unesco

Smarter, Smaller, Stronger:

Resource-Efficient Generative AI & the Future of Digital Transformation

Cite as: UNESCO (2025). “Smarter, Smaller, Stronger: Resource-Efficient Generative AI & the Future of Digital Transformation”

Published in 2025 by the United Nations Educational, Scientific and Cultural Organization (UNESCO), 7, Place de Fontenoy, 75007 Paris, France.

CI/DIT/2025/ER/01 Rev.

Authored by: Leona Verdadero[1] *, Ivana Drobnjak[2] *, Hristijan Bosilkovski[2]*, Zekun Wu[23], Emma Fischer, and María Pérez-Ortiz[2].

[1] UNESCO
[2] University College London
[3] Holistic AI

* First three authors contributed equal work.

Printed by: UNESCO
Printed in France

Acknowledgements: The authors would like to acknowledge Clare O’Hagan for her contributions to the research and analysis.

Graphic design: Sonia Savci
Cover design: Sonia Savci
Illustrations: Sonia Savci

Smarter, Smaller, Stronger: Resource-Efficient Generative AI & the Future of Digital Transformation

EXECUTIVE SUMMARY

Artificial Intelligence (AI) holds immense potential to support global efforts to reduce environmental impact by optimizing energy use, enhancing resource management, and improving climate modeling and prediction[1]. However, the accelerated rise of generative AI, particularly Large Language Models (LLMs), has brought new and urgent resource challenges. The exponential growth in computational power needed to run these models is placing increasing strain on global energy systems, water resources, and critical minerals, raising concerns about environmental sustainability, equitable access, and competition over limited resources.

Achieving ecological resilience in generative AI is not solely a matter of reducing energy consumption, it is about unlocking broader opportunities, expanding equitable access, and enabling scalable, impactful innovation. A fundamental shift toward AI systems that are "clean by design," with energy and resource efficiency integrated from the outset, is essential. This requires developing models that are not only high-performing but also lighter, more efficient, and environmentally sustainable, particularly as generative AI becomes a foundational layer of our digital infrastructure. Embracing energy- and resource-efficient AI is key to ensuring that the digital transformation advances in a way that is both inclusive and ecologically responsible, capable of scaling across diverse global contexts.

To turn this vision into reality, addressing generative AI's sustainability challenges demands sustained commitment and collective action. Policymakers, industry leaders, and the scientific community must prioritize the development of AI systems that are both energy-efficient and accessible, particularly in low-resource contexts. This report offers three key recommendations to support that shift: (a) mobilize public and private investment, along with strategic partnerships, to drive the development and adoption of clean by design AI systems that embed efficiency from the outset; (b) create incentives and standards, such as sustainability labels and green procurement criteria, that encourage transparency and promote eco-conscious design and usage across the AI ecosystem; and (c) enhance AI literacy to build critical awareness of generative AI's environmental footprint and foster more intentional and conscious engagement.

Through a combination of original experiments and data insights, this report illustrates how practical techniques can help translate that vision into action**.** Methods such as quantization and prompt optimization reduced the energy consumption of large language models by up to 75% without compromising accuracy. Moreover, in tasks that are specialized and repetitive, such as translation or summarization, replacing a large general-purpose model with smaller, task-specific models led to

energy reductions of up to 90%, while maintaining strong performance. These findings offer a tangible, scalable pathway toward a smarter, more accessible, and more resource-efficient AI future.

UNESCO's commitment to advancing the right to information, equitable access to knowledge, and ethical digital transformation underpins this report. As generative AI becomes a foundational layer of digital infrastructure, it is essential to ensure that its development supports sustainability, inclusion, and the public interest. This report provides evidence-based insights into how energy- and resource-efficient AI can help achieve these goals, especially in low-resource settings.

It aligns with UNESCO's broader efforts to strengthen information integrity, promote open and inclusive digital ecosystems, and support countries in navigating emerging technologies in line with human rights and environmental priorities. It also contributes to global initiatives such as the *Global Roadmap on Information as a Public Good in the Face of the Environmental Crisis* and the *Global Initiative on Information Integrity on Climate Change*, which highlight the need to combat disinformation, ensure access to reliable environmental data, and empower citizens and media with the knowledge and tools they need.

By embedding sustainability and transparency into AI systems, this report supports a wider vision: one where digital transformation is not only innovative, but also inclusive, rights-affirming, and environmentally aligned.

KEY MESSAGES

1. **The rapid expansion and widespread adoption of mainstream generative AI technologies is placing growing pressure on global energy and resource systems, raising serious concerns about long-term environmental sustainability and resource efficiency.**

Training state-of-the-art large language models (LLMs) and general-purpose models consumes approximately 50 GWh of electricity, comparable to the annual electricity use of some developing countries[2]. Even more alarming is the energy footprint of LLM model inference, the energy consumed when users interact and prompt with LLMs, which grows exponentially as generative AI tools become embedded in everyday life.

To illustrate the scale of energy use, let us consider ChatGPT, one of the most popular large language models in the world, and as of mid-2025, the fifth-most visited website in the world, ranking just after Instagram and ahead of X[3]. While ChatGPT is a proprietary, closed-source model, publicly available estimates offer insight into its energy demands. As of June 2025, ChatGPT receives approximately 1 billion queries daily, each using around 0.34 Wh of electricity, about what it takes to power a high-efficiency LED lightbulb for a few minutes, according to OpenAI CEO Sam Altman[4]. That adds up to roughly 310 GWh per year, which is comparable to the annual electricity consumption of over 3 million people in Ethiopia, where average per-capita use is around 96 kWh/year[5].

Data centers, or AI factories, where AI models are developed and implemented, are the driving force behind the environmental impact of AI[6]**.** Compute demand by AI is doubling every 100 days[7], driving a proportional increase in energy use. The International Energy Agency (IEA) estimated that data

center electricity consumption has grown by around 12% per year since 2017, more than four times faster than the rate of total electricity consumption[8].

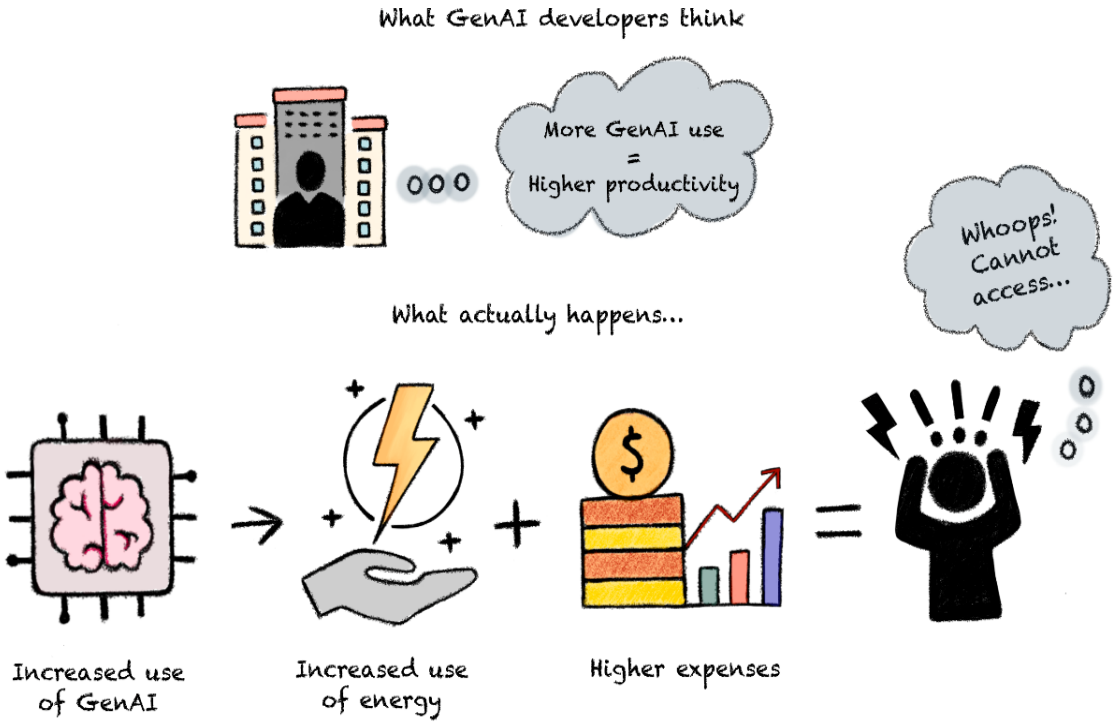
Beyond electricity, the water usage associated with training and running LLMs adds another layer to the environmental burden. Data centers that support AI operations require vast quantities of water for cooling, and significantly, much of this is fresh, potable water, further intensifying global water stress. Many of these buildings use millions of gallons of water (often fresh, potable water) per day in their cooling operations[9].

Projections indicate that the global water consumption of major AI players like Google, Microsoft, and Meta could triple by 2027. This isn't limited to cooling alone: water is also extensively used in the manufacturing and construction of AI hardware, particularly for cooling electronic components during production. AI demand is expected to consume between 4.2 and 6.6 billion cubic meters of water by 2027, surpassing Denmark's total annual water withdrawal of 4–6 billion cubic meters[10].

These demands are not just an environmental issue, they are a resource allocation challenge. In low-resource settings, where reliable energy and water are already scarce, the expansion of energy-intensive AI infrastructure competes directly with critical societal needs.

2. The current state of Generative AI development and deployment could deepen digital divides, excluding the same communities most at risk from environmental harm.

The benefits of generative AI are disproportionately accessible to those in regions with advanced digital infrastructure (e.g. high-performance computing) where most of AI development and adoption are taking place. In contrast, an estimated 2.6 billion people are offline in 2025, accounting for 32 per cent of the world's population[11]. Out of this total, 1.8 billion people live in rural areas. In Africa, only 5% of AI talent has access to adequate computing resources[12], and only a handful of nations in the region host supercomputing infrastructure for generative AI applications. Data center investments, while increasing, still represent less than 1-1.5% of global capacity[13], and their expansion risks exacerbating water scarcity and energy inequality, particularly in drought-prone or energy-constrained areas. In resource-constrained settings, the reliance on fossil fuels and limited renewable



energy options to power AI infrastructure further deepens environmental and social disparities, making resource-efficient AI practices an even greater imperative.

