KE HUB GREEN AI SANDPIT · OPENING TALK

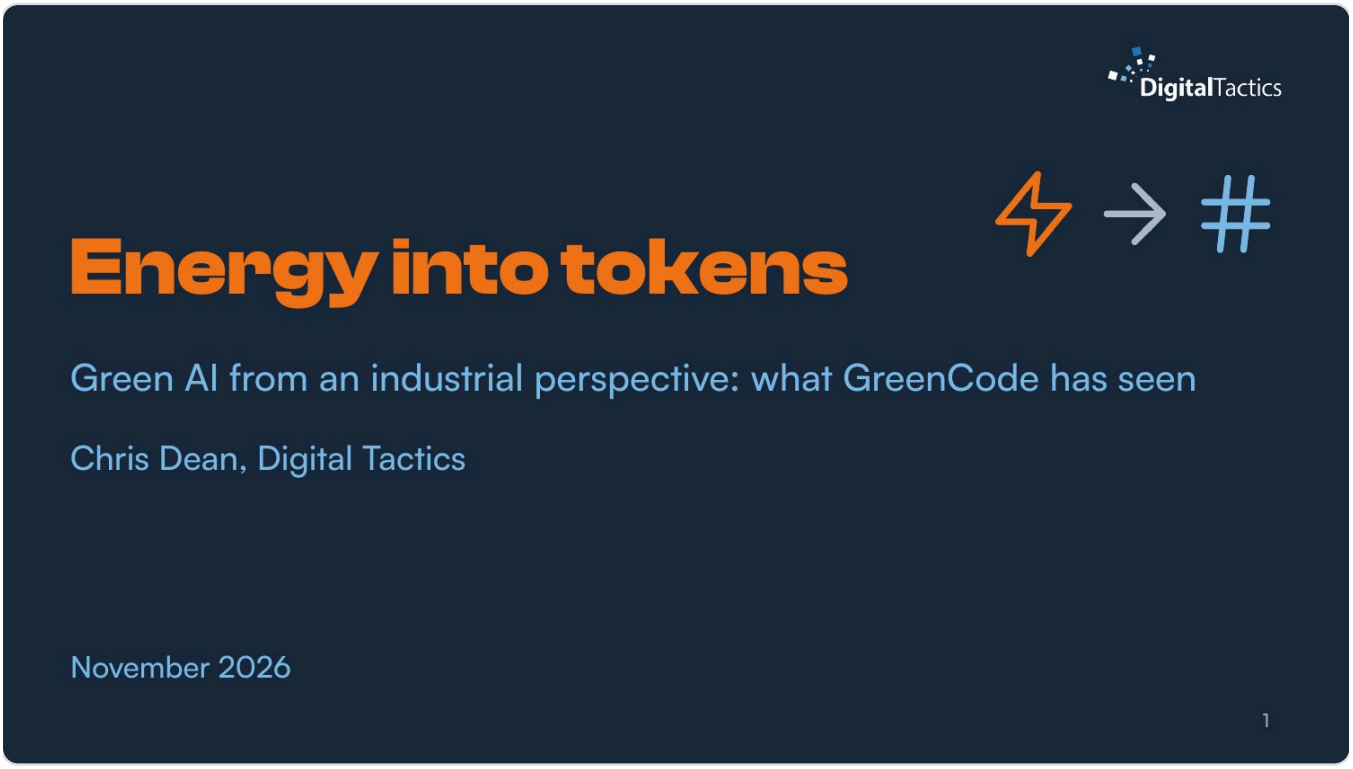
Energy into tokens

Green AI from an industrial perspective: what GreenCode has seen

This is the talk that opened the KE Hub Green AI Sandpit, Mathematical Sciences for Sustainable Artificial Intelligence, at STFC Rutherford Appleton Laboratory, Harwell, on 18–19 November 2026. Each slide is followed by the speaker's notes, lightly edited for reading. Sources are given at the foot of each slide.

