and system. In other words, the allocation metric should relate the cause and effect between product and system level impacts.

> Example: To determine the carbon impact for a pair of jeans we must understand the emissions attributable to the input materials, such as fabric. Through allocation, the total emissions from the fabric mill can be allocated by mass to the fabric produced in the mill, since we understand that the final output mass of fabric produced in the mill heavily influences the inputs, such as raw material and energy, used in the milling process (Greenhouse Gas Protocol, 2011).

AFNOR Spec 2314 covers this topic to establish allocation rules in alignment with the recommendations of the Digital Services PCR, where support services are allocated with a top-down method using a Power Usage Effectiveness (PUE)[24] uplift factor. Green Software Foundation establishes an allocation approach for embodied carbon within the Software Carbon Intensity specification, allocated with both a time and resource share.

AWS's recent Customer Carbon Footprint Methodology v3.0 establishes a comprehensive multi-tier allocation approach beginning with cluster-level emissions which are aggregated and allocated to rack-level emissions based on planned power draw, which in turn are allocated to specific cloud services with a usage-based allocation method and finally allocated to customer accounts based on physical allocation, where possible, and economic allocation as a fallback (Amazon Sustainability, 2025).

6.3.2. Allocating emissions to the functional unit level for training and inference

Allocation of emissions to AI model training

We first evaluate and propose an approach to allocate AI system emissions to model training, as shown in Table 4, following the schematic in Figure 14 and unit of analysis defined in Section 6.2.3. This

[23] This concept is covered in detail in the GHG Protocol Product Standard, Chapter 9.

[24] PUE is a metric used to measure data centre efficiency by relating the overhead electricity consumption of a data centre (i.e., from support services like MEP) to its IT electricity consumption

approach is proposed primarily for model developers or data centre operators, such as hyperscalers, who have access to detailed data related to IT and support services energy consumption.

Table 4. Allocation of emissions to AI model training

Emission source	AI model training
Operational (IT hardware)	Direct measurement of IT hardware energy consumption (including IT idle capacity) based on training time (no allocation required)
Operational (Support services)	Physical allocation of support services energy consumption by applying a PUE uplift on energy consumed by IT hardware
Embodied (IT hardware)	Physical allocation of amortised embodied emissions based on training time
Embodied (Building structure, support services)	Physical allocation of amortised embodied emissions based on planned power draw and provisioned time

Operational emissions of IT hardware should be captured through direct measurement of the IT hardware used for training, where possible, and the operational emissions of support services are accounted for by application of PUE of the data centre. The embodied emissions are first amortised, then allocated based on training time for IT hardware.

The amortised embodied emissions for building structure and support services are first allocated to the rack level using planned power draw relative to the building or cluster's total power capacity and then allocated from rack level to training based on provisioned time.

Table 5. Amortisation approach for embodied emissions of IT hardware and building infrastructure

Amortisation of embodied emissions of IT hardware The intent is to capture all embodied emissions associated with IT hardware and allocate them to the services that use that hardware over the lifetime of the hardware, so that no embodied emissions are left unallocated. The proposed approach amortises the embodied emissions of the IT hardware over the useful life of the hardware. For example, embodied emissions of the GPUs are amortised by GPU-hours over the 2-to-5-year life of the GPU[25]. The amortised emissions are then allocated with a time-based allocation.
Amortisation of embodied emissions of building structure and MEP *Amortisation and allocation of embodied emissions of the building structure and MEP is a complex topic and likely to be minimally material in many cases (see Table 3). It is therefore acceptable to rely on estimated proxies developed per inference or per token in the near-term.* Following a similar approach to the amortisation of IT hardware, the embodied emissions of building structure and MEP are amortised over the useful life of the building structure and MEP equipment, where the useful life is consistent with conventions established by financial depreciation schedules (typically around 30 years for building structure and 5-15 years for MEP).

Allocation of AI system emissions to AI inference by lifecycle stage

We next evaluate and propose an allocation approach of AI system emissions to inference, as shown in Table 6. This table is presented as a hierarchy, as we acknowledge that capabilities to use the proposed techniques will depend on the access to and availability of necessary data.

Allocation of model creation emissions to AI inference

To allocate training emissions to inference, we propose the approaches as shown in Table 6,[26] with option 1a being most desirable and option 2b being least desirable.

For model developers, providers and data centre operators, we strongly suggest option 1 as the preferred option and we expect that the necessary total model creation emissions data will be attainable. We further explore options 1a and 1b in Appendix 2 and conclude that an allocation of model training emissions per unit of energy related to inference is a stronger choice, principally because we expect inference to be the most material lifecycle stage and as the intended audience of these impact figures

[25] The GPU life here is illustrative and many vary depending on accounting practices and real-world use

[26] Tokens were the first option considered for allocating training emissions to individual prompts, due to their wide use in text-based models to approximate computational effort. They are not, however, well suited to video generation. For diffusion-based video models, model configuration factors — particularly the number of denoising steps — are significantly more influential than token count in determining inference energy (Li, 2024). We also note the non-linear relationship between token count and energy consumption as a weakness in this metric (Delavande, Pierrard, & Luccioni, 2025).

are ultimately the users, this approach does a better job of guiding users' decision making towards lower impact outcomes.

For *model users*, the approach selected will ultimately depend on the availability of data and will be highly dependent on the transparency of data from model developers and model providers. In the absence of data required to use option 1, we suggest a proxy-based approach using established estimates of model training emissions per unit of energy related to inference or per inference allocation. Currently, we are not aware of any such proxies that exist for video generation specifically, so we recognise this as an opportunity for ongoing exploration.

Table 6. Allocation of AI system emissions to AI inference

<table>
<tr><th rowspan="2">Emission source</th><th colspan="4">AI system lifecycle stages</th></tr>
<tr><th>Model creation (inc. training)</th><th colspan="2">Model operating lifetime</th><th>Model end-of-life (e.g., retirement and decommissioning)</th></tr>
<tr><th></th><th></th><th>Inference</th><th>Ongoing re-training and re-testing</th><th></th></tr>
<tr><td>Operational (IT hardware)</td><td rowspan="3">1. Allocate measured or estimated total model creation emissions

a) per unit of energy related to inference*

b) per inference*

2. Using established proxies to estimate model creation emissions

a) per unit of energy related to inference

b) per inference</td><td>1. Per inference direct measurement**
2. Per inference prediction model**
3. Measured total inference energy consumption allocated per MP-frame
4. Established proxies (e.g., per MP-frame proxy)</td><td rowspan="3">1. Allocate measured or estimated total re-training and re-testing emissions

a) per unit of energy related to inference*

b) per inference*

2. Using established proxies to estimate ongoing retraining and re-testing emissions

a) per unit of energy related to inference

b) per inference</td><td rowspan="3">Allocated per inference*</td></tr>
<tr><td>Operational (Support services)</td><td>Apply PUE uplift</td></tr>
<tr><td>Embodied (IT hardware)</td><td>1) Amortisation and allocation based on provisioned time
2) Allocate using established proxies</td></tr>
</table>

Embodied (Building structure, support services)		1) Amortisation and allocation based on planned power draw 2) Allocate using established proxies		

**Based on share of historical, forecasted or estimated figures for total inference energy consumption or total number of inferences over model lifetime.*

***No allocation required as energy consumption is measured or estimated per inference*

Allocation of inference-related emission sources to AI inference

For the inference lifecycle stage, specifically operational emissions from IT hardware, we propose four methods to determine IT hardware energy consumption, presented from most-desirable to least-desirable, alongside the recommended practitioner type for each method.

Table 7. Hierarchy to determine IT hardware energy consumption per inference

Method	Recommended practitioner type
1. Per inference direct measurement of IT hardware energy consumption	Data centre operators
2. Bottom-up prediction model which estimates energy consumption per inference	Model providers
3. Direct measurement of inference-related IT hardware consumption during a defined operating period (such as a year) and allocation based on suitable metric (e.g., megapixel-frames)	Data centre operators (if option 1 is not practical)
4. Use established proxies (e.g., energy consumed per MP-frame proxy or per second of video at specified resolution)	Model users
Note on model orchestration and monitoring: Importantly, the energy consumption captured here primarily focuses on the video generation model itself but should also include all relevant IT energy consumption related to 'orchestrating' and monitoring the model, such as those involved in prompt screening and filtering for security and reliability purposes. We do not propose in specific detail how to allocate various components of the prompt processing pipeline as we need further engagement and discussion with model providers and data centre operators on this topic to gain a clearer understanding of this process. In principle, a similar approach should be followed as described in Table 7.	

Once the per inference energy consumption has been determined, then the energy consumption of support services is allocated by applying a PUE uplift. Emissions are then determined by applying the appropriate emission factor for electricity.

For embodied carbon emissions of IT hardware, building structure and support services, we propose one of two methods:

1) Amortisation and allocation approach
2) Use of established proxies

We recognise that the amortisation and allocation approach will likely be most suitable for data centre operators who have better visibility of the IT and structural infrastructure required to determine embodied carbon emissions. For model providers and model users, who do not have access to this information, they should rely on use of established proxies per video, as these become developed. Currently, we are not aware of any such proxies that exist for video generation specifically, so we recognise this as an opportunity for ongoing exploration.

Allocation of ongoing re-training and re-testing emission sources to AI inference

This is an area that we expect continued evolution as models evolve both in how they are trained and how they are operated. We propose an allocation approach aligned with that which is established for model creation and thus requires an estimation of projected energy consumption from re-training and re-testing over the lifetime of the model. This estimation can be refined with historical data throughout the model's lifetime. We also highlight this as an area for further exploration and discussion with model providers and data centre operators to better understand how this field is evolving and refine the approach.

Allocation of model end-of-life emissions sources to AI inference

For model end-of-life emissions, we suggest a simplified allocation per inference. We do not expect this lifecycle stage to be particularly material, compared to model inference and model training.

7. Case study: AI video generation in a visual effects production process

This section analyses a real-world application of AI video generation as a case study, in line with a core objective of this report to contextualise AI emissions by comparing AI solutions to counterfactual non-AI solutions, with specific relevance for the digital media industry.

The use of AI video generation in video production is varied and rapidly evolving, so this case study aims to demonstrate how the carbon impact methodology explored in Section 6 could be applied to evaluate the lifecycle emissions to produce a scene and could be used as an assessment tool when evaluating different production approaches.

Furthermore, this case study provides a view of the lifecycle emissions for one specific use case and as each production will have its own unique requirements, the results are not indicative of the use of AI video generation in production generally.

7.1. Learnings from existing production-related emissions studies

Several relevant emissions studies (shown in Table 8) were reviewed to identify learnings that should be applied to our own analysis, with the following takeaways.

Table 8: Relevant industry case studies reviewed

Study Title	Author
Comparison of GHG Emissions from Scenes of On-Location and Virtual Productions	**ICF International, for Sony Pictures** (ICF, 2022) (ICF, 2023)
Virtual vs. Conventional Production for Film and Television: A Comparative Life Cycle Assessment	**Institute of the Environment and Sustainability (IoES) at UCLA, for Sustainable Production Alliance** (IoES, 2023)
Impact of Growth of Data Centres on Energy Consumption	**Europe Economics, for Department for Energy Security and Net Zero (DESNZ)** (Europe Economics, 2025)

1. **The baseline should reflect current production practices**

Any comparisons should be made against production methods that reflect current practices, not against outdated methods. The use of virtual production, e.g. LED stages with computer-generated backgrounds, is now a commonly used method in digital media production. Similarly, contemporary techniques should be considered as counterfactuals where they represent normal industry practice (i.e. in cases where production budgets and logistical constraints would typically result in using virtual production).