3. Addressing AI interactions, also known as inference, is key for building smarter, resource-efficient generative AI.

A paradigm shift is urgently needed in how we build and scale generative AI. The current trajectory, focused on ever-larger, more complex models, is creating barriers to innovation and accessibility, particularly in low-resource contexts. It is also driving up computing costs, slowing performance, and increasing emissions from energy-hungry infrastructure.

To reverse this trend, we must rethink AI systems through the lens of "clean by design", embedding energy and resource efficiency into the architecture of AI from the start. Sustainability must shift from an add-on to a foundational design principle, driving the creation of models that are leaner, faster, and more accessible by default.

This is especially urgent when it comes to inference, the stage when people use generative AI tools in real time. While training large models consumes massive power, it is now the billions of daily user prompts that make up the bulk of AI's environmental impact. Despite the urgency of the issue, practical solutions to reduce the energy consumption of AI inference are only just beginning as the field is evolving.

This report introduces three practical approaches to advancing resource-efficiency at the inference level, showing how small innovative solutions can deliver big environmental and accessibility gains:

a. **Shrink the model**: Reducing the size of models (compression) brings efficiency in bits and bytes. Model compression techniques like quantization can achieve energy savings of up to 44% by reducing computational complexity. It also reduces the cost of running LLMs by shrinking their size and making them faster.

b. **Say more with less:** Shorter prompts and responses lead to smart saving. Streamlining input queries and response lengths can reduce energy use by over 50%. Shortening inputs and outputs also reduces cost of running LLMs.

c. **Small is powerful:** Small language models (SLMs) can be powerhouses: Adopting smaller, fine-tuned models for specific applications can deliver energy savings of up to 90% while maintaining high accuracy. In addition to energy efficiency, small models are more accessible in low-resource environments with limited connectivity, offer faster response times, and are cost-effective.

To address the urgent need for resource-efficient and cleaner generative AI, the report recommends:

a. **Mobilize innovation to advance resource-efficient AI systems.** Accelerate the development and deployment of "clean by design" AI through bold public and private investment, strategic partnerships, and multi-stakeholder collaboration. This requires channeling support toward innovative approaches that embed energy and resource efficiency into AI systems from the outset, particularly at the inference stage where everyday energy use occurs. Dedicated research consortia, innovation hubs, and accelerators focused on sustainable AI should be prioritized to scale practical, cost-effective solutions. As major investment has

gone into developing frontier models, equal ambition must now be directed toward building resource-conscious systems that are accessible, efficient, and tailored for real-world use, especially in underserved settings.

b. **Incentivize transparency and eco-conscious innovation across the AI ecosystem.** Encourage more eco-conscious design and usage by establishing clear incentives, sustainability standards, and public procurement criteria that reward resource-efficient AI, particularly at the inference stage, where energy usage is most frequent and scalable. Like energy labels on appliances, visible indicators, such as efficiency ratings or environmental impact disclosures, can help users make informed choices and motivate developers to innovate toward sustainability. Developers and operators should also commit to transparent reporting of energy use, carbon emissions, and water consumption. Regulatory frameworks and independent audits can help ensure accountability and drive continuous improvement, making sustainability a foundational principle of AI innovation.

c. **Promoting AI literacy is essential**. Supporting education initiatives that increase user understanding of AI interaction and engagement while raising awareness of the environmental costs of generative AI can empower policymaking, decision-making and promote eco-conscious usage of these technologies. AI literacy initiatives include fostering critical thinking on AI's benefits and risks, integrating sustainability and green principles into AI literacy curricula, and equipping policymakers, developers, and consumers with the knowledge and guidance on sustainable AI usage and engagement.

AI & ENVIRONMENT

Artificial intelligence (AI) is a double-edged sword in environmental sustainability.

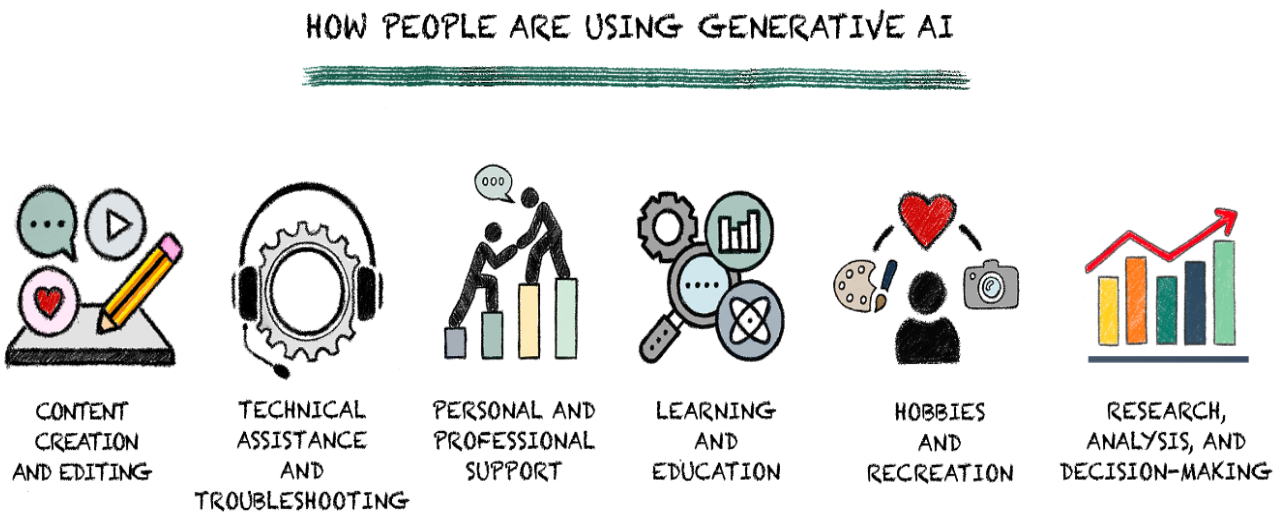
Artificial Intelligence (AI) has undergone a remarkable transformation over the past decade, transitioning from specialized applications to becoming a cornerstone of modern life, significantly affecting our digital interactions. Its ability to process vast amounts of data, recognize patterns, and make intelligent decisions has unlocked new possibilities across industries, including fostering sustainability strategies in several sectors of our economies. Today, AI powers tools ranging from voice assistants, chatbots, fraud detection, personalized educational platforms, language translation to advanced medical diagnostics, revolutionizing the way we live and work.

However, this rapid expansion comes with a complex relationship with the environment. On one side, AI shows immense potential to address global challenges, particularly in advancing climate change solutions on adaptation and mitigation. Much of these applications are in predictive AI tools, which use historical data to forecast future events or trends. For instance, predictive AI enables precision agriculture by analyzing weather patterns and soil conditions to optimize crop yields. It helps monitor deforestation through satellite imagery, predict the impacts of climate change, and optimize energy grids by forecasting demand and supply[14].

On the other side, the recent explosion of generative AI, a type of AI that creates new content, such as text, images, videos or music, often requires significant computational power and consumes substantial resources, including energy and water, for cooling the data centers that power these models. At the core of generative AI applications are large language models (LLMs), a subset of AI systems designed to understand and generate human-like text. LLMs are trained on vast datasets containing billions of sentences, requiring immense computational resources and energy. Notable advancements include BERT (by Google, 2018) and GPT-3 (by OpenAI, 2020), LLaMA (Meta, 2023), Claude (Anthropic, 2023), and Gemini 1.5 (Google DeepMind, 2024), which have revolutionized natural language processing and made it easier for machines to understand and generate human-like text.

The use of generative AI is set to grow exponentially as more companies, governments and organizations adopt it to drive efficiency and productivity. There are many use cases, spanning a vast number of areas of domestic and work life. The use of this technology is as wide-ranging as the

problems we encounter in our lives, including content creation, learning and education, technical assistance and troubleshooting, research and analysis, and hobbies, to name a few[15].



This duality positions AI as both a powerful ally and a potential adversary in the fight for a sustainable future. The environmental demands of AI, particularly generative AI, exacerbate resource consumption, deepening global inequalities, especially between regions with abundant computing power and data and those with limited access. As this study will show, addressing these challenges requires reconciling the growing ubiquity of AI with its resource-intensive nature.

To address the environmental impact of AI, industry and the scientific community are pursuing a range of innovative solutions under the umbrella of resource-efficient and green compute strategies. Solutions include efforts to improve the efficiency of AI models through optimizing hardware, improving algorithms, and considering renewable energy sources to offset their carbon footprint. Advances in energy-efficient chips, data center cooling techniques, and software optimization may also help reduce energy demand over time.

It is imperative to recognize the environmental cost of our daily digital interactions. While each individual search or prompt might seem negligible, collectively, they contribute to significant energy use and carbon emissions. To fully understand and address these impacts, it's essential to examine each phase of the AI system lifecycle–from inception to retirement–where different sustainability challenges and opportunities arise.