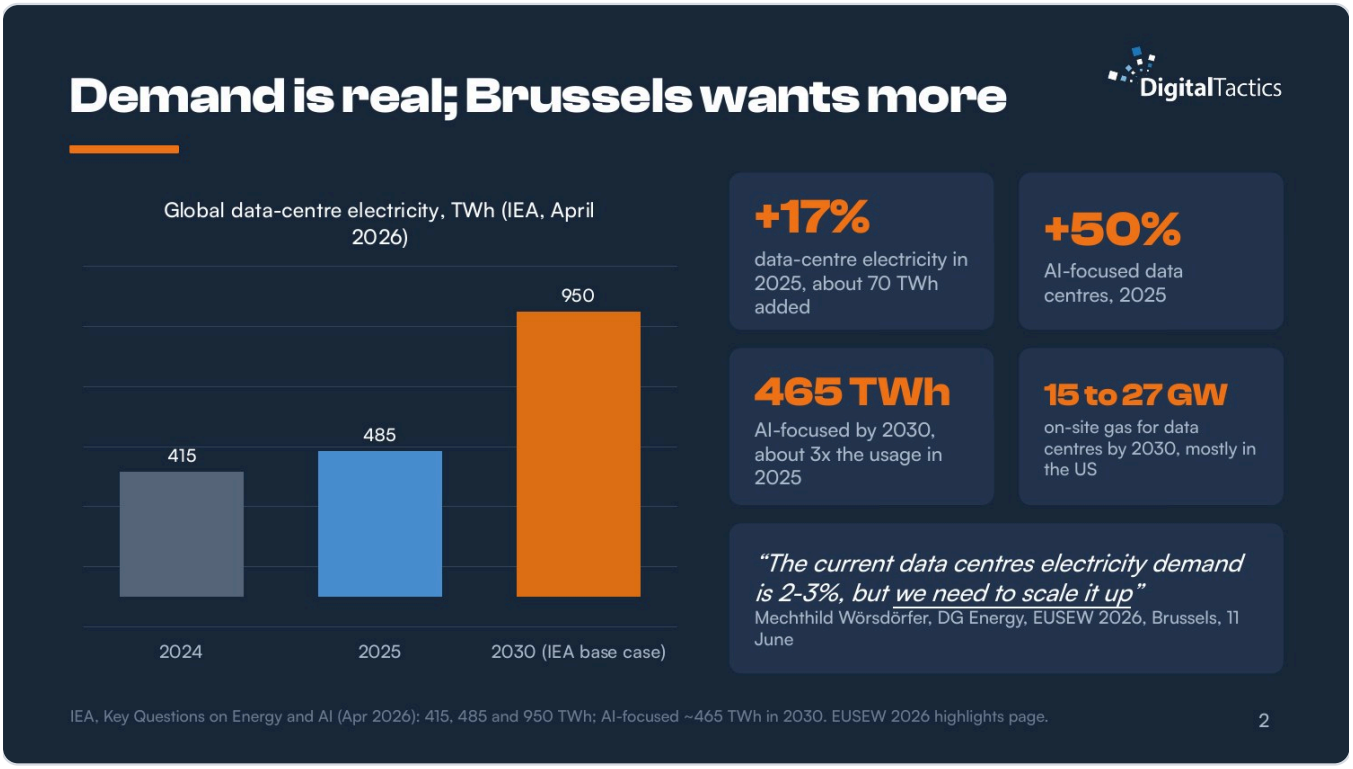
Chris Dean, Digital Tactics
GreenCode (ITEA 4 project 23016) · greencode.ai

01 Energy into tokens



Good morning. I'm Chris Dean. I run Digital Tactics, a software and AI consultancy in Hove, and I lead GreenCode, a European research project on energy-efficient software and AI, twenty-one partners in eight countries.