2. **Performing scenario analysis strengthens transparency and credibility**

Scenario analysis helps assess uncertainty and test assumptions to get a more comprehensive view of study outcomes. It helps strengthen conclusions by considering a range of plausible outcomes.

Assessing the variation and uncertainty of parameters in both scenarios provides a transparent view on the level of confidence in findings of the study. It also guides the study, and those seeking to build on it, on where to focus future assessments.

3. **Assessment boundaries should be set carefully and consider indirect emissions where appropriate**

In Section 6 we set out the elements that should be included in the boundary when assessing generative AI emissions, including direct emissions from the chips operating the model as well as indirect emissions from support services. Likewise, relevant indirect emissions need to be included in the boundary for the counterfactual scenario excluding the use of generative AI.

7.2. Analytical approach used for the case study

7.2.1. General principles

To ensure a credible assessment between a production approach using generative AI and a contemporary production approach (also referred to as the counterfactual), we have considered the principles of a fair comparison between product carbon footprints, such as those in The Greenhouse Gas Protocol Product Standard, Appendix A (WRI & WBCSD, 2013).

General principles for product carbon footprint comparison used in the case study

- The unit of analysis should be identical
- The system boundaries and temporal boundary should be equivalent
- The same allocation methods should be used for similar processes
- The data types used and the data quality and uncertainty of data should be reported and assessed to determine if a fair comparison can be made
- The temporal and geographical representativeness of the inventories should be assessed to determine if a fair comparison can be made

7.2.2. Interpretation and limitations of the study

How this study should be used and interpreted	• This study is used to **provide context for current real-world applications of AI video generation in media production**, including how the selection of VFX approach is considered among alternative approaches • This study **provides an estimate for a specific use case** of AI video generation in media production to generate background imagery for a scene • This study **demonstrates application of the methodology explored in Section 6** • This study represents the use of generative AI at **this moment in time**. The technology and its application are evolving

	rapidly so, in the future, a similar use case might be delivered differently, with different results.
How this study should *not* be used and interpreted	• This case study is **not a *general* representation of the carbon impact of AI video generation in media production,** as scenes, processes and creative requirements can vary significantly leading to different outcomes
Limitations of the study	• The case study **boundary does not evaluate physical production emissions before the VFX process begins** (these are equivalent processes, see boundary in Section 7.4.1) • **For cloud-based video generation, this case study relies on a single energy prediction model** from (Delavande, Pierrard, & Luccioni, 2025) • Assumptions used for the allocation of training emissions and energy intensity of inference **introduce uncertainty into the results and conclusions**

7.3. Selection of the case study

The use of generative AI in video production workflows is both wide ranging and fast evolving. In aiming to maximise the utility of the case study, we created four criteria to select the scene described below, which are summarised in Table 9.

CASE STUDY

Modification of the background for a fast-moving scene

For a streaming series, a fast-moving scene was filmed in a location with a visually simple background. Generative AI tools were used to change the background to a more complex setting that would have been logistically difficult to achieve during the physical shoot.

Table 9: Criteria used to select a generative AI case study

Criteria	Description	Rationale for selected case study
Representativeness	The case study should be representative of what is currently being done in the digital media industry	Background adaptation from a physical shoot was identified as a **representative** use case among DIMPACT companies
Credibility of comparison	The counterfactual approach should be a credible and plausible alternative	GenAI was used by the studio with a full VFX rendering approach

	to the selected AI video generation approach	considered as the back-up. This makes the comparison to a plausible counter-factual more **credible**
Data availability	Suitable data and production expertise should be available to inform the modelling approach	**Quantitative data was available** from the studio, along with the opportunity to speak with the VFX team directly
Study relevance	The use of AI video generation should be a material part of the overall production of the content	Use of AI video generation was identified as a **material part of the scene**

7.4. Methodological overview

Accounting methods referenced	Lifecycle carbon accounting approach referencing the GHG Protocol Product Standard and ISO 14067
Functional unit	Lifecycle emissions of the visual effects process per 'final pixel' scene (53-second scene @ 24 fps/2160p). *Note: final pixel refers to the finished video asset which will be seen by the viewer*
Boundary	Cradle-to-grave
Timeframe	Assumes emissions intensities representative of 2025
Geography	Scenarios assume production activities occur in France, the UK and the US[27]
Included activities and data	**Artist computers** (operational and embodied carbon of computer equipment and software), based on estimates of artist time required per shot provided by the VFX studio. **Generative AI models for image and video generation via cloud services** (including both operational and embodied carbon for training and inference). Image generation data was captured through the quantity of images used for the entire VFX process. Video generation data was captured through the total number of credits used for the entire VFX process. **Local compute from artist workstations** (operational and embodied carbon). For local video generation, the total actual electricity consumed by artist workstations was provided by the VFX studio. For the counterfactual using

[27] We note that it is common practice for productions to be shot in one location and post-production done elsewhere. Knowing the actual location of the work is critical to accurately reflect the local grid's carbon intensity.

	traditional VFX rendering, total estimated render hours and electricity consumption was provided based on two scenarios. **AI upscaling models** (including both operational and embodied carbon for training and inference). An estimation of the number of upscaled videos generated per shot was provided by the VFX studio.
Data quality and uncertainty	**Generative FX:** Data quality = Medium-to-High, Uncertainty = Medium-to-High **Traditional VFX:** Data quality = Medium, Uncertainty = Medium Further detail on data quality and uncertainty can be found in Appendix 5.
Exclusions	End-of-life assumed immaterial for IT hardware and software

7.4.1. Boundary of the study and visual effects processes

The processes that have been considered within the production of the scene start with the physical recording of the footage – the 'physical shoot' – and end with the final post-production step required to generate a 'final pixel' scene that will be seen by the viewer.

This production process is considered using two approaches:

1. **Generative effects (Generative FX or Gen FX):** Visual effects process after receipt of the 2D plates (the recorded content) conducted using a hybrid approach of generative AI technologies alongside certain traditional VFX methods and tools. This approach uses image generation, video generation and AI video upscaling, alongside traditional VFX methods (like compositing), to produce the final pixel scene.
2. **Traditional visual effects (Traditional VFX):** Visual effects process after receipt of the 2D plates, conducted using traditional visual effects methods with 3D rendering. This approach relies primarily on professional workstations to render a modelled 3D environment with high quality lighting, which then undergoes compositing and colour management to produce the final pixel scene.

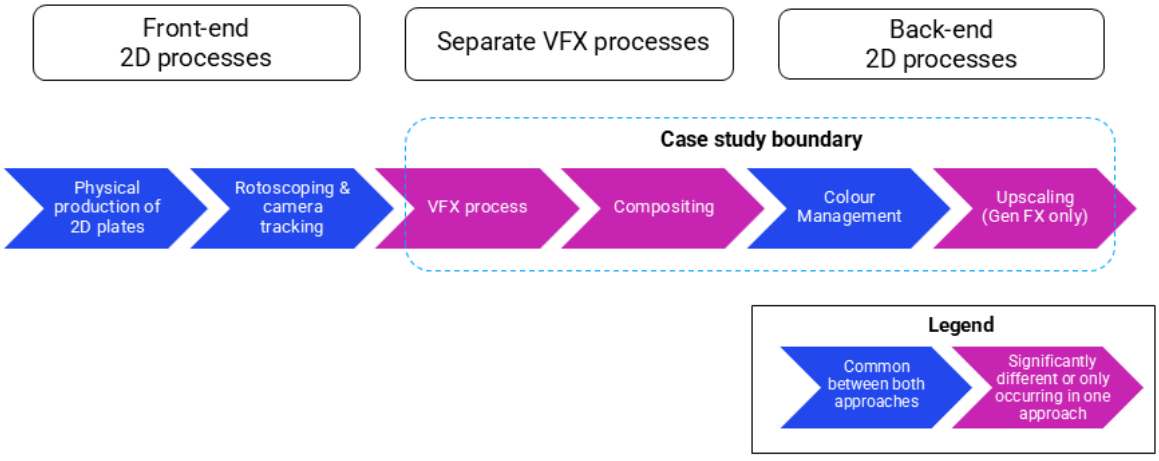


Figure 16: Process map of production of final pixel scene for the selected case study including the selected case study boundary

Through discussions with the VFX studio who performed the work, we established that there is little-to-no variation between front-end 2D processes between these two approaches. Therefore, we omitted these processes from the case study boundary. The boundary also excludes pre-production activities before the physical shoot, such as concept design and storyboarding.

One of the key differences between the two approaches is the management of video resolution throughout the process. In the generative effects process, generated videos are produced at 720p before undergoing AI upscaling to the final resolution of 2160p. In the traditional VFX process, rendering occurs natively at 2160p, so there is no need for upscaling.

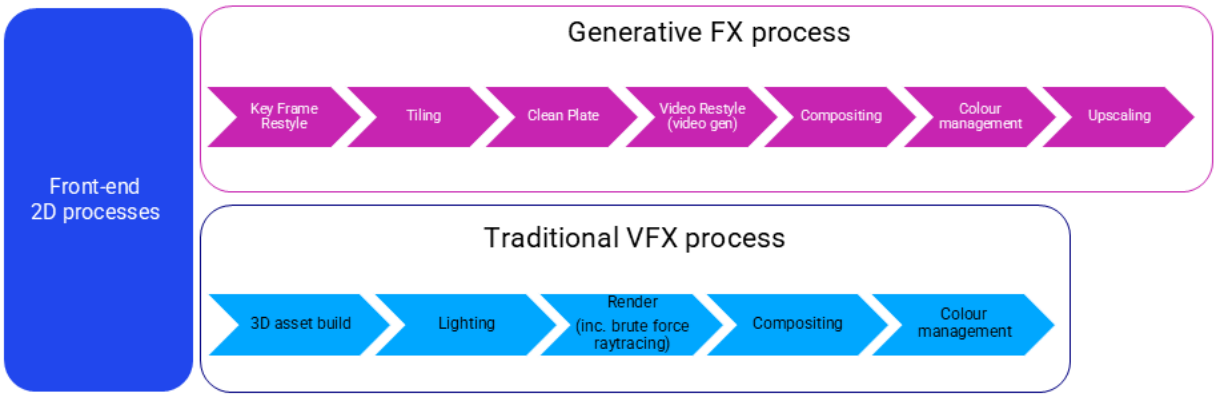


Figure 17: Detailed process map showing the key differences between the two approaches

7.4.2. Use of generative AI in the generative effects process

In the generative effects process, different forms of generative AI were used as outlined below:

- Image generation, via cloud services, during key frame restyle stage
- Video generation, via cloud services, during clean plate stage
- Video generation, run locally on the VFX studio's workstation computers, during video restyle
- Video upscaling, via cloud services, during the upscaling stage

Video generation which was conducted using a mix of third-party SaaS tools and the studio's own workstations. Initial storyboarding and imagery were created with off-the-shelf general tools due to their accessibility and ease of use. Over the duration of the project, the team shifted to using in-house infrastructure for more specialised tools for the more technical, fine-tuned end-product.