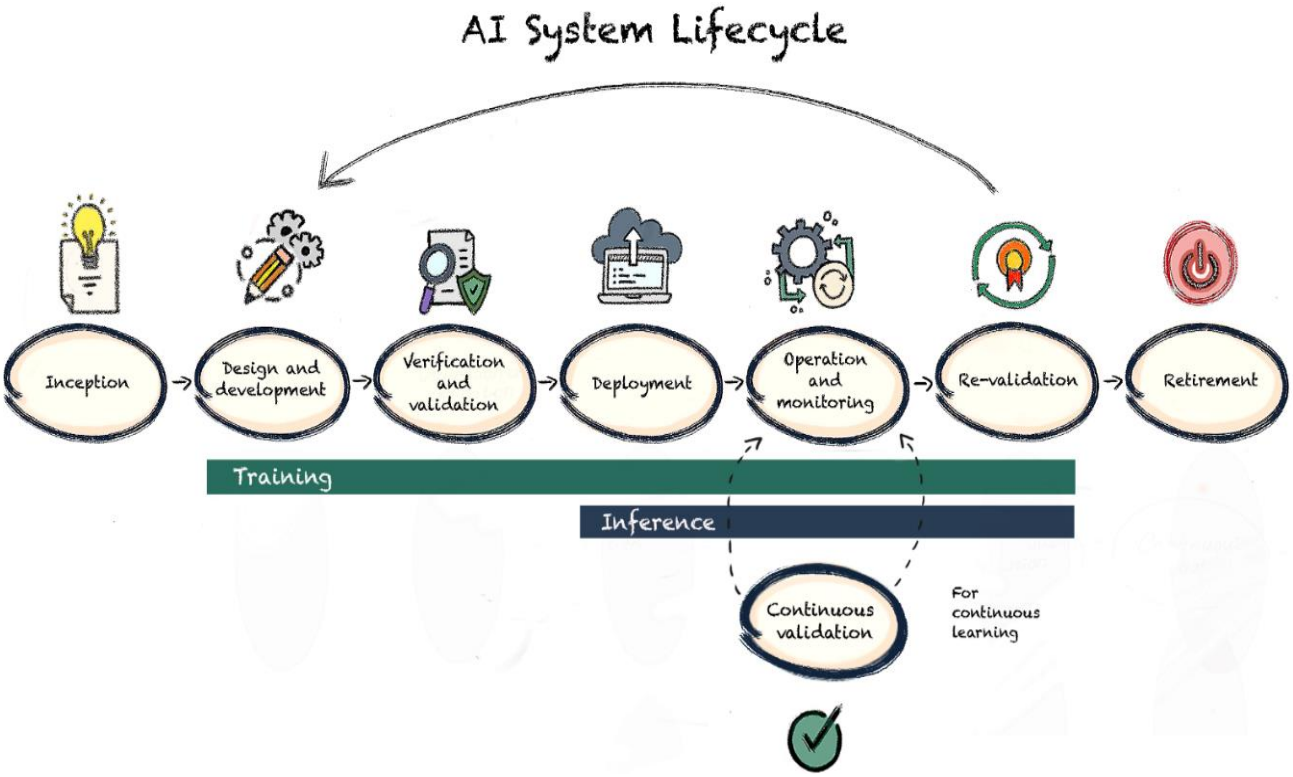
THE AI LIFECYCLE

The AI System Lifecyle: Each stage contributes to the environmental footprint

The **AI System Lifecycle**[16] outlines the key stages involved in developing, deploying, and maintaining an AI system, ensuring its effectiveness, reliability, and adaptability over time.

It begins with **Inception**, where the problem is defined, objectives are set, and feasibility is assessed. This is followed by **Design and Development**, where engineers select appropriate algorithms, collect and prepare data, and build the AI model. The system then undergoes **Verification and Validation**, a critical phase of rigorous testing to ensure accuracy, fairness, and robustness before deployment. In the **Deployment** stage, the AI model transitions into real-world use, processing new data and generating insights. However, the lifecycle does not end there; **Operation and Monitoring** is essential to track system performance, detect potential issues, and implement necessary adjustments. Over time, **Re-validation** is conducted to reassess the model's effectiveness, particularly as data patterns evolve. Eventually, when an AI system becomes outdated or no longer meets requirements, it reaches the **Retirement** stage, where it is decommissioned or replaced.

Throughout this lifecycle, two key processes ensure continuous improvement: **Training**, which spans from design to deployment, enabling the model to learn from data, and **Inference**, which allows the system to generate predictions in real-world scenarios. Additionally, **Continuous Validation** plays a crucial role in maintaining accuracy and reliability through iterative feedback loops and ongoing refinement. This structured approach ensures that AI systems remain aligned with ethical, regulatory, and performance standards, supporting responsible AI development.



ENERGY-EFFICIENT

Energy-efficient Computing: Focusing on AI Inference to drive efficiency and accessibility.

AI Inference, or interacting with the model, is becoming the largest cumulative energy footprint over the model's lifecycle. While training is the most energy-intensive stage on a per-event basis, the cumulative energy impact of inference often surpasses it over a model's lifecycle. Training involves processing massive datasets and running complex computations on high-performance hardware, consuming significant resources in a relatively concentrated timeframe. However, once the model is deployed, inference occurs repeatedly, potentially billions or even trillions of times, as the model is used by applications serving large user bases. Each individual inference requires less energy than training, but the sheer scale of usage, coupled with the need for continuous operation across global data centers, can result in a far greater overall energy footprint.

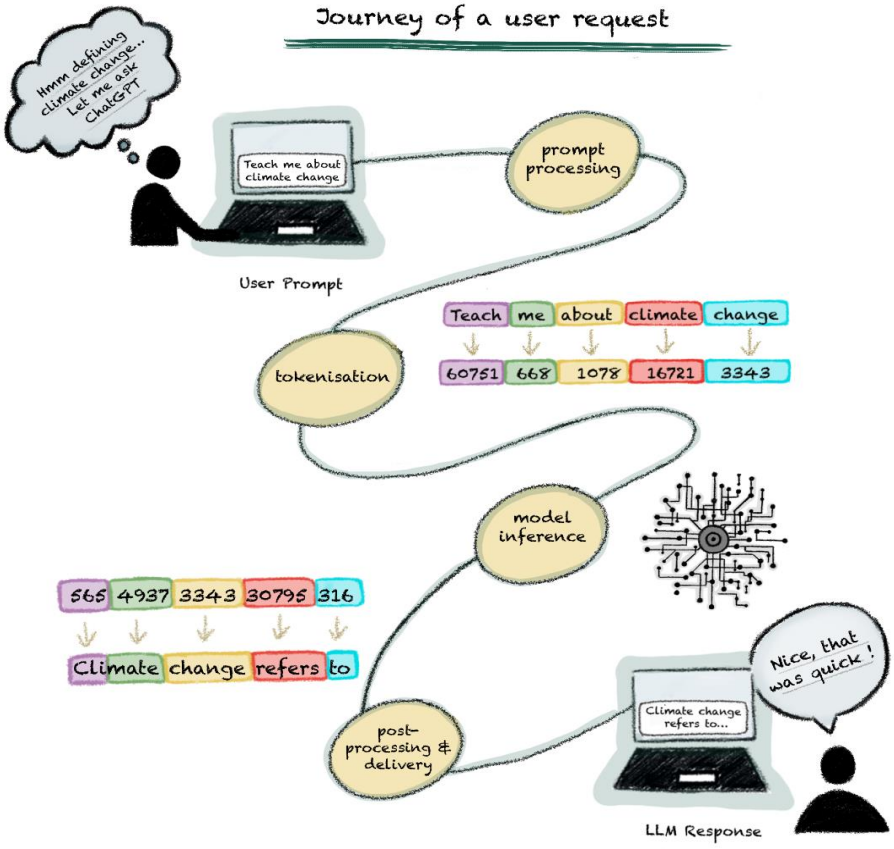
As the growing ubiquity of Generative AI is fueling environmental strain, it is crucial to address inference. As AI tools become deeply embedded in consumer devices, from chatbots like ChatGPT and Claude to AI-powered search engines such as Google AI Overview and advanced business tools like Microsoft 365 Copilot – AI is increasingly "one button away", making usage almost inevitable. For example, Open AI has reported that more than 500 million people use ChatGPT each week[17] with 92% of Fortune 500 companies using OpenAI's product[18]. However, most research to date has focused on the computational intensity of the development processes of AI models, and the exploration of usage is still in early stages.

This section explores the energy demands associated with AI inference and outlines accessible computing strategies that can substantially reduce environmental impact without compromising accuracy or performance. To enable empirical testing, the analysis required an open-source model capable of being run locally on independent hardware. This excluded proprietary, closed-source systems such as ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google DeepMind), which are not publicly accessible. The model selected for experimentation was LLaMA 3.1 8B Instruct, developed by Meta AI. LLaMA 3.1 8B is one of the most widely adopted and accessible open-source models due to its suitability for a range of optimization tests. A selection of additional open-source small models was also included (listed in the Appendix). As the scale of AI technologies continues to expand, enhancing energy efficiency is not only an environmental priority but also critical for ensuring equitable access to AI in underserved settings.

The journey of a single prompt

To fully grasp the energy and resource consumption involved in using Large Language Models (LLMs), it is essential first to map out and understand the journey of a user's request. This journey encompasses all the steps taken from the moment the user sends their query to the final response generated by the LLM. It starts with the user prompt, which is the text entered by the user. The system then moves through several stages:

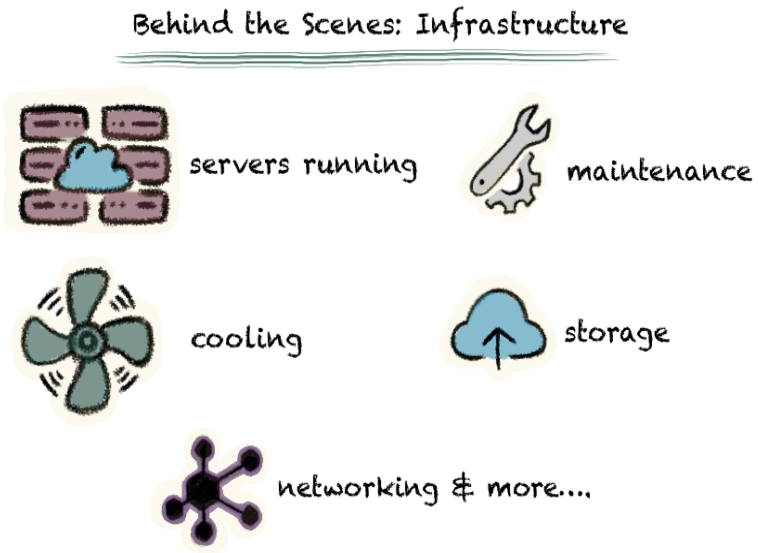
1. **Prompt Processing:** The input is interpreted and prepared for the next steps.
2. **Tokenization**: The prompt is broken down into smaller units called tokens, which the model can understand and process.
3. **Model Inference:** The LLM uses its trained parameters (aka model weights) to analyze the tokens and generate a response by predicting the next word or phrase.
4. **Post-Processing and Delivery:** The raw output is formatted, refined, and sent back to the user in a readable form as the LLM response.