02 Demand is real; Brussels wants more



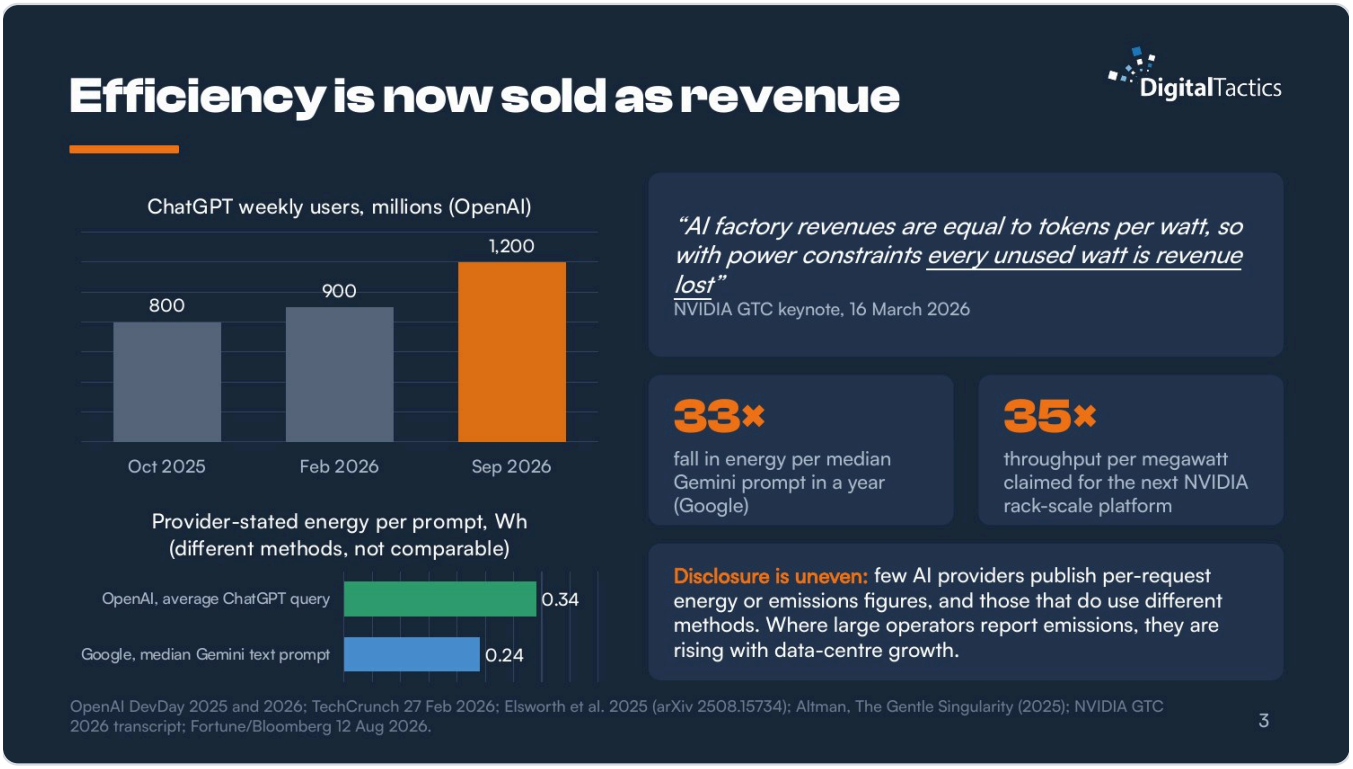
I'm going to give you the industrial view of green AI. Not what we'd like to be true, but what we've seen in two years of running a project on it.

Start with demand. The International Energy Agency's April update puts data centres at almost 500 terawatt hours in 2025, up seventeen per cent in a single year. The AI-focused part grew fifty per cent. That's roughly the electricity consumption of a mid-sized European country, added in twelve months. Their projection is a doubling by 2030, to around three per cent of world electricity.

Here's the part that surprised me from the Sustainable Energy Week sessions in June. I didn't make it to Brussels, I followed the streamed sessions. The Commission's energy directorate didn't talk about data centres as a problem to cut. The Deputy Director-General said demand is two to three per cent today "but we need to scale it up". Data centres as flexible load that helps the grid. That is the official European framing now.

And the way the extra demand is being met, at least in the United States, is gas. The IEA counts fifteen to twenty-seven gigawatts of on-site gas by 2030, and a seventy per cent jump in turbine orders last year. So the energy system is bending around AI, not the other way round.

03 Efficiency is now sold as revenue



Use is no longer a niche. OpenAI said in September that ChatGPT has 1.2 billion weekly users. It was 800 million a year before. One product.