7.4.3. Key assumptions used in the analysis

Video generation	All videos were generated at 720p and 24 fps. Videos generated in the cloud are assumed to be 8 seconds in duration and are represented using the model developed for the sensitivity analysis in Section 5 (see Appendix 3 for further detail). Videos generated locally are represented using measured electricity consumption of workstations, with a training and embodied carbon uplift applied.

	Conversion from cloud-services 'credits' to seconds used for video generation assumes a conversation rate of 15 credits per second of video based on Runway ML API pricing.
Training uplift	A training uplift is applied to all cloud-based generative AI (including upscaling) in the case study, ranging from 10% of inference (low) to 100% of inference (high), consistent with the sensitivity analysis performed in Section 5. A training uplift is applied to energy used by locally run video generation models, based on low, medium and high assumptions of total model training emissions and total model inference by user population (see Appendix 3 for figures used).
Video upscaling	Video upscaling is estimated using a carbon intensity estimate per MP-second. This assumes Crystal Video model pricing of $0.1/MP-second at 30 fps, a conversion to equivalent tokens (based on pricing between $5 and $15 per 1M tokens) and applying token-based estimates of energy consumption (0.35 Wh to 0.70 Wh per 1K tokens).
Local compute	For local compute associated with video generation and rendering, we assumed the use of local workstations[28]
Embodied carbon of workstation and artist computers	The embodied carbon of workstation and artist computers is estimated based on workstation specifications provided by the VFX studio and allocated based on hours of use. Workstation lifetime is assumed to be 5 years with an active usage rate of 50%
Electrical grid intensity	Scenarios assume France, UK and US average electrical grids
Software development	Software development emissions are included in emissions from artist computers
Emissions intensities	A list of emissions intensities used in the analysis is included in the Appendix

[28] In other cases, it is plausible that this compute may be performed via local or third-party render farms in a separate geographical location

7.5. Case study results

The results of the case study analysis are presented below in Figure 18 where we evaluated two generative FX scenarios and one traditional VFX scenario using average the emissions intensity of electricity in France, the UK and the US. The US grid is a useful reference point as many AI data centres are concentrated in the United States; however, we also evaluate other grids to show the relative effect of different grid regions and emissions intensities on the carbon impact.

We see that in the base case for the generative FX approach, an emissions range of approximately 900 to 2,100 $kgCO_2e$ per final pixel scene is estimated using the US average grid.

In the generative FX scenario studied for this case study, the project team was relatively efficient with its iteration process generating an estimated 2,137 videos (we refer to this as the base case). Under other circumstances,[29] it is possible that the project team could have used generative FX to iterate many more times and increase the total impact for the project, which is represented by the higher iterations scenario, with approximately three times the number of videos generated relative to the base case[30] (6,525 videos generated in the higher iterations scenario). In the higher iterations scenario for the generative FX approach using the US average grid, emissions increase to a range of approximately 2,000 to 5,500 $kgCO_2e$.

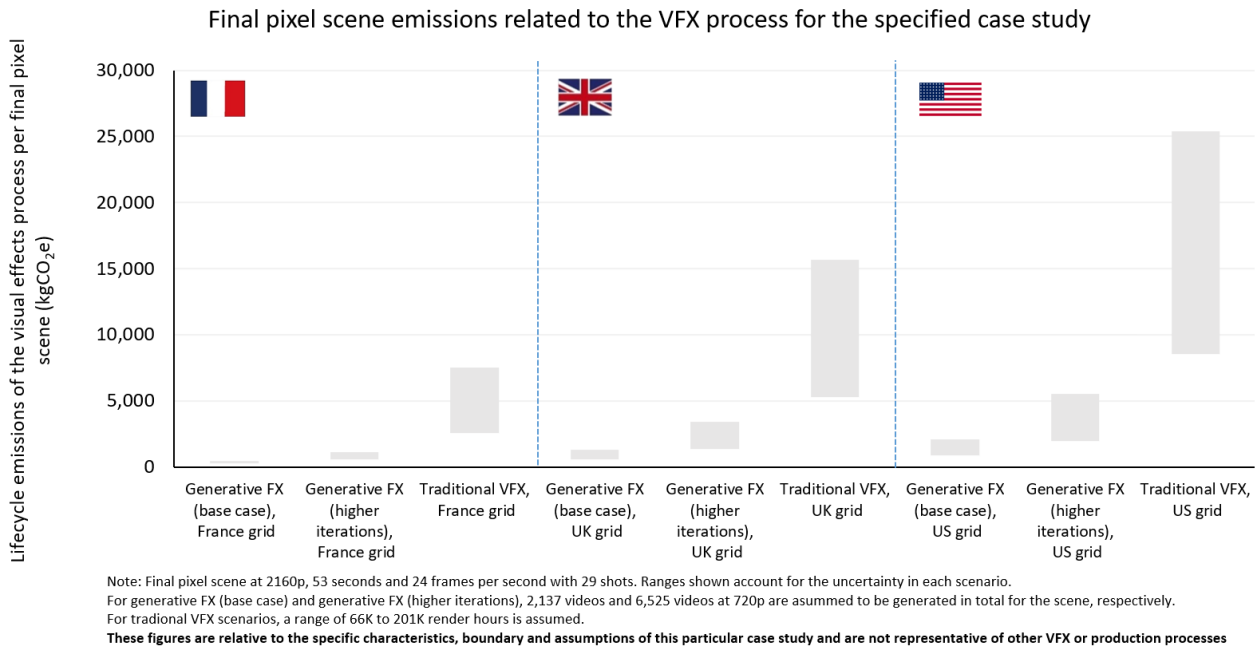


Figure 18. Final pixel scene emissions related to the VFX processes evaluated in the case study

The traditional VFX scenario in this analysis was calculated from historical data at the studio from recent, comparable productions that had actual VFX shot data for final projects. This was then used to

[29] It is worth noting that in some cases, there is a minimum number of 'iterations' contractually mandated to accommodate feedback from producers, studios and other parties. It would be interesting to explore opportunities to streamline this process and reflect that contractually.

[30] This is consistent with the range of rendering hours evaluated for the traditional VFX approach

determine a plausible range of total render hours required, which is assumed to be approximately 66,000 to 201,000 hours. The large range in the traditional VFX scenario comes from both the human variable in production that drives the number of iterations for a project (i.e., a producer asking for a change to a shot requiring it to be re-rendered) and from the relatively longer compute time required for VFX iterations compared to generative AI iterations.

The traditional VFX scenario shows an estimated range between approximately 8,500 and 25,000 $kgCO_2e$, which is nearly entirely driven by the significant amount of workstation compute required to render the scene elements in 3D, which can be seen in Table 10.

Table 10. Videos generated and workstation compute associated with each scenario

Activity	Generative FX (base case)	Generative FX (higher iterations)	Traditional VFX (low range)	Traditional VFX (high range)	Unit
Videos generated (cloud)	687	2,061			Number of videos
Videos generated (local)	1,450	4,350			Number of videos
Total videos generated	**2,137**	**6,411**			**Number of videos**
Upscaling videos	87	87			Number of videos
Local workstation electricity	794	2,381	16,791	51,087	kWh
Local workstation hours	2,935	8,805	66,145	201,248	Compute hours

Importantly, we also see that the emissions intensity of electricity has a pronounced effect on the resulting emissions across all scenarios, illustrating the importance of using electricity from renewable sources and from grids with lower average emissions intensities.

When looking at the breakdown in emissions by activity, as shown in Figure 19, we see that video generation performed on local workstations is the largest source of emissions, which is expected since the majority of videos generated during the generative FX process were generated locally (approximately two times the number of videos were generated per shot on local workstations as were generated using cloud services).

Upscaling is comparatively small, as expected, since the number of videos required for the upscaling process is significantly less at approximately three videos per shot in the scene. There is uncertainty regarding the emissions intensity of machine learning (ML) and AI upscaling models, but we expect it to be comparatively lighter than AI video generation given that the model only needs to 'upscale' using existing video information, rather than completely generate video from a prompt.

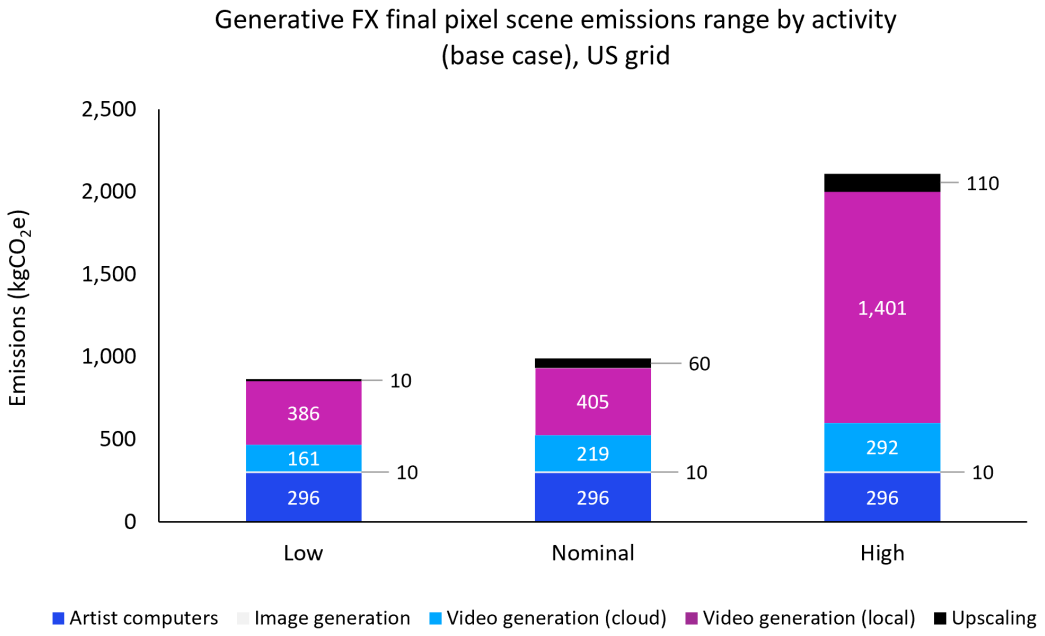


Figure 19. Generative FX final pixel scene emissions range by activity (base case), US grid

We also see that there is a relatively wide range of uncertainty in the estimates when viewed from low to high, which is primarily a reflection of the uncertainty around training emissions estimation and allocation, as seen in Figure 20, where nominally training accounts for 12% of total scene emissions, rising up to 57% at the high end of the uncertainty range.