While users only see the immediate interaction–typing a question and receiving a response–the underlying process relies on a vast and complex infrastructure. This includes extracting raw materials like rare earth metals, manufacturing components such as graphics processing units (GPUs), constructing data centers, and maintaining power and cooling systems. Each stage, from the mining of natural resources to the delivery of the final product, requires energy, water, and other resources.

When users query most AI models–whether through mobile apps or web interfaces–what happens behind the scenes remains largely opaque. Key factors such as which data center handles the request, how much energy is consumed in the process, and the carbon intensity of the energy sources involved are typically known only to the companies operating the models.

Reducing energy expenditure of AI use can be achieved by optimizing the entire prompt journey, from input formulation to output generation. Techniques like simplifying prompts, using concise inputs, and managing LLM response length help reduce tokenization and processing overhead. During inference, methods like quantization (e.g., integer format weights) and dynamic scaling (using smaller models for simpler tasks) significantly lower energy use. Post-inference strategies, such as refining outputs incrementally and leveraging prior responses, further minimize redundant computations. These approaches ensure efficiency and sustainability without compromising user experience.

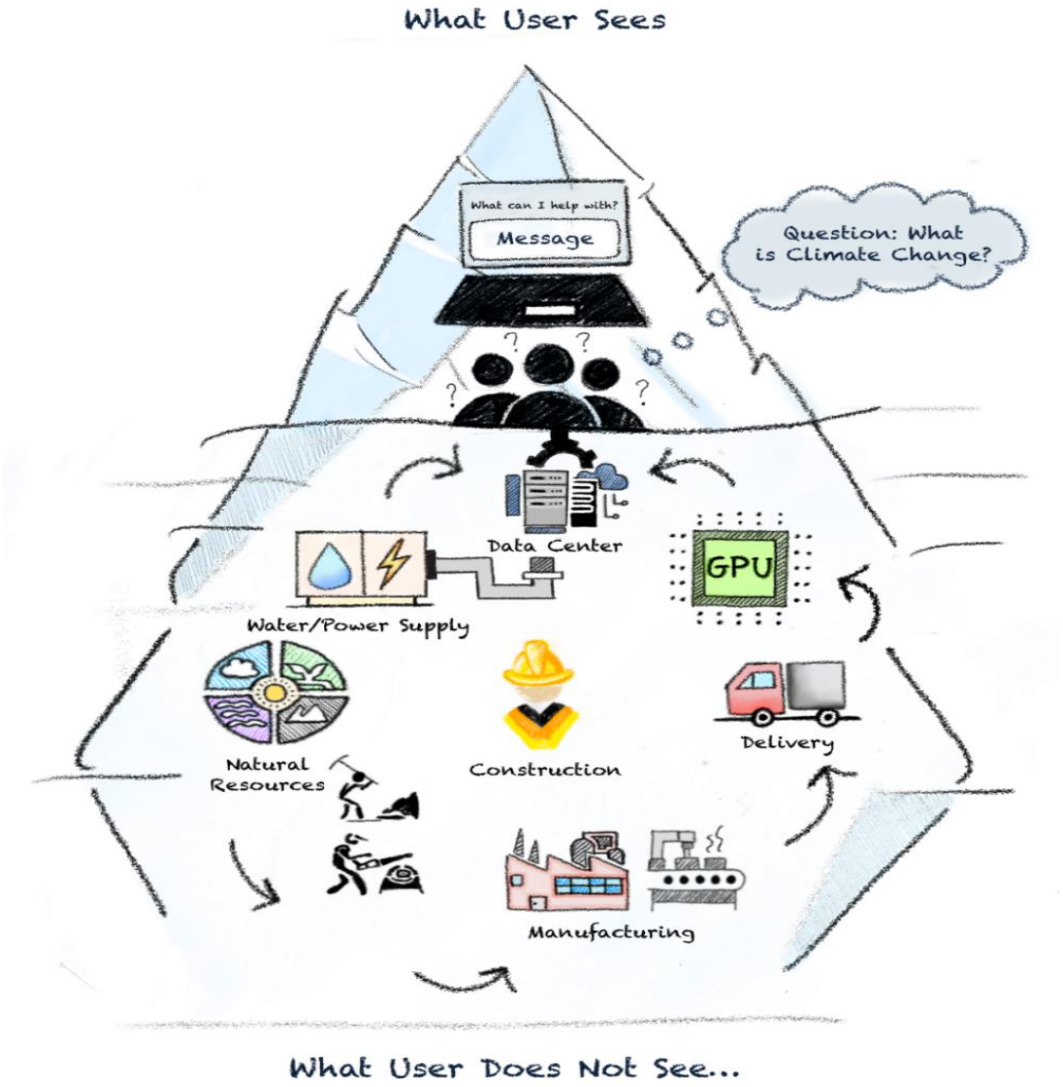


ENERGY-EFFICIENT AI EXPERIMENTS

The following section explores some of these actionable strategies that can be applied to reduce energy costs. The experiments were conducted using open-source models and tools, with the full list and further details on the methodology and data provided in the Appendix.

Energy-efficient Compute Strategy 1: As more compute means energy usage, simplifying and optimizing LLM computations can achieve energy savings.

LLMs generate responses by processing vast amounts of data through interconnected layers of "neurons," much like a brain. These neurons communicate using numbers (called "weights") that determine how the model learns and responds. Optimizing these weights can improve energy efficiency without greatly affecting performance. Experiments were done on Meta AI's LLaMA 3.1 8B Instruct, a state of the art language model designed for summarizing, translating, and answering questions. With 8 billion parameters, it is a large language model adept at handling complex inputs, such as summarizing long documents or answering detailed questions.



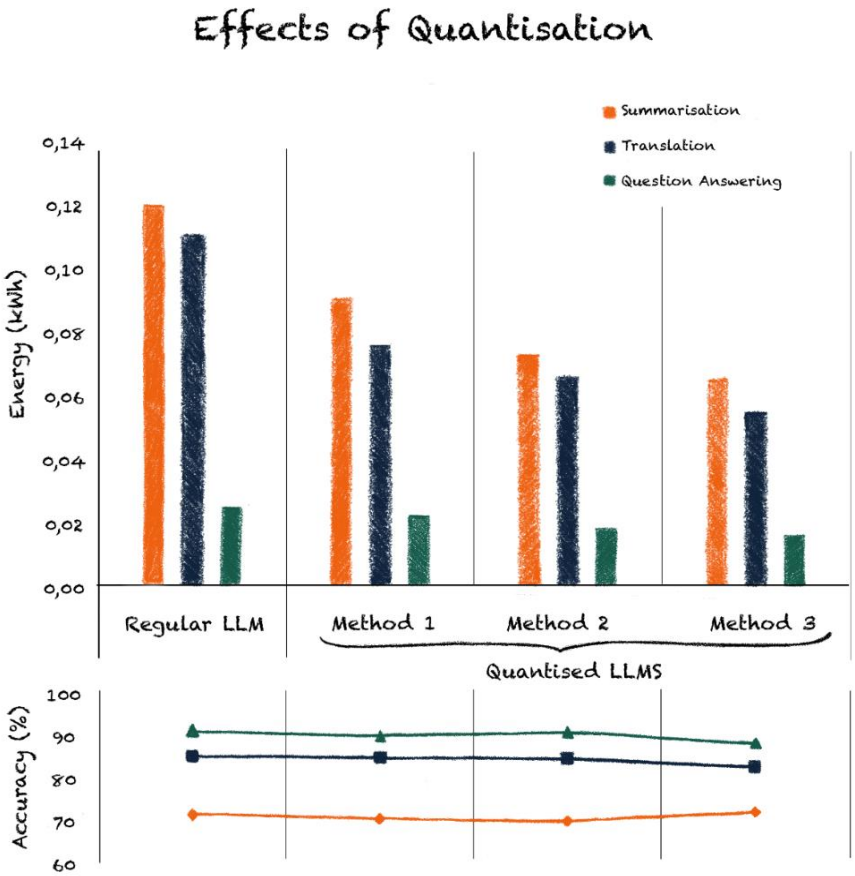
To measure efficiency and performance, the model was tested on three tasks:

1. **Summarization** (e.g., condensing BBC articles into summaries).
2. **Translation** (e.g., English to Hindi).
3. **Question Answering** (e.g., answering questions from science textbooks).

The Appendix shows a few examples of the tasks. The energy consumption was evaluated using CodeCarbon, an open-source tool designed to estimate the carbon emissions associated with computational processes by monitoring power usage, hardware specifications, and the carbon intensity of the energy grid. Performance was assessed using BERTScore, a metric that measures the semantic similarity between model-generated text and human-written answers. BERTScore leverages a pre-trained language model (BERT) to evaluate the meaning and context of the text beyond exact word matches. Together, these tools provided a comprehensive assessment of both environmental impact and task performance.

Energy-efficient Technique: Quantization

Quantization is a technique that reduces the precision of numbers used by machine learning models, thereby accelerating computations and reducing energy consumption. This process can be likened to rounding values, such as reducing 8.01 to 8 or 4.23 to 4, to simplify mathematical operations. Although quantization leads to lower energy consumption, it may also impact the accuracy of the model. To strike a balance between energy savings and performance, three distinct techniques were tested (more details can be found in the Appendix):



1. **Bits and Bytes Quantization (BNBQ)**: This method employs low precision (4 bits) for all model weights, potentially sacrificing accuracy on larger numbers.

2. **Generalized Post-Training Quantization (GPTQ)**: This technique adjusts the precision across the model, preserving essential information to maintain overall accuracy.

3. **Activation Aware Quantization (AWQ)**: This approach prioritizes high precision for critical model components, while reducing precision for less important parts.