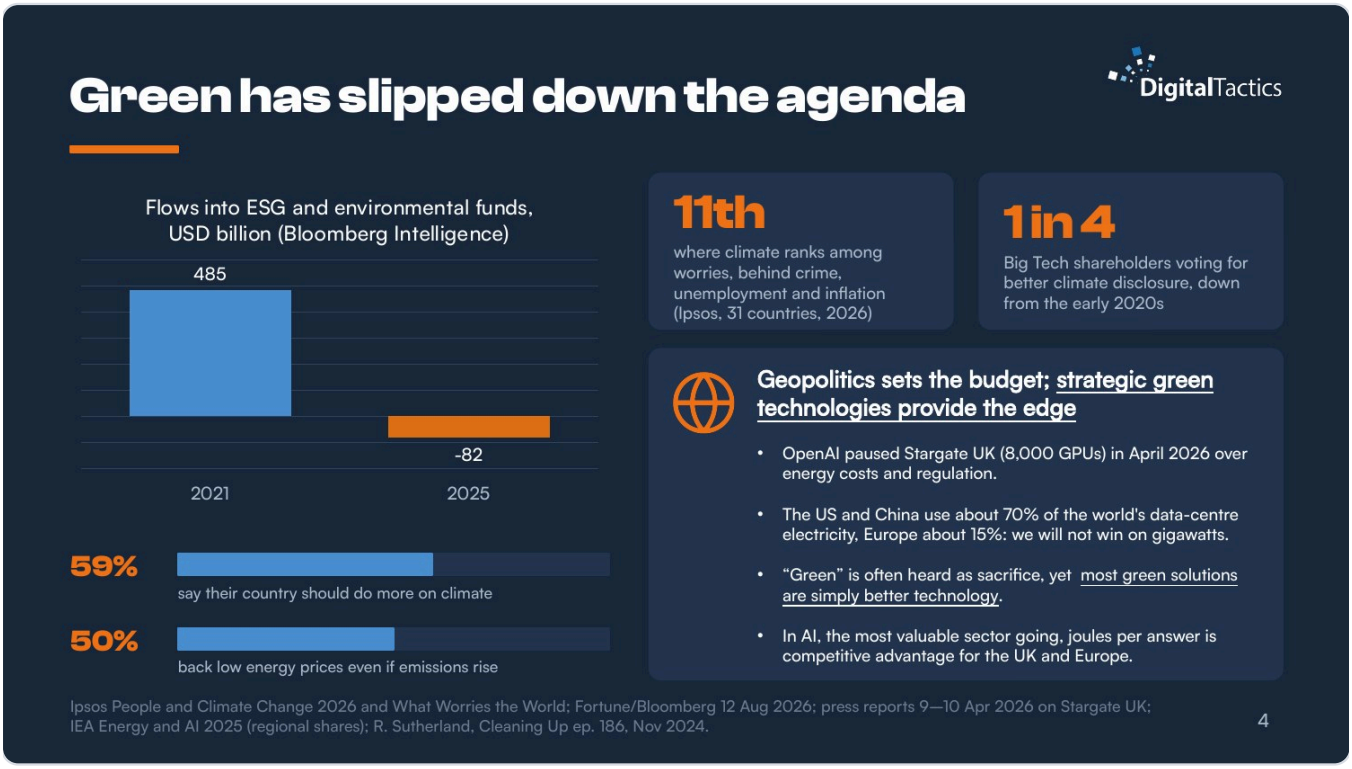
Now listen to how the industry talks about energy. At GTC in March, Jensen Huang said every AI factory is power constrained by definition, and NVIDIA's keynote spelled out what follows: AI factory revenue equals tokens per watt, so every unused watt is revenue lost. That is not a climate statement. It's a revenue statement. Efficiency is being sold as throughput per megawatt: thirty-five times more, NVIDIA claims, for its next rack-scale platform.

The model companies have started publishing per-prompt numbers. Google measured a median Gemini text prompt at a quarter of a watt-hour, and said it fell thirty-three-fold in a year, partly from software work. OpenAI's number is about a third of a watt-hour, from a blog post, no method published. Those are the provider's own figures, on different methods, so you can't compare them. And AI-focused data-centre electricity still went up fifty per cent last year, because use grew faster than efficiency.

The other half of the picture: disclosure is uneven. Few providers publish per-request figures, no two use the same method, and where the large operators do report emissions, they're rising.

So the industrial reading is: efficiency is real, it's measured in tokens per watt, and it's about revenue. The word green is nowhere in it.

04 Green has slipped down the agenda



I'll be straight about the mood. Green has slipped.

Ipsos surveyed thirty-one countries this year. Climate sits eleventh on the list of worries, behind crime, unemployment and inflation. Half of people would back low energy prices even if emissions rise. Fifty-nine per cent still say their country should do more, so it's not opposition. It's priority.

Money followed. ESG funds took in about four hundred and eighty-five billion dollars in 2021 and lost eighty-two billion last year. About a quarter of Big Tech shareholders now vote for better climate disclosure, down from the early twenties. One survey of sustainability professionals found nearly two thirds of companies had scaled back or rethought how they talk about it.

And the geopolitics landed on our doorstep. In April OpenAI paused its Stargate UK project, eight thousand GPUs, citing energy costs and regulatory uncertainty. The reporting put UK industrial power at about four times US prices. Energy security and trade: that's what sets budgets now.