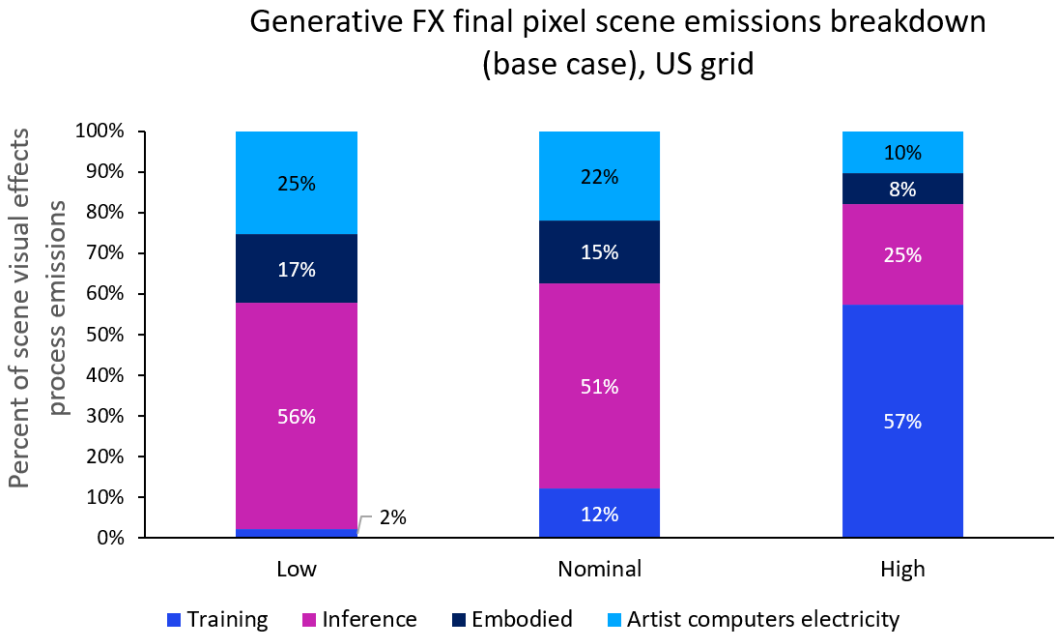


Figure 20. Generative FX final pixel scene emissions breakdown (base case), US grid

Finally, we also see that embodied carbon is a significant element of the total scene emissions across both generative FX and traditional VFX approaches (shown in Figure 21).

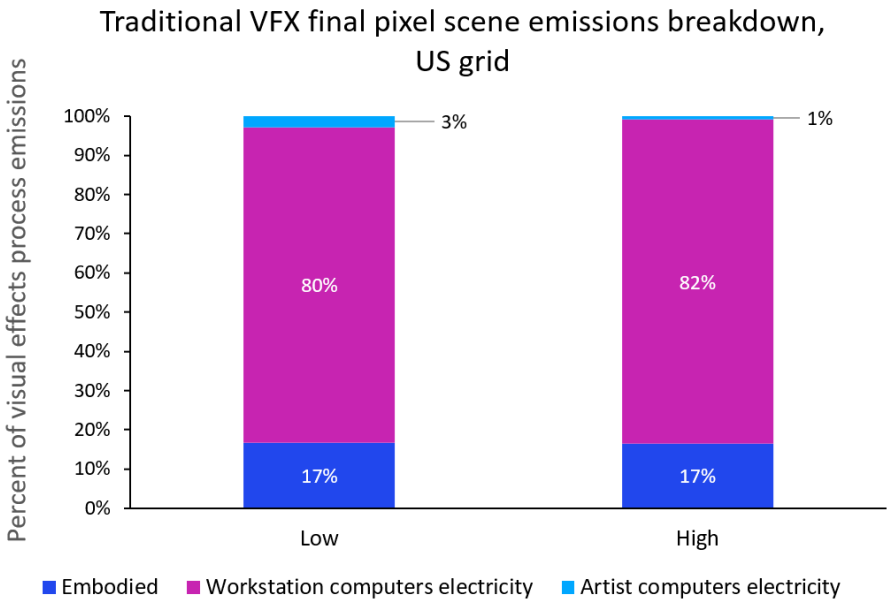


Figure 21. Traditional VFX final pixel scene emissions breakdown, US grid

7.6. Case study findings and discussion

In assessing this case study, we have demonstrated an application of the methodology proposed in Section 6 to understand its utility and limitations, while providing context for how production studios are using and adopting generative AI tools and production methods. Below we discuss our findings and interpretation of the study.

The uncertainty due to the estimation of training emissions and inference for closed models is significant and further demonstrates the need for a consistent measurement and reporting approach through a mechanism such as Product Category Rules

We see that relying on assumptions for the emissions associated with model training can have a meaningful effect on the results of the study, with emissions from video generation nearly tripling between low and high uncertainty bounds. This uncertainty (which is both a result of the estimation method and boundary associated with measuring training emissions, as well as how these emissions are allocated) makes it difficult to confidently evaluate generative AI services, tools and production approaches, further underscoring the importance of having consistent and transparent methods to measure and report the carbon impact of AI video generation and other generative AI technologies.

Careful use of AI video generation in the production process may *potentially* reduce energy consumption and carbon impact

In this case study, we've seen that the use of generative AI on local workstations could plausibly reduce the workstation energy required for the scene relative to the estimated workstation energy for traditional 3D rendering, demonstrating the potential for energy and carbon efficiency improvements under the right conditions. However, these results are not indicative of all scenarios and use cases and can't be generally applied to other uses.

To illustrate this point, consider Figure 22, where a production team located in the UK might consider using cloud-based AI video generation at 1080p (via US-based data centres) as an alternative to a traditional VFX approach using their local workstations in the UK. In this case, given the uncertainty around the estimations, the two approaches are not significantly different in terms of carbon impact for the final pixel scene, as a result of the reduced carbon intensity of the UK electrical grid relative to the US electrical grid, and the use of higher resolution video generation.

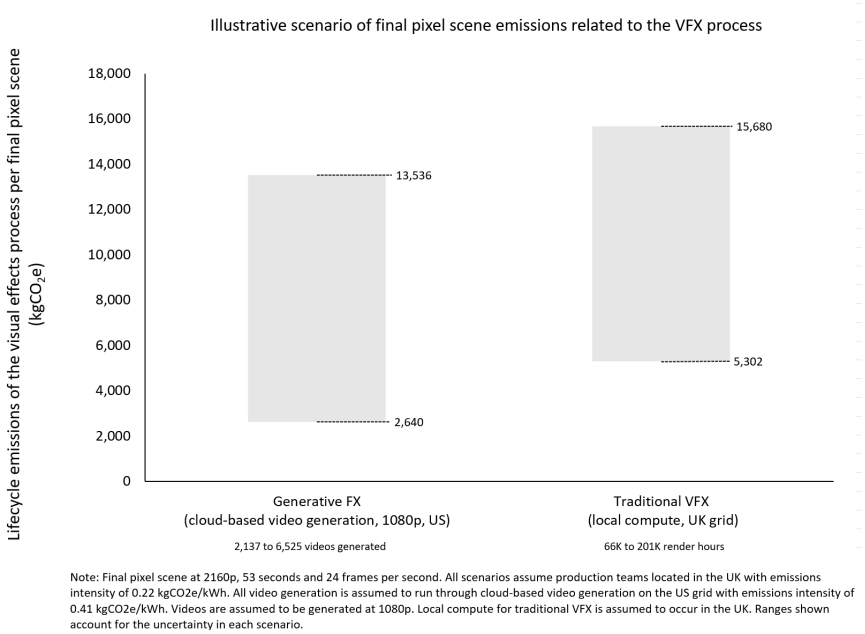


Figure 22. Illustrative scenario of final pixel scene emissions related to the VFX process

Furthermore, some applications of generative AI may potentially reduce physical production-related emissions, particularly if specialised equipment is involved, such as for aerial shots or those involving significant travel. Conversely, there are scenarios where generative AI is creating new possibilities entirely, where emissions from its use could be fully additional against a counterfactual scenario that never existed. Examples of this include the generation of realistic historical scenes for unscripted documentaries, which would not have had a plausible counterfactual given the production budget.

In other cases, traditional methods may simply be more efficient than newer generative ones. Each use case needs to be considered on its own merits against what would have realistically happened otherwise.

Local compute can give production studios a clearer view of the energy used for inference

In the case study, the production studio used open-source video generation models on local workstations for the bulk of the AI video generation. Interestingly, this allowed the production team to capture actual energy consumption required to perform inference and generate the videos, allowing for better monitoring and management of energy and carbon impacts as a result.

The use of upscaling in production of final-pixel scenes should be explored further

We also learned that by using upscaling (AI upscaling in this case) at the end of the production process, the production team was able to generate videos more efficiently at lower resolutions. Our analysis shows that this approach has potential for efficiency improvements relative to generating video at higher resolutions, but we relied on assumptions that introduce uncertainty into the conclusions.

Nevertheless, we view this as a potential area for further exploration to understand the trade-offs in the use of ML and AI upscalers paired with AI video generation.

In the absence of specific carbon impact data for AI video generation, we recommend studios collect the best data they can and call on AI model developers to bridge the current transparency gap

When models are run locally, production teams should monitor and collect data relating to energy usage and compute time. If running these models in a hyperscaler cloud environment, studios should review what data is already accessible through the customer carbon footprint portals provided by AWS, Azure, Google Cloud and others. In some cases, separate projects may need to be setup to track AI usage separately from other cloud services.

If using closed models such as Google Veo or Runway, there is likely to be limited visibility of energy or carbon emissions from model inference. To begin tracking and estimating these impacts, companies can consider using online leaderboards such as ML Leaderboard and AI Energy Score. The resulting figures will have a wide range of uncertainty, but can still be useful to begin to implement a measurement and monitoring system that can be refined over time as methods evolve.

To determine lifecycle impacts, additional assumptions and uplifts may be needed to account for model training, embodied emissions, idle capacity and non-IT hardware, depending on the data collected or on the methodology of cloud customer carbon footprint reports.

Finally, we suggest that DIMPACT companies and digital media industry companies request data from model providers directly, in line with the boundary in Figure 14 and proposed functional units in Section 6.2.4 to demonstrate demand for these metrics.

Production studios need guidance on best practice to help navigate this space

Given that we are still early in the evolution of generative AI technologies and tools, production studios are still exploring and experimenting with different approaches. Through discussions with production teams during the development of this report, we learned of other ways that generative AI is being used, which include digital set extensions developed with generative AI and sparse set restyling, which uses generative AI to recreate complex environments digitally.

We've shown that even with today's best methods, there is still uncertainty in estimating the carbon impact of their use. To help production teams navigate this space, we recommend the development of best practice guidance for the sustainable use of generative AI in production processes. This guidance should be grounded in comparable assessment approaches, both for generative AI use and for alternative production methods.[31]

As we have done in this report, we suggest knowledge sharing and peer learning to document and discuss use cases, examples and guardrails for the use of generative AI in digital media production, which considers environmental impact alongside other impacts in a holistic way.

[31] We note the forthcoming updates to the BAFTA Albert tool will be helpful in this regard.

7.7. Recommendations for media and production companies

Based on the case study, we make the following recommendations for media and production companies to better monitor and manage their carbon impact.