Results: Reduced Energy Consumption with Maintained Accuracy

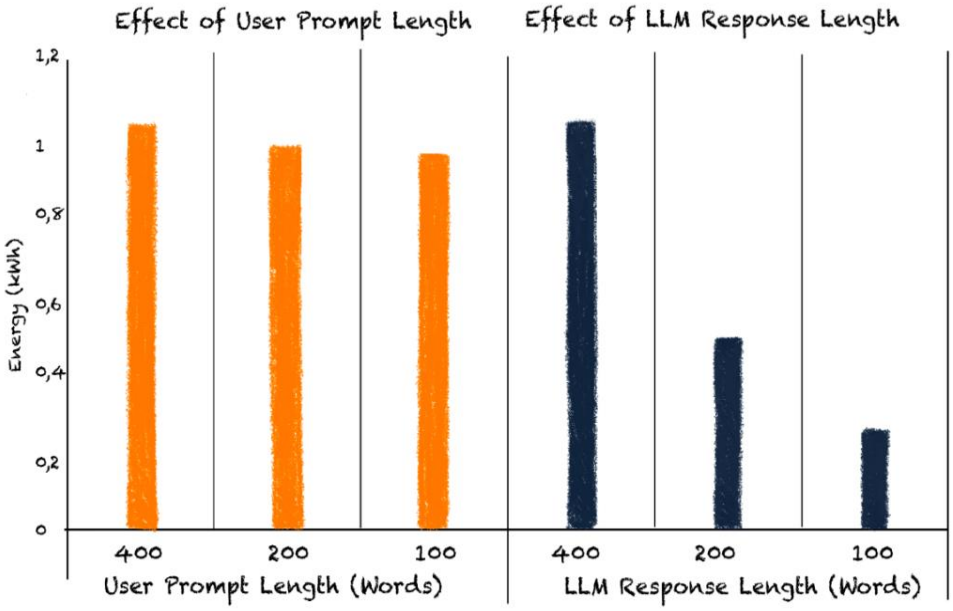
Quantization resulted in significant energy savings while preserving model accuracy:

- **BNBQ**: Achieved a **22%** reduction in energy consumption.
- **GPTQ**: Led to a **35%** reduction in energy consumption.
- **AWQ**: Delivered a **44%** reduction in energy consumption and even outperformed the unquantized model on certain tasks.

While question answering requires more energy than more complex tasks such as summarization or translation, the energy savings with quantisation were similar across all tasks.

Energy-efficient Compute Strategy 2: Optimizing User Prompt and LLM Response Lengths

Energy efficiency during inference is significantly influenced by the length of both the user prompt and the LLM response. Longer inputs and outputs require more tokens to be processed and stored, which increases energy consumption. Tokens are the digital representations of words or parts of words, and tokenization allows models to handle various languages and new words effectively by breaking down text into smaller, reusable units.



User Prompt Length: The energy consumption during inference is directly impacted by the length of the user input prompt. For example, a short prompt like "Summarize this article" (3 tokens) requires minimal energy, whereas a longer prompt, such as "Here is a detailed article about climate change impacts, including specific data points and historical context. Summarize it in five key points" (27 tokens), demands more computation. The model must encode and analyze the additional tokens, resulting in higher energy use.

LLM Response Length: The length of the LLM response also significantly affects energy consumption. For instance, a concise response like "Climate change impacts include rising temperatures, sea levels, and extreme weather" (11 tokens) requires fewer computations compared to a detailed explanation like "Climate change affects the planet by causing rising temperatures, increasing sea levels, and leading to more frequent extreme weather events such as hurricanes and droughts, which disrupt ecosystems and human societies" (36 tokens). Each additional token requires the model to process all previously generated tokens to predict the next one, further increasing energy demands.

Energy Consumption Analysis: Energy consumption (in kWh) was measured for 1,000 prompts across various combinations of input and output lengths, expressed in tokens ranging from 128 tokens (approximately 110 words in English) to 1024 tokens (approximately 850 words or 8 paragraphs). Language differences also play a role, as languages like European languages, Arabic, and Hindi use 1.2–1.5 tokens per word, whereas languages like Chinese, Japanese, and Korean require 1.5–2 tokens per character. These variations significantly impact the computational resources needed for different languages.

The results show that using a longer user prompt (400 words) and receiving a longer LLM response (400 words) spends the most energy (orange bars). Halving the user prompt length to 200 words

reduces the energy expenditure by 5%. On the other hand, halving the LLM response length to 200 words (dark blue bars) reduces energy consumption by 54% (see Appendix for the data). With each further reduction of the user prompt and LLM output length, the energy saving is further compounded.

Notably, the length of the LLM response has a far greater impact on energy consumption than the user input prompt length. This emphasizes the importance of strategies to reduce response length, which can substantially lower energy demands. Some practical approaches to streamline responses include:

1. **Adopt and Mainstream Innovative Solutions:** Emerging technologies, such as prompt compression tools, present effective strategies for enhancing AI efficiency. Research institutes and AI accelerators are developing prompt compression tools that streamline AI workflows. These tools can reduce input size to AI models, decrease response time and costs without compromising output quality. Integrating such technologies can significantly improve the energy efficiency of LLMs while maintaining performance standards.
2. **Promote User Awareness and AI Literacy:** Increasing user awareness of the energy consumption associated with prolonged interactions with AI models, such as chatbots and companions, is critical. Extended dialogues can result in considerable energy expenditure, yet many users remain unaware of this impact. Educating users on the environmental implications of their interactions with AI systems, and encouraging concise engagements, will help mitigate unnecessary energy consumption. Raising awareness will be essential as AI technologies continue to expand across various applications.

Energy-efficient Compute Strategy 3: Small Language Models (SLMs) vs. Large General-Purpose Models

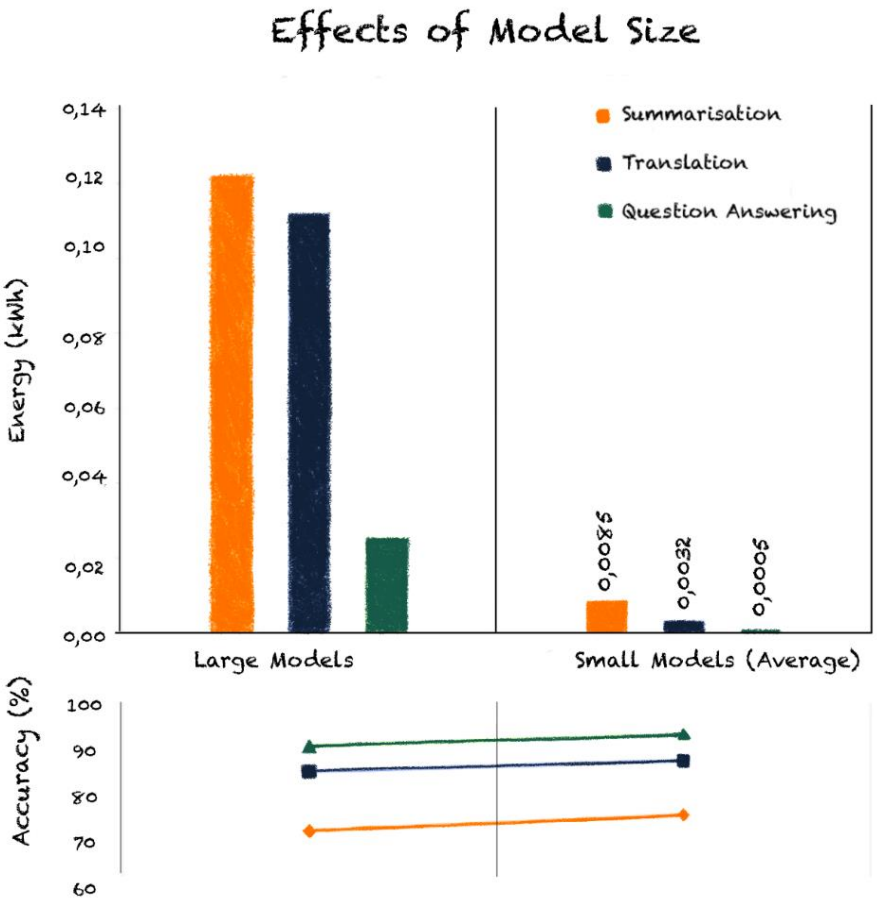
Model size plays a critical role in balancing computational efficiency, energy consumption, and accuracy. A key strategy for optimizing these factors involves replacing large, general-purpose large language models (LLMs) with smaller, fine-tuned models that are specifically optimized for designated tasks. Fine-tuning adapts a pre-trained model using domain-specific datasets, tailoring it for particular applications such as legal document analysis or medical diagnosis. While these smaller models may not match the broad capabilities of larger LLMs, they excel in targeted applications, significantly reducing both computational and energy costs. This approach proves to be particularly effective in low-resource environments where computational efficiency is a priority.

In this experiment, a comparison was made between the unquantized (original) LLaMA model and task-specific models that had been fine-tuned for each respective task. The task-specific models were selected based on their relevance to common applications and their popularity within the Hugging Face community, ensuring that the evaluation reflected models that are widely recognized and used.

The fine-tuned models were evaluated on two types of datasets: in-context (an extension of the dataset used for fine-tuning) and out-of-context (a slightly varied dataset that tested the same task but introduced some differences). This comparison aimed to determine whether specialized models could deliver higher performance and greater energy efficiency than a larger, general-purpose model.

The results, as shown in the figure below, indicate that task-specific models significantly outperform the general-purpose LLaMA model when tested on their fine-tuned datasets. These specialized

models consumed between 15 to 50 times less energy while producing higher-quality outputs on the in-context datasets. These findings have major implications for low-resource settings, where smaller models, delivering similar or even better performance than larger models, are more viable and cost-effective for smaller infrastructures.



However, the performance of the smaller models decreased when applied to unfamiliar datasets, highlighting the trade-off between task-specific optimization and generalization. Additionally, they were found to be less suitable for applications that require multi-tasking capabilities.

The optimal approach involves combining smaller models for well-defined tasks with available fine-tuning data, especially by leveraging the growing ecosystem of open-source AI. Platforms like Hugging Face, Replicate, and Mistral offer pre-trained models and collaborative tools that empower developers to build efficient, tailored solutions without starting from scratch. This open-source movement is central to enabling more resource-conscious innovation, accelerating access, and fostering transparency across the AI ecosystem.

This strategy demonstrates that smaller, fine-tuned models can provide significant energy and performance benefits for specific applications, particularly in low-resource settings. However, limitations in generalization and multi-tasking should be considered. A balanced approach, leveraging small task-specific models, a “collection of small models” approach, and large general-purpose models when necessary, will offer the best optimization for energy and performance.

