WANT TO KNOW ABOUT AI?

THE AI ASSEMBLY

🖐 Your AI Footprint:

A 5-Step Guide to Energy-Efficient Prompting

www.theaiassembly.com

WANT TO KNOW ABOUT AI?

🖐 Your AI Footprint:
A 5-Step Guide to Energy-Efficient Prompting

The race to build bigger, faster AI models is consuming massive amounts of energy and water. While corporations debate the future of "gigawatt clusters" and "Manhattan-sised" data centers, you, the user, have the immediate power to reduce the environmental cost of AI.

At The AI Assembly, we believe that becoming an expert user means becoming an efficient user. Every time you interact with a generative AI tool (like a chatbot or image generator), you initiate a burst of computation that consumes energy.

Here are five practical, immediate steps you can take to be a more mindful and sustainable AI user:

Match the Model to the Job: Don't Use a Sledgehammer for a Nail

Not every task requires the world's largest, most expensive AI model.

- **The Problem**: Using a massive, trillion-parameter model (like GPT-4) to summarise a two-paragraph email is like using a supercomputer for simple arithmetic. It wastes power./
- **The Fix**: Choose the smallest, most specialised model that can handle your task. Use smaller, fast models (like GPT-3.5 or specialised, open-source models) for simple classification, summarisation, or short text generation. Research shows that for repetitive tasks, switching to a smaller, specialised model can lead to up to 90% energy reduction (UCL News Report).

💡 Pro Tip: If your company offers different AI tiers (e.g., "fast mode" vs. "advanced mode"), use the faster, less powerful option unless complexity is absolutely necessary.

WANT TO KNOW ABOUT AI?

🖐 Your AI Footprint:
A 5-Step Guide to Energy-Efficient Prompting

Master the Minimum Viable Prompt (MVP)

Every word you type into an AI, and every word it generates, is converted into data chunks called tokens, which directly translate to energy consumption.

- **The Problem**: Long, rambling prompts filled with unnecessary context or conversational filler are inefficient.
- **The Fix**: Be Concise and Clear. Get straight to the point by clearly stating the task, format, and goal.

Inefficient Prompt ❌
"Hi there, I was wondering if you could help me out. Can you tell me what the three biggest benefits of using cloud computing are, but please keep it simple for a beginner."

Efficient Prompt ✅
"List the 3 main benefits of cloud computing, explained simply for a beginner. Use bullet points."

Control the Output Length: Stop the Over-Generation

The model uses the most energy when it's generating its response, known as "inference." The longer the output, the higher the energy cost.

- **The Problem**: Models often "hallucinate" or continue writing long after you've received the useful information.
- **The Fix**: Set Strict Limits. Use specific constraints in your prompt to prevent unnecessary token generation:
  - "Summarise in one paragraph."
  - "Provide a maximum of 5 bullet points."
  - "Limit your response to 100 words."

The Essential Action: If you get the answer early, hit the "Stop Generating" button immediately. You are saving power right then and there.

WANT TO KNOW ABOUT AI?

🖐 Your AI Footprint:
A 5-Step Guide to Energy-Efficient Prompting

Ask Hard Questions First: Check for Necessity

Generative AI uses significantly more computing power than traditional digital tools because it has to create a response from scratch, not just retrieve a pre-existing answer.

- **The Problem**: Many users default to an energy-intensive AI model for tasks that could be solved by a simple search, a spreadsheet, or an internal document.
- **The Fix**: Pause and Assess. Before hitting send, ask yourself: "Is this question or task absolutely necessary to put to a generative AI model?"
    - If you need a simple fact, a quick definition, or a link to an existing resource, use a highly optimised search engine or an internal resource first.
    - Save the energy-intensive AI for tasks that require true generation, synthesis, or creative drafting—the jobs only a complex model can do

Make It Repeatable: Train Your Workspace Assistant

The most efficient way to use AI for repetitive tasks is to teach it your task once, and then only give it the small pieces of information that change.

- **The Problem**: Re-explaining your company's tone, rules, and required format every day burns massive amounts of repeated energy.
- **The Fix**: Build or Use a Specialised Assistant in your workspace.
    - Use your available tools (like a "Gem" or "Agent") to pre-train a mini-AI just for your task (e.g., a "Team Update Writer Assistant").
    - This specialised Assistant already knows the context, tone, and format because you taught it once.
    - You only have to input a couple of short facts each time, saving the big AI from having to "re-learn" your entire job every single day. This is the ultimate power saving technique for business use!