Measure what you can	
☐	Track local workstation electricity consumption and hours
☐	Leverage existing customer carbon footprint tools from cloud service providers where possible (which may require creating separate projects to track generative AI use specifically)
☐	Request data from model providers and cloud service providers directly, in line with the checklist in Section 8.2, boundary established in Figure 13 and proposed functional units in Section 6.2.4, to demonstrate demand for these metrics
☐	Use KPIs to track process efficiency, e.g., number of videos generated by model and resolution; total number of iterations required per shot and per final-pixel scene
☐	Account for training and embodied carbon impact, in addition to inference
☐	Account for uncertainty
Work with peers and industry groups such as BAFTA albert and the Sustainable Entertainment Alliance to align on reporting practices	
☐	Incorporate estimations of emissions associated with generative AI within wider 'production emissions' calculators
☐	Align on consistent data request content and format when requesting data from model providers
☐	Engage in discussion to establish consistent measurement methods through Product Category Rules
☐	Report using a location-based accounting approach Optionally report on a market-based emissions accounting approach
☐	Include a methodological statement for transparency, including: • To what extent these figures are measured or estimated • Measurement method, key assumptions and uncertainties • Any omissions

Use lower carbon sources of electricity	
☐	Use cloud services and render farms located in lower carbon grid regions, where possible
☐	Use renewable electricity for your own operations and workstations
Explore and share learnings around production efficiency	
☐	Share learnings around approaches for production efficiency, such as using smaller more efficient models for early iterations
☐	Use available energy and carbon data to share with production teams the effects of iterations on carbon impact
☐	Investigate the use of efficient machine learning and AI upscalers as a potential way to reduce the resolution of videos generated
Contribute to and support the development of guidance for the sustainable use of AI with peers	

8. Future trends and recommendations

AI video generation and generative AI overall are evolving rapidly. From the inception of this research to publication, the field has already changed significantly, with the growing use of agents and the mothballing of Sora. We are conscious therefore that this report represents a snapshot in time rather than an evergreen analysis. In this section, we explore how this disruptive new technology may evolve and provide recommendations on managing its carbon impacts.

8.1. Future trends impacting the growth and environmental impact of generative AI

How AI models are architected and trained is evolving quickly and may bring energy efficiency improvements

Advancements in model architectures are continuously evolving. We've already seen this in the development of video generation models built on diffusion transformer architectures. Considering the technical and economic constraints involved in deploying AI systems, it's plausible that new architectures that we're not yet aware of will be developed to improve video quality, performance and energy efficiency.

Training is also changing alongside model architectures. Industry practice is moving toward continuous retraining, with models updated iteratively over time rather than trained once. At the same time, techniques such as distillation allow smaller models to reduce the compute required to produce task-specific systems with higher-quality outputs (IBM, 2024). Finally, the physical location of training may change as advances in decentralised techniques are making it more feasible to spread the training load across geographies and take advantage of underutilised compute resources (IEEE Spectrum, 2026).

Future demand for generative AI will look different from today's demand and underscores the importance of regularly reviewing what 'best climate practice' looks like

Until recently, demand for generative AI was shaped largely by human-driven prompting and querying, yet new uses and applications are constantly being developed, such as reasoning models and agentic tasks. These new applications can be more energy-intensive, as is the case with long reasoning and agentic queries, which can increase energy consumption by an order of magnitude relative to short queries (Oviedo, et al., 2026). If usage trends change consistently toward more energy-intensive tasks, this could have ripple effects on aggregate energy demand.

At the same time, while we are seeing energy efficiency improvements per individual task (IEA, 2026d), these do not guarantee reduced system-level energy demand. Consistent with the Jevons' paradox, efficiency improvements can lead to more frequent and widespread use, enabling total AI-related energy consumption to grow even as models become more efficient. This mix of factors underscores the importance of monitoring outlooks for energy consumption and establishing guardrails to prevent unintended consequences.

Supply-side constraints and bottlenecks across the value chain are influencing the pace and nature of digital infrastructure rollout

Supply constraints are emerging across the AI value chain. A key challenge is access to IT equipment, where supply is struggling to keep up with demand. Most recently, constraints in the supply of high-bandwidth memory (HBM) have become a critical bottleneck in the supply chain and are expected to continue through the end of 2027. Frontier AI GPUs like Nvidia's Blackwell chips have a 12-month+ 'wait list' to purchase (TokenRing AI, 2025) while the 'big 4' hyperscalers are developing their own in-house alternatives to supplement or circumvent this supply.

As highlighted in the IEA's 2026 report, 'Key Questions on Energy and AI', access to electricity is also becoming a constraint for how quickly AI systems can be deployed, particularly as data centre operators seek to secure large amounts of reliable power for new AI data centres. Long queues for grid connections are increasingly pushing data centre operators towards natural-gas power generation to meet their electricity needs. This presents a risk that deployment of AI data centres will further increase emissions from the sector, yet also presents an opportunity for innovation if policymakers and investors can work out a path to put in place requirements and incentives to ensure that data centre electricity demand can be met with clean energy at the pace that it is growing.

We're in the venture-funded experimental phase of generative AI – stable business models will emerge in time

The deprecation of Sora by OpenAI earlier this year hinted that the company was unable to identify a sure route to monetising this technology and was unwilling to continue subsidising the high energy costs of delivering a response to a user prompt with reasonable latency.

As companies continue to experiment with these tools, it will become clearer where they add sufficient value to justify a fee covering the 'true cost'[32] of the service beyond the current investment-subsidised rate. This, in turn, will affect usage of these tools, with the potential that usage will decrease outside professional production companies if the cost pass-through makes it unaffordable for the 'casual' user.

Future AI-related energy demand is likely to come from local and edge devices, shifting where and how carbon impact can be mitigated

The carbon impact of AI technologies will increasingly depend on how they are deployed. This includes decisions about where compute resources are located and how they are powered, as the same AI workload can have very different emissions outcomes depending on local grid mix, access to low-carbon electricity, and cooling requirements (MIT Technology Review, 2025) and whether its run in the cloud, on the edge or on local devices.

As the orchestration layer becomes more sophisticated, carbon impacts will depend increasingly on how workloads are routed across geographies and systems

This reinforces the need to develop frameworks and methods to monitor and measure total system impacts and to be forward-looking in assessing the risks and opportunities presented by a shift in

[32] Meaning true 'financial' cost. Of course, there is also the externality of the environmental cost which is not currently reflected in the financial cost – but there is insufficient space to discuss that topic here.

demand. Product Category Rules will need to consider defining a boundary that encompasses wherever the inference compute takes place (i.e., potential edge nodes or on-device). This will allow standardised measurement and support efforts to optimise across the whole system.

Recommendations to address the carbon impact of AI video generation

Taken together, the above trends suggest that the future carbon impact of AI video generation, specifically, and AI technologies, generally, will be shaped as much by patterns of use, infrastructure constraints and deployment choices as by improvements in efficiency. This reinforces the need for close coordination across value chain participants, policymakers, and investors to monitor these dynamics and to shape interventions which nudge the trajectory of the overall system in a greener direction.

We therefore make the following recommendations, supported by checklists in Section 8.2, to guide industry stakeholders on the actions they can take towards increased transparency.

- **For the technology sector and LCA community**: We recommend constructive dialogue and consultation with the AI technology sector to develop a consistent approach in the form of product category rules (PCR) for AI technologies, such as video generation. This report provides a methodological foundation to build on and test with stakeholders, and we invite continued engagement on developing this further, with the aim of improving transparency of carbon impacts of generative AI. We've engaged with prospective partners who could participate in this process and advocated for the development of PCRs for AI services.
- **For the digital media industry:** We recommend the development of best practice guidance for the sustainable use of generative AI in production processes, which considers impacts from a holistic perspective, including environmental impacts. This should build on the case study findings in Section 7.6. As we have done in this report, we suggest knowledge sharing and peer learning to document and discuss use cases, examples and guardrails for the use of generative AI in digital media production. As noted at the end of Section 7, we recommend developing standardised ways of requesting data from model providers and methodologies for including these emissions within wider carbon accounting estimation tools for production.

8.2. Checklists for industry stakeholders to drive transparency

To close existing gaps and enhance comparability, we propose strengthening the transparency and disclosure of the carbon impact of generative AI models for model developers as well as model providers and data centre operators. These recommendations have been summarised into checklists that these organisations can begin to follow with a view towards improving measurement methodologies and transparency in carbon emissions measurement and reporting.

Reporting and transparency checklist for model developers	
☐	**Engage in discussion to establish consistent measurement methods through Product Category Rules**
☐	**Report energy use and emissions transparently and consistently on model cards**
☐	Follow the GHG Protocol accounting principles
☐	Report using a location-based accounting approach
☐	Optionally report on a market-based emissions accounting approach
☐	**Include emissions related to the full AI system lifecycle, by lifecycle stage:**
☐	Initial training (including experimentation)
☐	Inference (including operation and monitoring)
☐	Ongoing re-training
☐	Both operational and embodied emissions of all lifecycle stages
☐	**Include energy and emissions specific to inference for key functions of the model (e.g., text generation, coding, video generation)**
☐	**Include a methodological statement for transparency, including:**
☐	To what extent these figures are measured or estimated
☐	Measurement method, key assumptions and uncertainties
☐	Any omissions

Reporting and transparency checklist for model providers and data centre operators	
☐	**Engage in discussion to establish consistent measurement methods through Product Category Rules**
☐	**Allow cloud customers to see AI-related usage as a separate reporting category in customer carbon footprint tools**
☐	Incorporate best practice measurement methods as established by PCRs
☐	Include emissions related to the full AI system lifecycle
☐	Report using a location-based accounting approach
☐	Optionally report on a market-based emissions accounting approach
☐	**Provide guidance to customers on what and how they can currently track this data from existing carbon footprint tools**
☐	**Include a methodological statement for transparency, including:**
☐	To what extent these figures are measured or estimated
☐	Measurement method, key assumptions and uncertainties
☐	Any omissions

References

AFNOR. (2024). *Spec 2314: General framework for Frugal AI.* La Plaine Saint-Denis Cedex: AFNOR.

AI Energy Score. (2026). Retrieved from Hugging Face: https://huggingface.github.io/AIEnergyScore/#faq

Altman, S. (2025). *The Gentle Singularity*. Retrieved from https://blog.samaltman.com/the-gentle-singularity

Amazon Sustainability. (2025). *AWS Customer Carbon Footprint Methodology v3.0.* Retrieved from AWS: https://sustainability.aboutamazon.com/aws-customer-carbon-footprint-tool-methodology.pdf

Arena. (2026). *Leaderboard*. Retrieved from Arena: https://arena.ai/leaderboard

ARUP. (2025). *Circular thinking for data centres.* Retrieved from ARUP: https://www.arup.com/insights/circular-thinking-for-data-centres/

Axios. (2025). *America's data center growth hot spots, mapped*. Retrieved from Axios: https://www.axios.com/2025/12/18/data-center-growth-map-states

BBC. (2025). *OpenAI video app Sora hits 1 million downloads faster than ChatGPT*. Retrieved from BBC: https://www.bbc.com/news/articles/crkjgrvg6z4o

BBC. (2026, March 25). *OpenAI closes Sora video-making app and cancels $1bn Disney deal*. Retrieved from BBC: https://www.bbc.com/news/articles/c3w3e467ewqo

Caravaca, F., Cuevas, A., & Cuevas, R. (2025). *From Prompts to Power: Measuring the Energy Footprint of LLM Inference.*

Carbon Trust. (2025, August 28). *The renewable route for data centre expansion*. Retrieved from Carbon Trust: https://www.carbontrust.com/news-and-insights/insights/the-renewable-route-for-data-centre-expansion

Delavande, J., Pierrard, R., & Luccioni, S. (2025). *Video Killed the Energy Budget: Characterizing the Latency and Power Regimes of.*

Elsworth, C., Huang, K., Patterson, D., Schneider, Ian, Sedivy, R., & al, e. (2025). *Measuring the Environmental Impact of Delivering AI at Google Scale.* Mountain View, CA: Google.