Translating green compute strategies into relatable energy consumption units

Impact of LLM Optimization on Energy Consumption in Relatable Units

Putting the energy-efficient compute strategies into practice, this part focuses on translating the energy consumption of LLMs into familiar and relatable units to underscore the potential impact of optimization. This combines findings from the energy consumption of the LLaMa model, as outlined in Strategies 1-3, with global usage statistics for ChatGPT, at the time this brief was written the most widely used LLM-based chatbot. This approach bridges experimental results with real-world scenarios, illustrating both the consequences of inefficient LLM usage and how optimization can yield significant energy savings in easily understandable terms.

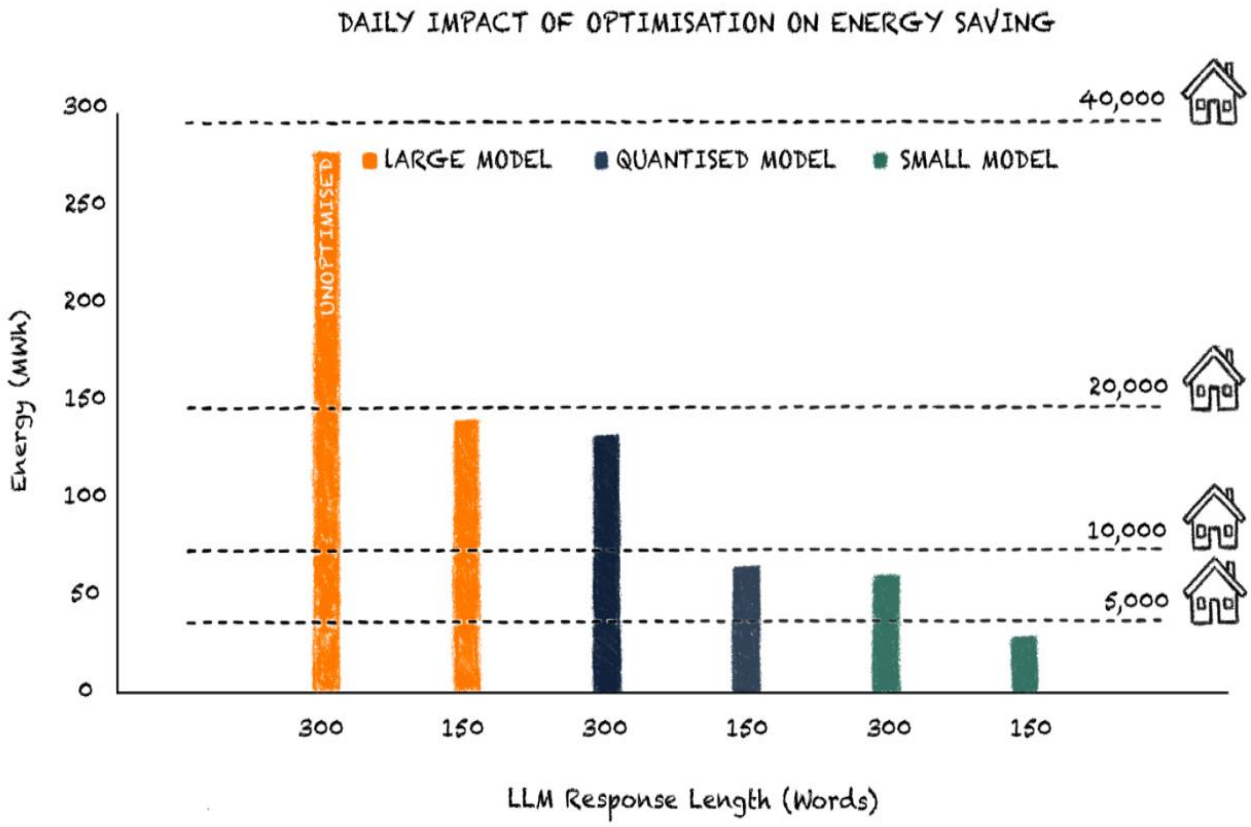
For the purposes of this analysis, the "concept explanation" task was selected, where a user requests an LLM to explain a specific concept. This task represents nearly 35% of user queries. The following input prompt was used: "Explain the concept of reinforcement learning, emphasizing its core principles, components (like agents, environments, and rewards), and typical applications. Keep the explanation accessible to someone with basic knowledge of artificial intelligence." This prompt was designed to assess the model's factual knowledge and ability to follow instructions.

Energy consumption was then calculated for a single LLM model response to this question using the Lllama model described earlier. This figure was multiplied by the number of daily user prompts for concept explanation queries. Based on global usage statistics for ChatGPT, which shows approximately 1 billion daily requests, and assuming 35% of these to be concept explanations, we get a total number of requests of this type to be 350 million. This allowed for the estimation of total daily energy consumption for concept explanation tasks globally.

The findings, illustrated in the next figure, demonstrate the energy savings across various optimization scenarios. Reducing the LLM response from 400 tokens (roughly 300 words) to 200 tokens (approximately 150 words) could save enough electricity to power approximately 20,000 average UK households per day (assuming 7.4 kilowatt-hours per household per day [19]. Furthermore, using a quantized model yields similar results, saving the equivalent of an additional 10,000 households' worth of electricity.

When both strategies–quantized models and reduced response lengths–are combined, the energy savings are compounded, reducing energy expenditure by up to 75%, equating to approximately 30,000 U.K. households per day.

Interestingly, the use of smaller models yields the largest energy savings. Even without reducing response length, smaller models save more energy than the combined approach above. **When response length is also reduced, the energy consumption drops to fewer than 4,000 households, representing a 90% reduction from the original 38,000 households.**



However, it is important to note that while smaller models outperform larger ones for specific tasks, they may not perform as well for general-purpose applications. As such, optimizing model size requires careful consideration to maintain output quality.

These findings reinforce that sustainability does not always require sacrificing performance. With smart design choices, AI can become dramatically more efficient.

FUTURE OF GEN AI: Smaller, Smarter, Stronger

New Frontiers in Efficient AI: New Architectural Strategies

Beyond individual tweaks, new architectural strategies are emerging that rethink how generative AI models are structured.

One promising direction is the **Mixture of Experts (MoE)** approach. Rather than relying on a single model to handle every task, MoE systems activate only a few specialised small models, or "experts," for each query. This targeted use of compute avoids unnecessary energy use and allows for scale without a proportional increase in cost - much like consulting just the right expert instead of convening an entire panel for every question. **Multi-agent systems** offer another compelling strategy. In these architectures, multiple smaller models, each with distinct capabilities, work together to solve complex tasks. Inspired by human collaboration, this approach enables more flexible and modular AI systems that can be tailored to specific domains or responsibilities.

A third strategy gaining traction is **sparse and conditional computation**, where models learn to activate only the components needed for a particular input. These architectures, such as sparse transformers and early-exit models, can significantly reduce computational waste while maintaining high performance. In parallel, **retrieval-augmented generation (RAG)** provides a way to combine small generative models with search and memory. Instead of trying to store all knowledge within the model itself, RAG systems retrieve relevant external information on demand—enabling more accurate and efficient responses without massive parameter counts.

Finally, **neurosymbolic and brain-inspired architectures** are being explored as long-term solutions. These systems integrate symbolic reasoning with neural networks or take cues from the brain's energy-efficient mechanisms, such as spiking neurons or event-driven processing. Though still in their early stages, they open new possibilities for interpretable, low-power AI.

These emerging approaches hold great promise, though many remain under active development. Mixture of Experts systems must be carefully calibrated to ensure that expert models are used efficiently. Multi-agent architectures face challenges related to coordination, communication, and system stability. Sparse models, RAG techniques, and brain-inspired designs require further progress in infrastructure, scalability, and integration into real-world applications.

To move these ideas from lab experiments to practical deployment, sustained investment will be critical—along with robust evaluation standards and ecosystem support. When mature, these

architectural innovations promise a new generation of AI systems that are not only highly capable and intelligent, but also efficient and sustainable by design.

Recent research from NVIDIA on small language models[20] further underscores growing industry momentum behind modular, energy-aware AI systems capable of efficient reasoning and collaboration. Such innovations reflect a broader shift toward performance-optimized AI that doesn't come at the expense of sustainability.

RECOMMENDATIONS

To address the urgent resource-efficiency and accessibility challenges posed by Generative AI, we propose the following three main recommendations, accompanied by targeted actions for **policymakers, the technology sector**, and **end users**:

1. Mobilize innovation to advance resource-efficient, "clean by design" AI systems

Bold public and private investment are essential to accelerate the development and deployment of AI systems that are sustainable by design. This includes channeling support toward innovations that improve both energy and resource efficiency throughout the AI lifecycle, from training to inference. Prioritizing "clean by design" approaches ensures that efficiency and sustainability are not add-ons, but foundational design principles.

Improving both energy and cost efficiency in AI inference operations is vital to ensuring the accessibility of generative AI, particularly for underserved communities.

- **Policymakers should:**
    - Establish national R&D funding programs to support research into energy-efficient AI techniques, such as model compression, quantization, and inference optimization. These should level the playing field between academia, startups, and large industry players.
    - Establish targeted funding & incentives: Provide grants or tax relief for energy-efficient AI solutions to encourage their adoption. Similar to clean energy subsidies, targeted incentives can help bridge the gap between traditional AI deployments and greener alternatives.
    - Public Procurement: Governments could lead by example by adopting (and requiring) low-carbon AI systems in public services. This signals the viability of such solutions and validates them for industry use.
    - Standards & Regulatory Clarity: Establish benchmarks or guidelines around AI energy usage. Clear criteria can encourage responsible innovation whilst maintaining a level playing field.
    - Mandate environmental risk assessments for the development and deployment of AI systems, particularly large-scale or high-energy models. These assessments should evaluate impacts on electricity grids, water usage, and emissions, especially in regions with fragile infrastructures or scarce resources.

- Require companies to publicly disclose environmental risk assessments as part of AI model approval, similar to impact assessments required in other high-risk sectors (e.g., construction or energy).
- Collaborate with international bodies to create and promote global benchmarks for energy-efficient AI.