Epoch AI. (2025). *Over 30 AI models have been trained at the scale of GPT-4*. Retrieved from Epoch AI: https://epoch.ai/data-insights/models-over-1e25-flop#:~:text=Our%20estimate%20of%20GPT%2D4,across%20a%20range%20of%20benchmarks.&text=Compute%20estimated%20using%20parameter%20count%20and%20dataset%20size.,-More%20information%20is

Europe Economics. (2025). *Impact of Growth of Data Centres on Energy Consumption*. Retrieved from https://assets.publishing.service.gov.uk/media/689d9dc487bf475940723f6c/impact-of-growth-of-data-centres.pdf

Faiz, A., Kaneda, S., Wang, R., Osi, R., Sharma, P., Chen, F., & Lei, J. (2024). LLM Carbon: Modeling the End-to-End Carbon Footprint of Large Language Models. *ICLR 2024.*

Global Energy Monitor. (2026, January). *Betting big on data centers, U.S. now leads world for new gas power development*. Retrieved from Global Energy Monitor: https://globalenergymonitor.org/report/betting-big-on-data-centers-u-s-now-leads-world-for-new-gas-power-development/

Global Energy Monitor. (2026). *Global Oil and Gas Plant Tracker*. Retrieved from Global Energy Monitor: https://globalenergymonitor.org/projects/global-oil-gas-plant-tracker

Google. (2024). *Google Environmental Report 2024*. Retrieved from Google: https://www.gstatic.com/gumdrop/sustainability/google-2024-environmental-report.pdf

Greenhouse Gas Protocol. (2011). *Product Life Cycle Accounting and Reporting Standard.* World Resources Institute and Business Council for Sustainable Development.

IBM. (2024). *What is knowledge distillation?* Retrieved from https://www.ibm.com/think/topics/knowledge-distillation

IBM. (2026). *What are diffusion models?* Retrieved from IBM: https://www.ibm.com/think/topics/diffusion-models

ICF. (2022). *Comparison of GHG Emissions from Scenes of On-Location and Virtual Productions*. Retrieved from https://sonypicturesgreenerworld.com/sites/sonypicturesgreenerworld.com/files/2022-09/Sony%20Pictuares_Virtual%20Production%20GHG%20Analysis_2022_2.pdf

ICF. (2023). *Revised Analysis Comparing GHG Emissions from Scenes of On-Location and Virtual Productions*. Retrieved from https://sonypicturesgreenerworld.com/sites/sonypicturesgreenerworld.com/files/2024-03/Revised%20Virtual%20Production%20Analysis%20Results%20Explanation_2023%5B1%5D.pdf

IEA. (2025). *CO2 emissions associated with electricity generation for data centres by case, 2020-2035*. Retrieved from IEA: https://www.iea.org/data-and-statistics/charts/co2-emissions-associated-with-electricity-generation-for-data-centres-by-case-2020-2035

IEA. (2025a). *Energy and AI.* France: IEA.

IEA. (2025b). *Data centre electricity consumption in household electricity consumption equivalents, 2024*. Retrieved from IEA: https://www.iea.org/data-and-statistics/charts/data-centre-electricity-consumption-in-household-electricity-consumption-equivalents-2024

IEA. (2025c). *Energy and AI: Executive Summary*. Retrieved from IEA: https://www.iea.org/reports/energy-and-ai/executive-summary

IEA. (2026d). *Key Questions on Energy and AI.* Paris: IEA.

IEEE Spectrum. (2026, April 7). *Decentralized Training Can Help Solve AI's Energy Woes*. Retrieved from IEEE Spectrum: https://spectrum.ieee.org/decentralized-ai-training-2676670858

IoES. (2023). *Virtual vs. Conventional Production for Film and Television: A Comparative Life Cycle Assessment*. Retrieved from https://www.ioes.ucla.edu/wp-content/uploads/2023/06/UCLA-IoES-Practicum-SPA-Virtual-Production-Final-Report-2023.pdf

ITU and WBA. (2025). *Greening Digital Companies 2025.* Geneva and Amsterdam: ITU and WBA.

Joshi, K. (2026). *The AI Climate Hoax: Behind the Curtain of How Big Tech Greenwashes Impacts.*

Kamiya, G. &. (2025). *Data Centre Energy Use: Critical Review of Models and Results.* EDNA - IEA 4E TCP.

Ketan Joshi. (2026). *LinkedIn*. Retrieved from LinkedIn: https://www.linkedin.com/posts/ketanjoshi1_this-is-a-huge-admission-coca-cola-revealed-activity-7406978626123956224-Kw8V/

Koomey, J., & Masanet, E. (2026). Understanding AI Energy Use. *Field Actions Science Reports*.

Li, B. Y. (2024). Carbon in motion: Characterizing Open-Sora on the sustainability of generative AI for video generation. *ACM SIGENERGY Energy Informatics Review*, 160-165.

Luccioni, A. S., Viguier, S., & Ligozat, A.-L. (2022). *Estimating the Carbon Footprint of BLOOM, a 176B Parameter Language Model.*

Luccioni, S. A., Jernite, Y., & Strubell, E. (2024). *Power Hungry Processing: Watts Driving the Cost of AI Deployment.*

Luccioni, S., Gamazaychikov, B., Alves de Costa, T., & Strubell, E. (2025). Misinformation by Omission: The Need for More Environmental Transparency in AI.

Meta. (2025a). *Llama 3.1 Model Card*. Retrieved from Github: https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/MODEL_CARD.md

Meta. (2025b). *Llama 4 Model Card*. Retrieved from Github: https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md

Mistral AI. (2025). *Our contribution to a global environmental standard for AI*. Retrieved from Mistral.ai: https://mistral.ai/news/our-contribution-to-a-global-environmental-standard-for-ai

MIT Technology Review. (2025, May 20). *We did the math on AI's energy footprint. Here's the story you haven't heard.* Retrieved from MIT Technology Review: https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/

ML.ENERGY. (2026). *The ML.ENERGY Leaderboard*. Retrieved from ML.ENERGY: https://ml.energy/leaderboard/

Nicholas Stern, M. R.-B. (2025). Green and intelligent: the role of AI in the climate transition. *Nature*.

Oviedo, F., Kazhamiaka, F., Choukse, E., Kim, A., Luers, A., Nakagawa, M., . . . Ferres, J. M. (2026). Energy use of AI inference, efficiency pathways, and test-time scaling. *Joule*.

Passoni, R., Ronchini, F., Comanducci, L., Serizel, R., & Antonacci, F. (2025). *Diffused Responsibility: Analyzing the Energy Consumption of Generative Text-to-Audio Diffusion Models.*

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., . . . Dean, J. (2021). *Carbon Emissions and Large Neural Network Training.*

Profile IT Solutions. (n.d.). *Best Practices for Data Center Area Sizing Per Rack Based on Power Density*. Retrieved from Profile IT Solutions: https://www.profileits.com/best-practices-for-data-center-area-sizing-per-rack-based-on-power-density/

Rhodium Group. (2025). *Preliminary US Greenhouse Gas Emissions Estimates for 2025*. Retrieved from https://rhg.com/research/us-greenhouse-gas-emissions-2025/

Sasha Luccioni, T. A. (2025, September 17). *What kind of environmental impacts are AI companies disclosing? (And can we compare them?)*. Retrieved from Hugging Face: https://huggingface.co/blog/sasha/environmental-impact-disclosures

Schneider Electric. (2015). *Guidelines for Specification of Data Center Power Density.* Retrieved from Schneider Electric: https://www.se.com/id/en/download/document/SPD_NRAN-69ANM9_EN/

Shehabi, A. S. (2024). *2024 United States Data Center Energy Usage Report.* Berkeley, CA: Lawrence Berkeley National Laboratory.

Shin, R. U. (2025). *Complete Guide to Five Generative AI Models*. Retrieved from Coveo: https://www.coveo.com/blog/generative-models/

The Wall Street Journal. (2025). *Google's First AI Ad Avoids the Uncanny Valley by Casting a Turkey*. Retrieved from The Wall Street Journal: https://www.wsj.com/articles/googles-first-ai-ad-avoids-the-uncanny-valley-by-casting-a-turkey-dafd3662?mod=rss_Technology

TokenRing AI. (2025, December 29). *Nvidia's Blackwell Dynasty: B200 and GB200 Sold Out Through Mid-2026 as Backlog Hits 3.6 Million Units*. Retrieved from Financial Content: https://markets.financialcontent.com/stocks/article/tokenring-2025-12-29-nvidias-blackwell-dynasty-b200-and-gb200-sold-out-through-mid-2026-as-backlog-hits-36-million-units#google_vignette

UNEP. (2025). *Emissions Gap Report 2025: Off target – Continued collective inaction.* Nairobi: UNEP.

Uptime Institute. (2025). *Density choices for AI training are increasingly complex*. Retrieved from Uptime Institute: https://journal.uptimeinstitute.com/density-choices-for-ai-training-are-increasingly-complex/

WRI & WBCSD. (2013). *Product Standard*. Retrieved from Greenhouse Gas Protocol: https://ghgprotocol.org/product-standard

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., & et, a. (2022). *Sustsainable AI: Environmental Implications, Challenges and Opportunities.* Facebook AI.

Yiran Lei, J. F. (2026, April 6). *The Energy Cost of Execution-Idle in GPU Clusters*. Retrieved from arXiv: https://arxiv.org/pdf/2604.04745

Tables

Figures

APPENDIX

Appendix 1: Additional data and figures

Table 11. Energy and carbon emissions from training of six selected LLMs

Model	Model release year	Parameters (billion)	Energy (MWh)	Carbon emissions (tCO_2e)	Boundary	Source
T5	2019	11	86	47	GPU, CPU, RAM, network interface, fans and PUE	Patterson et. al, 2021
Switch	2020	1,500	179	72	GPU, CPU, RAM, network interface, fans and PUE	Patterson et. al, 2021
GPT3	2020	176	1,287	552	GPU, CPU, RAM, network interface, fans and PUE	Patterson et. al, 2021
BLOOM	2022	176	433	30	GPU only	Luccioni et. al, 2022
Llama 3.3	2024	70		11,390	GPU only	Meta, 2025
Llama 4	2025	17		1,999	GPU only	Meta, 2025

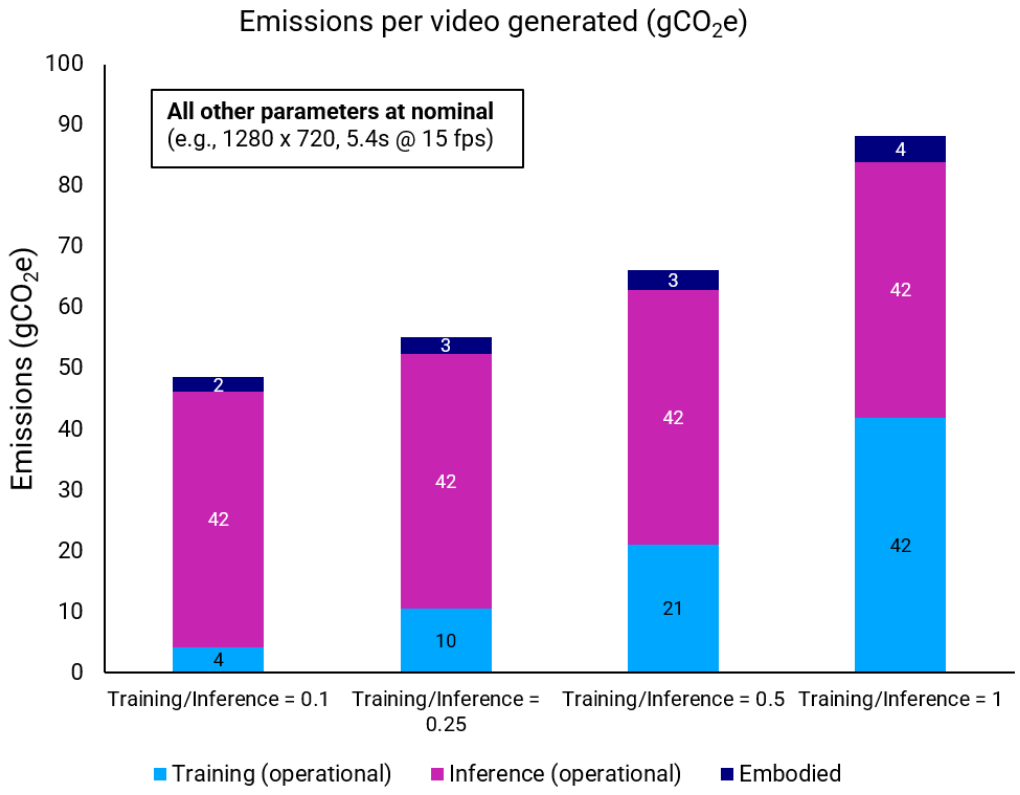


Figure 23. Nominal range values from sensitivity analysis by training, inference and embodied carbon at 720p

Appendix 2: Sensitivity analysis of allocation approaches for model creation (i.e., training)

We evaluate two approaches to allocating emissions from model creation to the lifecycle impact of the generated video, as proposed in Table 6: 1) allocation per unit energy and 2) allocation per inference

Table 12. Comparison of allocation approaches for model creation

	Characteristics	Considerations
Per unit energy allocation	• Lifecycle impact is variable and linked to inference energy consumption • Proportion of lifecycle emissions due to training is constant across resolutions • Relates total lifecycle carbon impact more closely to inference, where users have better capacity to influence impact via video parameter selection and usage patterns	• Simpler to understand • Does not fully reflect the how the system responds to the user • Heavier inferences bear more of the burden • Possibly more challenging to collect or estimate necessary data to calculate impact
Per inference allocation	• Lifecycle impact is made up of 'fixed' and 'variable' portions • Proportion of lifecycle emissions due to training varies across resolutions • Approach is not sensitive to inference profile, beyond total number of inferences	• Simple and fair way to allocate emissions that are already 'baked in' to the system • Inconsistent with the selected functional unit of MP-frames • May lead to counterintuitive results in some edge cases, which negatively affect user behaviour

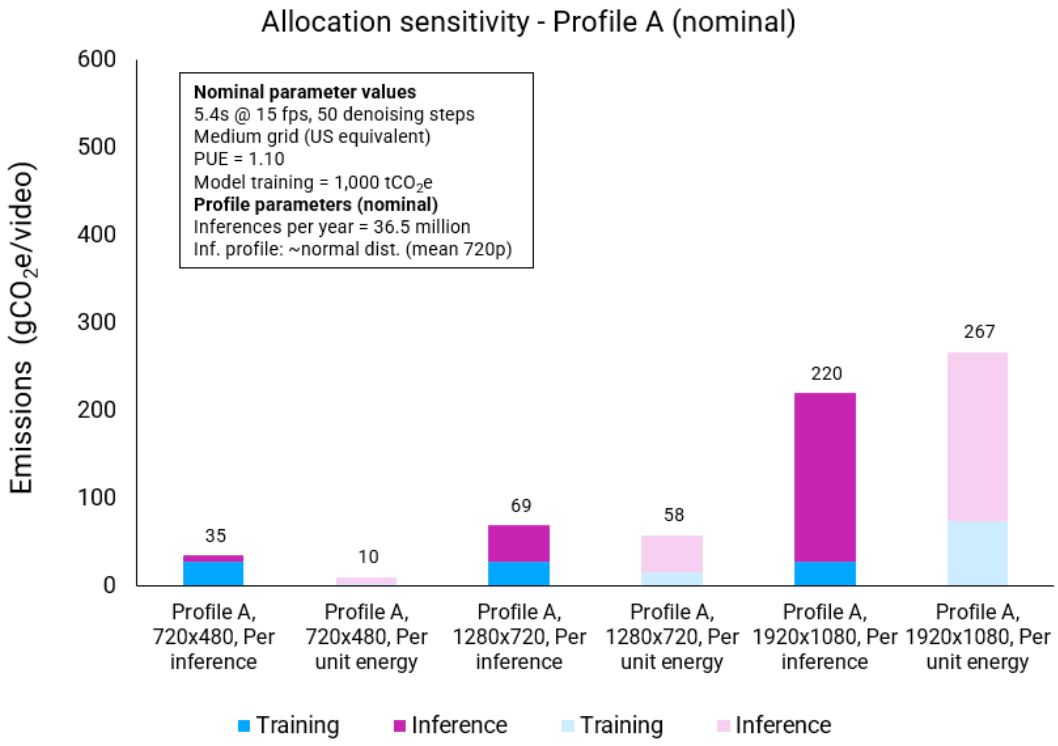


Figure 24. Comparison of allocation approaches

In Figure 24 we compare both approaches. We conclude that an allocation of model training emissions per unit of energy related to inference is a stronger choice, principally because we expect inference to be the most material lifecycle stage and as the intended audience of these impact figures are ultimately the users, this approach does a better job of guiding users' decision making towards lower impact outcomes. We observe this in assessing the difference in impacts at lower and higher resolutions, where the per inference approach effectively 'subsidises' the carbon impact at higher resolutions with the carbon impact at lower resolutions.

Appendix 3: Emissions intensities used in the case study

Activity	Label	Element	Carbon impact	Unit	Source
Image generation	Image generation, US	Inference	0.002	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, US	Training	0.001	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, US	Embodied	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, US	Total	0.002	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, France	Inference	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, France	Training	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, France	Embodied	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, France	Total	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, UK	Inference	0.001	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, UK	Training	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, UK	Embodied	0.000	$kgCO_2e$/image	CT analysis (see below)
Image generation	Image generation, UK	Total	0.001	$kgCO_2e$/image	CT analysis (see below)
Video gen (cloud)	720p, low, US	Inference	0.017	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, US	Training	0.002	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, US	Embodied	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, US	Total	0.020	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, US	Inference	0.017	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, US	Training	0.008	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, US	Embodied	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, US	Total	0.027	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, US	Inference	0.017	$kgCO_2e$/second	CT analysis (see below)

Video gen (cloud)	720p, high, US	Training	0.017	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, US	Embodied	0.002	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, US	Total	0.036	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, France	Inference	0.003	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, France	Training	0.000	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, France	Embodied	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, France	Total	0.004	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, France	Inference	0.003	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, France	Training	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, France	Embodied	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, France	Total	0.005	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, France	Inference	0.003	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, France	Training	0.003	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, France	Embodied	0.002	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, France	Total	0.007	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, UK	Inference	0.009	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, UK	Training	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, UK	Embodied	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, low, UK	Total	0.011	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, UK	Inference	0.009	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, UK	Training	0.005	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, UK	Embodied	0.001	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, nominal, UK	Total	0.015	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, UK	Inference	0.009	$kgCO_2e$/second	CT analysis (see below)

Video gen (cloud)	720p, high, UK	Training	0.009	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, UK	Embodied	0.002	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	720p, high, UK	Total	0.020	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, low, US	Inference	0.080	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, low, US	Training	0.008	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, low, US	Embodied	0.005	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, low, US	Total	0.093	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, nominal, US	Inference	0.080	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, nominal, US	Training	0.040	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, nominal, US	Embodied	0.006	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, nominal, US	Total	0.126	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, high, US	Inference	0.080	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, high, US	Training	0.080	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, high, US	Embodied	0.008	$kgCO_2e$/second	CT analysis (see below)
Video gen (cloud)	1080p, high, US	Total	0.169	$kgCO_2e$/second	CT analysis (see below)
Upscaling	Upscaling (low)	Inference	0.002	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (low)	Training	0.000	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (low)	Embodied	0.000	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (low)	Total	0.002	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (nominal)	Inference	0.005	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (nominal)	Training	0.005	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (nominal)	Embodied	0.001	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (nominal)	Total	0.010	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (high)	Inference	0.01	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)

Upscaling	Upscaling (high)	Training	0.009	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (high)	Embodied	0.001	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Upscaling	Upscaling (high)	Total	0.019	$kgCO_2e$/MP-second @ 30 fps	CT analysis (see below)
Local compute	Local compute, US	Electricity	0.410	$kgCO_2e$/kWh	IEA, Emissions Factors 2025
Local compute	Local compute, France	Electricity	0.064	$kgCO_2e$/kWh	IEA, Emissions Factors 2025
Local compute	Local compute, UK	Electricity	0.222	$kgCO_2e$/kWh	IEA, Emissions Factors 2025
Local compute	Local compute, training (low)	Training	0.001	$kgCO_2e$/kWh	CT analysis (see below)
Local compute	Local compute, training (nominal)	Training	0.026	$kgCO_2e$/kWh	CT analysis (see below)
Local compute	Local compute, training (high)	Training	1.280	$kgCO_2e$/kWh	CT analysis (see below)
Local compute	Local compute, embodied	Embodied	0.020	$kgCO_2e$/hour	CT analysis (see below)
Artist	Artist computers, US	Electricity	0.104	$kgCO_2e$/hour	Assuming 250W
Artist	Artist computers, France	Electricity	0.016	$kgCO_2e$/hour	Assuming 250W
Artist	Artist computers, UK	Electricity	0.056	$kgCO_2e$/hour	Assuming 250W
Artist	Artist computers	Embodied	0.020	$kgCO_2e$/hour	CT analysis (see below)
Artist	Artist computers	Software (development)	0.017	$kgCO_2e$/hour	CT analysis (see below)
Artist	Artist computers, US	Total	0.141	$kgCO_2e$/hour	CT analysis (see below)
Artist	Artist computers, France	Total	0.054	$kgCO_2e$/hour	CT analysis (see below)
Artist	Artist computers, UK	Total	0.094	$kgCO_2e$/hour	CT analysis (see below)

Image generation: assumes 2.907 kWh/1000 queries for image generation inference (Luccioni, Jernite, & Strubell, 2024) with uplift of additional 29% to account for other IT components and PUE (assuming PUE of 1.1). Embodied carbon assumed at 5% of total carbon impact.