- **AI Developers and the Technology Industry should:**
  - Proof of Concept & Pilot Programmes: Enterprises can start small by testing efficiency tools in isolated pilots before rolling them out widely. Positive results in cost and performance provide a compelling case for larger investment. This includes integrating energy-efficient techniques, such as model optimization, quantization, and task-specific model development, as standard practices.
  - Advance open-source sustainability frameworks: Support initiatives such as Hugging Face's *Reduce, Reuse, Recycle* and *Energy Score* efforts to promote transparent benchmarking, shared tools, and procurement standards for sustainable AI development.
  - Design and implement low-energy inference modes (e.g., "Eco Modes") for common AI tasks.
  - Collaboration with Start-ups: Large AI companies or cloud providers might partner with emerging start-ups to integrate next-generation efficiency features, speeding up mainstream adoption.
  - Open Knowledge-Sharing: Contribute to open-source libraries or best-practice guides on AI inference optimization. This fosters a community-driven approach and accelerates innovation across the board.
  - Conduct and publish environmental risk assessments before launching new models or infrastructure expansions (e.g., large data centers, major model updates). These assessments should cover energy intensity, water usage, emissions impact, and mitigation strategies.
  - Develop internal governance mechanisms to systematically assess and reduce environmental risks in product roadmaps and model iteration cycles.
- **General Users should:**
  - Support AI platforms that explicitly commit to energy-efficient practices.
  - Support research and advocacy efforts focused on making AI systems more sustainable and accessible.
  - Use AI tools judiciously by selecting smaller, task-specific models or tools that align with sustainable practices.

2. Incentivize transparency and eco-conscious innovation across the AI ecosystem

Transparency and accountability must become central to how AI is designed, deployed, and used. Clear sustainability incentives, visibility mechanisms, and procurement guidelines can drive more responsible behavior across the ecosystem, from model developers to end users.

- **Policymakers should:**
    - Create procurement standards that incentivize and reward sustainable-by-design models, especially in public sector adoption of generative AI tools.
    - Develop regulatory frameworks that require the disclosure of environmental metrics on energy usage, carbon emissions, and water consumption by AI providers.
    - Mandate standardized reporting formats for environmental performance data to ensure comparability and foster informed public and private sector decision-making.
    - Incentivize eco-conscious innovation via certifications, labeling systems (akin to appliance efficiency ratings), and tax incentives for sustainable AI.
- **AI Developers and the Technology Industry should:**
    - Transparently report the environmental footprint of their models and infrastructure, including inference-level energy use, water demand, and emissions.
    - Adopt sustainability scoring mechanisms (e.g., AI energy performance scores) that reflect real-time model efficiency across common tasks. These scores can support institutional procurement, compliance monitoring, and user awareness.
    - Publish supporting documentation that allows external reviewers to verify and replicate energy usage claims under reproducible test conditions.
    - Optimize data centers and AI pipelines for energy efficiency, and integrate renewable energy sources wherever feasible.
    - Embed eco-conscious design choices into development workflows, including quantization, sparsity, and modular inference approaches.
- **General Users should:**
    - Favor AI tools and platforms that are transparent about their environmental impact, including energy and resource use, and that clearly disclose sustainability practices.
    - Choose lightweight, purpose-built models over large general-purpose ones when practical, especially for routine tasks.
    - Look for models or applications that display sustainability labels, efficiency scores, or metrics aligned with public benchmarks.
    - Support platforms and services that embed eco-conscious design principles–such as quantization, sparsity, and modular inference–and prioritize sustainability as a core value.

3. Promote AI Literacy and Awareness of Generative AI's Environmental Costs

Education initiatives that highlight both the capabilities and environmental costs of Generative AI are critical to fostering informed decision-making.

- **Policymakers should:**
    - When setting standards, developing national policies and educational curriculum on AI, governments should support the integration AI information literacy to advance critical thinking of citizens and foster user empowerment. Policies should guarantee the training of citizens of all ages to understand AI (for e.g. information search techniques) and critically engage with Gen AI tools, including by raising awareness of its environmental and social impacts
    - Fund public awareness campaigns that explain the energy and resource costs of AI operations and encourage sustainable AI usage.
    - Collaborate with AI developers, researchers, and educators to create inclusive and accessible educational materials tailored to diverse audiences.
- **AI Developers and the Technology Industry should:**
    - Prioritize transparency and explainability in the development and deployment of generative AI tools, disclosing data on environmental impact, resource consumption, fairness, and safety.
    - Provide user-friendly transparency reports on the environmental impacts of their models, including energy consumption and carbon emissions per query.
    - Develop tools or dashboards to help users visualize the environmental impact of their interactions with AI systems.
    - Partner with educational institutions to create accessible AI training materials, focusing on sustainability and ethical use.
- **General Users should:**
    - Learn about the environmental impact of AI and use this knowledge to make informed choices about platform selection and usage patterns.
    - Engage with educational resources, workshops, or online courses on AI literacy and sustainability.
    - Advocate for sustainable AI by supporting companies that prioritize transparency and green practices.

CONCLUSION & FUTURE WORK

Tackling the resource demands of generative AI requires more than incremental fixes, it calls for a shift in how we design, deploy, and govern AI technologies. By addressing critical areas such as energy efficiency in AI inference, advancing AI literacy, and fostering greater transparency, we can begin to mitigate the environmental and accessibility challenges posed by today's AI systems.

A new paradigm is needed–one that prioritizes *sustainable by design* innovation, embedding efficiency, inclusivity, and environmental responsibility at the core of AI development. This means accelerating the deployment of smarter, more resource-conscious AI systems, particularly at the inference stage where the bulk of energy use occurs. It also means empowering users, policymakers, and developers alike through education, open information, and clear sustainability benchmarks.

Such transformation will require coordinated, cross-sectoral collaboration. Governments, technology developers, research institutions, and end-users each have a vital role to play in building an AI ecosystem that is not only powerful but also sustainable, equitable, and responsive to global challenges.

As part of its continued commitment to responsible AI, UNESCO will champion the adoption of energy- and resource-efficient AI solutions–such as small language models, agentic AI, frugal AI, and edge computing–as core enablers of sustainable digital transformation. Working in partnership with the scientific and technology communities, UNESCO will help develop practical policy tools and technical guidance to support low-resource environments, ensuring that innovation remains both inclusive and ecologically sound.

By steering AI toward a cleaner, leaner, and fairer future, we can unlock its full potential–not just for technological progress, but for the shared sustainability and resilience of our societies.

REFERENCES

[1] A. Dannouni, S. A. Deutscher, G. Dezzaz, A. Elman, A. Gawel, M. Hanna, A. Hyland, et al., “How AI can speed climate action,” Boston Consulting Group, Nov. 15, 2023.

[2] “Energy and AI report,” International Energy Agency, 2025.

[3] “AI energy usage and climate footprint: Big Tech’s impact,” MIT Technology Review, May 20, 2025.

[4] S. Altman, “The gentle singularity,” Sam Altman Blog, 2025.

[5] “Ethiopia country profile,” Enerdata, 2025.

[6] Electric Power Research Institute, “Powering intelligence: Analyzing artificial intelligence and data center energy consumption,” May 28, 2024.

[7] S. Zhu, T. Yu, T. Xu, H. Chen, S. Dustdar, S. Gigan, D. Gunduz, E. Hossain, Y. Jin, F. Lin, et al., “Intelligent computing: The latest advances, challenges, and future,” Intell. Comput., vol. 2, p. 0006, 2023, doi: 10.34133/icomputing.0006.

[8] “Energy and AI report,” International Energy Agency, 2025.

[9] “AI energy usage and climate footprint: Big Tech’s impact,” MIT Technology Review, May 20, 2025.

[10] “Artificial intelligence: How much energy does AI use?,” United Nations Western Europe (UNRIC), 2024.

[11] “Global internet use continues to rise but disparities remain, especially in low-income regions,” International Telecommunication Union (ITU), Nov. 27, 2024.

[12] A. Tsado and C. Lee, “Only five percent of Africa’s AI talent has compute power it needs,” United Nations Development Programme, Nov. 12, 2024.

[13] M. Donegan, “Data center investment on the rise in Africa,” TM Forum, May 23, 2024.

[14] A. Dannouni, S. A. Deutscher, G. Dezzaz, A. Elman, A. Gawel, M. Hanna, A. Hyland, et al., “How AI can speed climate action,” Boston Consulting Group, Nov. 15, 2023.

[15] M. Zao-Sanders, “How people are really using GenAI,” Harvard Bus. Rev., Mar. 19, 2024.

[16] ISO, “ISO/IEC TS 12791:2024(En), Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks,” 2024.

[17] I. Fried, “OpenAI says usage has doubled in the last year,” Axios, Aug. 29, 2024.

[18] E. Roth, “ChatGPT’s weekly users have doubled in less than a year,” The Verge, Aug. 29, 2024.

[19] C. Topping, “Average electricity usage in the UK: How many kWh does your home use?,” OVO Energy, Aug. 23, 2024.

[20] P. Belcak, G. Heinrich et al., "Small Language Models are the Future of Agentic AI," Arvix, June 2, 2025. https://doi.org/10.48550/arXiv.2506.02153.