Video generation (cloud): assumes inference GPU energy of 381 Wh and 1,847 Wh per 8-second video & 24 fps at 720p and 1080p, respectively, derived using the prediction model from Delavande et al. (Delavande, Pierrard, & Luccioni, 2025). Total inference energy (including other IT and PUE of 1.1) is estimated as

494 Wh and 2,391 Wh, respectively. Training assumes a range of 0.1x, 0.5x and 1.0x inference (low, nominal, high). Embodied carbon is estimated as 5% of total carbon impact on the US grid, equating to 25% of total carbon impact on France grid and 8.8% of total carbon impact on UK grid.

Upscaling: assumes pricing of 0.1 $/MP-second at 30 fps. Token pricing range of $5/million tokens to $15/million tokens. Energy per token derived from (Oviedo, et al., 2026) and estimated between 0.35 and 0.7 Wh/1k tokens, resulting in a range of 3.68 to 22.11 Wh/MP-second. Training is applied with a range of 0.1x to 1.0x inference (low to high). Embodied carbon is assumed to be 5% of total carbon impact.

Local compute (training): Allocation of training emissions for open-models used on local workstations was derived as follows:

Scenario	Training emissions (tCO2e)	Annual users	Annual videos per user	Energy per inference at 720p (Wh)	Total inference energy (kWh)	Allocated training emissions ($kgCO_2e/kWh$)
Low	50	420,000	1000	93	39,060,271	0.001
Nominal	100	84,000	500	93	3,906,027	0.026
High	500	16,800	250	93	390,603	1.280
Notes	Assumed range with reference to range from Fig. 5 of the report	Nominal is based on downloads last month of 6,826 for WAN 2.1 VACE on Hugging Face		GPU, CPU, RAM , WAN2.1-T2V-1.3B (Delavande, Pierrard, & Luccioni, 2025)		

Local compute and artist computers embodied: Derived from a bottom-up estimate of workstation computer embodied carbon based on specifications provided by the studio. Assume total embodied carbon impact of 447 $kgCO_2e$ per computer, 5-year lifetime and 50% active usage rate.

Software development: Estimated based on Adobe's 2024 reported emissions (Scope 1, 2 and upstream Scope 3 of 523,411 tCO_2e) and derivation of top-down emissions intensity per hour of user. Assumes 850 million monthly active users (MAU), 0.1 daily hours per MAU, totalling 31.025 billion annual hours.

Appendix 4: Equations for case study analysis

General terminology: *E* denotes emissions, *A* denotes activity, *I* denotes emission intensity

Artist computers

$$I_{artist\ computer,total} = I_{artist\ computer,elec.} + I_{artist\ computer,embodied} + I_{artist\ computer,software}$$

$$E_{artist\ computer} = A_{artist\ computer} \times I_{artist\ computer,total}$$

Where $A_{artist\ computer}$ is measured in hours of computer use

Image generation

$$I_{image\ generation,total} = I_{image\ generation,inference} + I_{image\ generation,training} + I_{image\ generation,embodied}$$

$$E_{image\ generation} = A_{image\ generation} \times I_{image\ generation,total}$$

Where $A_{image\ generation}$ is measured in total number of images generated

Video generation (cloud)

$$A_{video\ gen,cloud} = Q_{credits} \times C_{credits-to-videos}$$

Where $A_{video\ gen,cloud}$ is in seconds of video generated, $Q_{credits}$ is the quantity of credits used for video generation and $C_{credits\text{-}to\text{-}videos}$ is the conversion applied based on credit pricing per second of video generated

$$I_{video\ gen,cloud,total} = I_{image\ gen,cloud,inference} + I_{video\ gen,cloud,training} + I_{video\ gen,cloud,embodied}$$

$$E_{video\ generation} = A_{video\ generation} \times I_{video\ generation,total}$$

Video generation (local compute)

$$E_{video\ generation,local,inference} = A_{local\ compute,energy} \times I_{electricity}$$

Where $A_{local\ compute,\ energy}$ is measured in kWh based on electricity consumed by local workstations

$$E_{video\ generation,local,training} = E_{video\ generation,local,inference} \times U_{training}$$

Where $U_{training}$ is the uplift applied to account for training emissions

$$E_{video\ generation,local,embodied} = A_{local\ compute,hours} \times I_{local\ compute,embodied}$$

Where $A_{local\ compute,\ hours}$ is measured in workstation compute hours

$$E_{video\ generation} = E_{video\ generation,local,inference} + E_{video\ generation,local,training} + E_{video\ generation,local,embodied}$$

Upscaling

$$C_{videos-to-MP-frame} = MP_{video} \times D_{video}$$

Where, $C_{videos-to-MP-frame}$ is the conversion applied from quantity of videos to MP-frame, MP_{video} is the total megapixels of the video (e.g., 8.3 megapixels at resolution of 3840 x 2160 pixels), D_{video} is the duration of the video in seconds (e.g., 8 seconds).

$$A_{upscaling} = Q_{upscaled\ videos} \times C_{videos-to-MP-frame}$$

Where $A_{upscaling}$ is the total quantity of videos produced, expressed in total MP-frames, and $Q_{upscaled\ videos}$ is the quantity of upscaled videos produced

$$I_{upscaling,total} = I_{upscaling,inference} + I_{upscaling,training} + I_{upscaling,embodied}$$

$$E_{upscaling} = A_{upscalinig} \times I_{upscaling,total}$$

Traditional VFX rendering (local compute)

$$E_{rendering,local} = A_{local\ compute,energy} \times I_{electricity} + A_{local\ compute,hours} \times I_{local\ compute,embodied}$$

Appendix 5: Indicative uncertainty and data quality associated with case study

Data quality and uncertainty have been assessed for both the Generative FX and Traditional VFX scenarios, as detailed in Table 13 and Table 14 where:

- Uncertainty is assessed in relation to the main emissions intensities used due to the sensitivity of assumptions used to derive these.
- Data quality is assessed in relation to the activity data used, to reflect the discrepancy between primary data in some areas compared to estimates from desk research in others.

Table 13: Uncertainty of main emissions intensities used in case study analysis

Activity	Activities covered by emissions intensity	Indicative uncertainty – Generative FX	Indicative uncertainty – Traditional VFX	Notes
Image generation	All	Medium-to-high	N/A	Emission intensity derived from academic paper with uplifts to account for training and embodied emissions
Video generation (cloud)	All	Medium-to-high	N/A	Based on the Carbon Trust sensitivity analysis with formulae derived from a single study on text-to-video generation
Upscaling	All	High	N/A	Emission intensity derived from a per token estimate of energy consumption
Local compute	Electricity	Low	Low	Grid average electricity emissions intensity used
Local compute	Training (uplift)	High	N/A	Estimated uplift range used due to uncertainty associated with this metric
Local compute	Embodied carbon	Medium	Medium	Bottom-up estimate made based on specifications of workstations
Artist computers	All	Medium	Medium	Provided by studio, estimated for Traditional VFX
Total	**All**	**Medium-to-high**	**Medium**	**Overall assessment of uncertainty**

Table 14: Data quality of main activity data used in case study analysis

Activity	Indicative data quality – Generative FX[33]	Indicative data quality – Traditional VFX	Notes
Image generation	High (measured)	N/A	Number of images generated measured through SaaS platform for base case
Video generation (cloud)	Low-to-Medium (estimated)	N/A	Seconds of video generated relies on estimate of iterations per shot and conversion from credits
Video generation (local)	High (measured)	N/A	Measured local workstation electricity is used instead in the case study analysis.
Upscaling	Medium (estimated)	N/A	Relies on estimate of iterations per shot
Local compute (workstation electricity)	High (measured)	Medium (estimated)	Measured for Generative FX, estimated by studio for Traditional based on prior experience
Local compute (hours)	High (measured)	Medium (estimated)	Measured for Generative FX, estimated by studio for Traditional based on prior experience
Artist computers (hours)	High (measured)	Medium (estimated)	Provided by studio, estimated by studio for Traditional VFX
Total	**Medium-to-High**	**Medium**	**Overall assessment of data quality**

[33] Data quality classified as: Low = Significant use of proxies or assumptions; Medium = Moderate use of proxies or assumptions; High = Little to no use of proxies or assumptions.

Appendix 6: Generative AI model architecture and operation

Large language models: tokens are a proxy for energy consumption and output token length is a key driver.

Text generation models process so-called tokens. A token is a unique number for a word or part of a word that is pre-defined in an embedding table. As all words in the input are rewritten into equal-sized token vectors, the input prompt forms a matrix which has the dimensionality of the LLM model width. The `reading' of the input prompt (the prefill) is then carried out as a matrix multiplication in parallel. During decoding, when a response is produced, each output token is generated one at a time until the next predicted token is a stop token (which the model learned during training). Even though a cache of working memory helps to speed up this process, this iterative (so-called auto-regressive) approach still tends to be computationally more expensive (and thus energy consuming) than the parallel reading of the input.

In other words, for energy consumption, the length of the response matters more than the input text. This input text includes the entire context of the prompt, including chat history, and attached documents and images (images are converted to image tokens first). A short 'yes' or 'no' question usually requires less computation than an explanation. Asking the model to `explain its reasoning' usually increases the number of output tokens and therefore the energy used. Very long inputs become relevant drivers of energy consumption when users paste long documents.

Notably, the content of the message in only relevant in as far as it influences the length of the response, (i.e. the number of tokens generated before the stop token is produced). It does not matter if the output is a summary of quantum physics or a recipe for sponge cake; nor does it matter if the output is correct or not. Energy use is determined by the number of matrix operations performed. A long incorrect answer can use more energy than a short correct one.

Diffusion models: tokens are NOT a proxy for energy consumption

As we saw earlier, diffusion models start with noise and remove it ('de-noising') until an image emerges. They work primarily with 'continuous' rather than 'tokenised' data. At the level of working principles of diffusion models, the content of the image is largely irrelevant for image generation. More importantly, the denoising is iterative, starting with very noisy images before high-level structures become defined, and moving to less noisy images when details are being refined.

Within each step the model works on a fixed-size patches of the image through matrix operations, which are independent of the content of the pixels. What does matter is the resolution of the image as it is broken down into more patches to fit the size of the matrices. Much more important for the energy consumption is the number of denoising steps, which is being chosen by the model provider when the image generation model is deployed.

During Image-to-image generation the model partially noises the input image and then denoises the image, usually preserving some of the composition. This can be more efficient than text to image if the number of denoising steps is reduced. However, this is a deployment setting that the user cannot influence. The energy cost for image-to-image models is thus still mainly driven by resolution and the number of denoising steps. Learn more about types of generative AI models at (Shin, 2025) and diffusion models, specifically (IBM, 2026).

carbontrust.com

+44 (0) 20 7170 7000

Published in the UK: 2026