APPENDIX

A.1

Task	Model	Developer	Model Size (GB)	Model Parameters (millions)
All	Llama-3.1-8B-Instruct	Meta Llama	16.06	8,030
All	Llama-3.1-8B-Instruct-BNB	Hugging Quants	5.59	8,030
All	Llama-3.1-8B-Instruct-GPTQ	Hugging Quants	5.74	8,030
All	Llama-3.1-8B-Instruct-AWQ	Hugging Quants	5.86	8,030
Summarisation	pegasus-xsum	Google	2.28	570
Summarisation	bart-large-xsum	AI at Meta	1.63	406
Summarisation	distilbart-xsum-12-6	AI at Meta / Sam Shleifer	1.22	306
Summarisation	t5-small-xsum	Google T5 / Abhijit Das	0.24	61
Translation	opus-mt-en-es	Language Technology Research Group at the University of Helsinki (Helsinki NLP)	0.31	78
Translation	opus-mt-es-en		0.31	78
Translation	opus-mt-en-zh		0.31	78
Translation	opus-mt-zh-en		0.31	78
Translation	opus-mt-en-hi		0.31	76
Translation	opus-mt-hi-en		0.31	76
Translation	opus-mt-en-uk		0.30	76
Translation	opus-mt-uk-en		0.30	76
Question Answering	bert-large-uncased-whole-word-masking-finetuned-squad	Google BERT	1.34	335
Question Answering	bert-large-uncased-whole-word-masking-squad2	Google BERT	1.34	335
Question Answering	mdeberta-v3-base-squad2	Microsoft / Tim Isbister	1.11	278
Question Answering	roberta-base-squad2	Facebook AI / Deepset	0.50	125
Question Answering	electra-base-squad2	Facebook AI / Deepset	0.44	
Question Answering	tinyroberta-squad2	Facebook AI / Deepset	0.33	82
Question Answering	distilbert-base-uncased-distilled-squad	HuggingFace	0.27	66
Question Answering	distilbert-base-cased-distilled-squad	HuggingFace	0.26	65

A.2

Feature	BNBQ	GPTQ	AWQ
Precision	4-bit weights, 16-bit activations	4-bit weights, 16-bit activations	Mixed precision: higher precision for critical weights, lower for others
Adaptability	Static 4-bit quantisation	Layer-specific dynamic quantisation	Importance-aware quantisation based on activation impact
Focus	Compact storage, energy efficiency	Accuracy preservation, task-specific	Balancing performance and efficiency by prioritising critical components
Retraining	Not required	No retraining, but more task-specific	Not required but uses activation data for fine-tuning decisions
Ideal Use Case	General-purpose, storage-constrained	High-precision tasks with complex needs	Optimising energy efficiency while maintaining performance for critical tasks

A.3

All experiments were conducted on an NVIDIA GeForce RTX 3090 Ti GPU with 24GB of RAM, combined with a 12th Gen Intel Core i9-12900K CPU1 and 128GB of RAM. The Linux operating system (version 5.14.0-427.31.1.el9 4.x86 64) provided a stable environment for model execution, with Python version 3.10.14 used for managing the experiments.

Effect of quantisation:

	Unquantised	**Quantised**		
Energy (kWh)/1000 inferences		**BNBQ**	**GPTQ**	**AWQ**
Summarisation	0.122	0.091	0.074	0.066
Translation	0.112	0.077	0.067	0.056
Question Answering	0.025	0.023	0.019	0.016
Accuracy (%)				
Summarisation	70.5	69.5	69.0	71.2
Translation	84.2	83.8	83.7	81.7
Question Answering	90.0	89.1	89.9	87.3

Effect of model size:

	Large Model	Small Models
Energy (kWh)/1000 inferences		
Summarisation	0.122	0.008
Translation	0.112	0.003
Question Answering	0.025	0.001
Accuracy (%)		
Summarisation	70.5	74.2
Translation	84.2	86.6
Question Answering	90.0	92.7

Effect of lengths of user prompt and LLM response:

	Words		
Energy (kWh)/1000 inferences	**400**	**200**	**100**
User Prompt	1.031	0.980	0.955
LLM Response	1.031	0.476	0.248

Impact of optimisation on energy saving:

	Words	Energy (kWh) per inference	Daily energy usage, 350M inferences (MWh)
Large Model	300	0.0008	278.5
	150	0.0004	141.0
Quantised Model	300	0.00038	133.9
	150	0.00019	65.4
Small Model	300	0.00017	60.5
	150	0.00008	29.0

A.4 *Examples from the task data set (translations, summarizations etc.):*

- XSum:
  - **Prompt**:

    “They found that targeting a part of the brain called the parietal lobe improved the ability of volunteers to solve numerical problems. They hope the discovery could help people with dyscalculia, who may struggle with numbers. Another expert said effects on other brain functions would need checking. The findings are reported in the journal Current Biology. Some studies have suggested that up to one in five people have trouble with maths, affecting not just their ability to complete problems but also to manage everyday activities such as telling the time and managing money. Neuroscientists believe that activity within the parietal lobe plays a crucial role in this ability, or the lack of it. When magnetic fields were used in earlier research to disrupt electrical activity in this part of the brain, previously numerate volunteers temporarily developed discalculia, finding it much harder to solve maths problems. The latest research goes a step further, using a one milliamp current to stimulate the parietal lobe of a small number of students. The current could not be felt, and had no measurable effect on other brain functions. As it was turned on, the volunteers tried to learn a puzzle which involved substituting numbers for symbols. Those given the current from right to left across the parietal lobe did significantly better when given, compared to those who were given no electrical stimulation. The direction of the current was important - those given stimulation running in the opposite direction, left to right, did markedly worse at these puzzles than those given no current, with their ability matching that of an average six-year-old. The effects were not short-lived, either. When the volunteers whose performance improved was re-tested six months later, the benefits appear to have persisted. There was no wider effect on general maths ability in either group, just on the ability to complete the puzzles learned as the current was applied. Dr Cohen Kadosh, who led the study, said: "We are not advising people to go around giving themselves electric shocks, but we are extremely excited by the potential of our findings and are now looking into the underlying brain changes. "We've shown before that we can induce dyscalculia, and now it seems we might be able to make someone better at maths, so we really want to see if we can help people with dyscalculia. By Fergus WalshMedical correspondent, BBC News Read more in Fergus's blog "Electrical stimulation is unlikely to turn you into the next Einstein, but if we're lucky it might be able to help some people to cope better with maths." Dr Christopher Chambers, from the School of Psychology at Cardiff University, said that the results were "intriguing", and offered the prospect not just of improving numerical skills, but having an impact on a wider range of conditions He said: "The ability to tweak activity in parts of the brain, turning it slightly 'up' or 'down' at will, opens the door to treating a range of psychiatric and neurological problems, like compulsive gambling or visual impairments following stroke." However, he said that the study did not prove that the learning of maths skills was improved, just that the volunteers were better at linking arbitrary numbers and symbols, and he warned that researchers needed to make sure other parts of the brain were unaffected. "This is still an exciting

new piece of research, but if we don't know how selective the effects of brain stimulation are then we don't know what other brain systems could also be affected, either positively or negatively." Sue Flohr, from the British Dyslexia Association, which also provides support for people with dyscalculia, said the research was welcome. She said: "It's certainly an under-recognised condition, but it can ruin lives. "It makes it very hard to do everyday things like shopping or budgeting - you can go into a shop and find you've spent your month's money without realising it.'"”

- **Response**:
  “Applying a tiny electrical current to the brain could make you better at learning maths, according to Oxford University scientists.”

- SamSum
  - **Prompt**:
    “Dialogue: Claire: hey
    Claire: where's this conference about relations at school?
    Aaron: hey
    Aaron: it's in the teaching training thing
    Claire: oh, in Sesame Street?
    Aaron: yeah exactly
    Claire: thanks!
    Claire: are you going? i mean apart from your own lecture ;D
    Aaron: haha yeah i'm giving two, actually
    Aaron: and i'm kind of a host there as well
    Aaron: so guess i'll be looking to coffee, biscuits etc
    Claire: i see
    Aaron: i was thinking i could go to Jo's lecture
    Claire: Jo's... is it the one about stress?
    Aaron: no, that's Joe ;D Jo is having a workshop on teamwork
    Claire: ah, this one. yeah looks interesting
    Aaron: i know her, she's a really good coach
    Claire: i was thinking maybe mindfullness...
    Aaron: yeah definitely!! if you haven't been to any mindfullness workshops then it's a must!
    Aaron: i've taken it in Berlin last year, with the same guy
    Aaron: it was amazing, life-changing i'd say
    Claire: sounds wonderful
    Claire: one can easily tell you're one of the organisers ;D
    Aaron: hahah yeah but we did choose the best topics and presenters, believe me :D
    Aaron: we've checked out all those people and learnt from them and yeah, they're the best ;)
    Claire: ok ok i'm signing up :D see you there, during the breaks at least;)
    Aaron: ok bye bye;)”
  - **Response**: “Aaron will give two lectures during the conference about relations at school at the Sesame Street. Claire will participate in mindfulness workshop.”

- Opus:
    - **Prompt**: “The general view was that the Assembly should continue this work, notwithstanding the fact that efforts made in this area for more than a decade had not yielded any tangible results.”
    - **Response**: “La opinión general es que la Asamblea debería continuar esta labor, a pesar de que los esfuerzos realizados en este ámbito durante más de un decenio no han producido resultados tangibles.”
- Flores:
    - **Prompt**: “Although three people were inside the house when the car impacted it, none of them were hurt.”
    - **Response**: “A pesar de que había tres personas en el interior de la vivienda en el momento del impacto del automóvil, ninguna de ellas sufrió lesiones.”

- SciQ:
    - **Prompt**: “Context: Type 1 diabetes and other autoimmune diseases cannot be prevented. But choosing a healthy lifestyle can help prevent type 2 diabetes. Getting plenty of exercise, avoiding high-fat foods, and staying at a healthy weight can reduce the risk of developing this type of diabetes. This is especially important for people who have family members with the disease.
      Question: What disease is unpreventable in the type one form but may be prevented by diet if it is of the second type?
      Choices: 1) TB 2) diabetes 3) Cancer 4) Obesity”
    - **Response**: “diabetes”
- SQuAD:
    - **Prompt**: “Context: CBS broadcast Super Bowl 50 in the U.S., and charged an average of $5 million for a 30-second commercial during the game. The Super Bowl 50 halftime show was headlined by the British rock group Coldplay with special guest performers Beyonce and Bruno Mars, who headlined the Super Bowl XLVII and Super Bowl XLVIII halftime shows, respectively. It was the third-most watched U.S. broadcast ever. Question: Who was the main performer at this year's halftime show?”
    - **Response**: “Coldplay”

Smarter, Smaller, Stronger: Resource-Efficient Generative AI & the Future of Digital Transformation

Smarter, Smaller, Stronger: Resource-Efficient Generative AI and the Future of Digital Transformation explores how simple, evidence-based techniques can dramatically reduce the energy footprint of generative AI systems, without sacrificing performance.

Backed by original experiments from UNESCO and UCL, the report highlights three scalable strategies: model quantization, shorter prompts and outputs, and the use of smaller, task-specific AI models. These techniques can cut energy use by up to 90%, offering a practical roadmap for more sustainable and accessible AI.

As generative AI becomes embedded in daily life, this report offers timely solutions to ensure digital transformation is both equitable and climate-conscious.