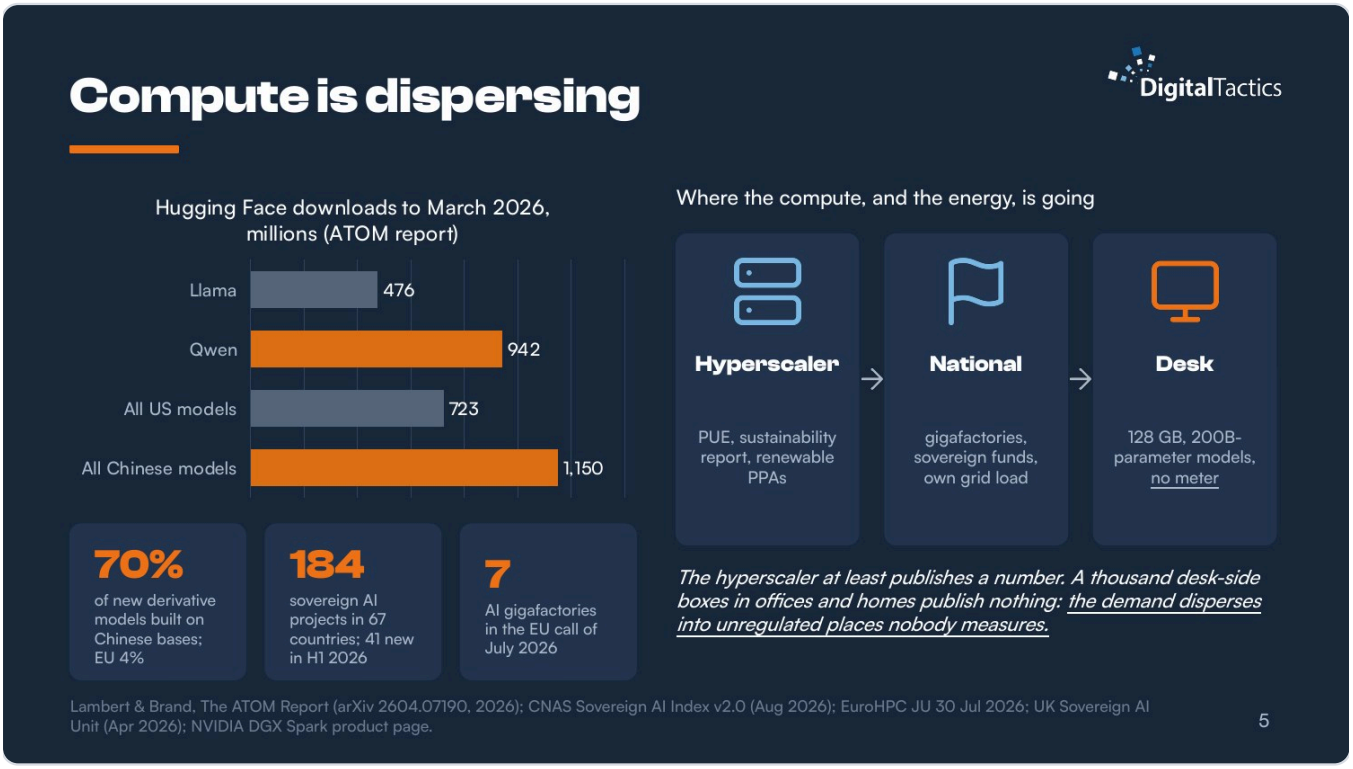
In GreenCode we learned this early, and I've said it to investors: the commercial interest is in cost, speed and quality. Sustainability is the icing on the cake. If you want green AI to happen, it has to ride on something the buyer already wants.

Part of the reason it slipped is tone. Green is often heard as sacrifice. Rory Sutherland makes the point that it is too often presented as self-sacrifice. Industry doesn't buy that framing, and it shouldn't have to, because most green solutions are objectively better technologies. An electric motor wastes less than an engine. An LED wastes less than a bulb. Efficient code is faster, cheaper to run and easier to maintain than bloated code. That is what GreenCode measures.

Now put that next to AI, the sector the markets value most right now. The United States and China will out-build us on compute. Between them they use about seventy per cent of the world's data-centre electricity; Europe about fifteen. The UK has just watched a data-centre project pause over power prices. We will not

win on gigawatts. We can win on joules per answer. Marry the green concern to the commercial one and it stops being a hair shirt and becomes a competitive advantage, for the UK and for Europe.

05 Compute is dispersing



Three movements at once, and they point the same way.

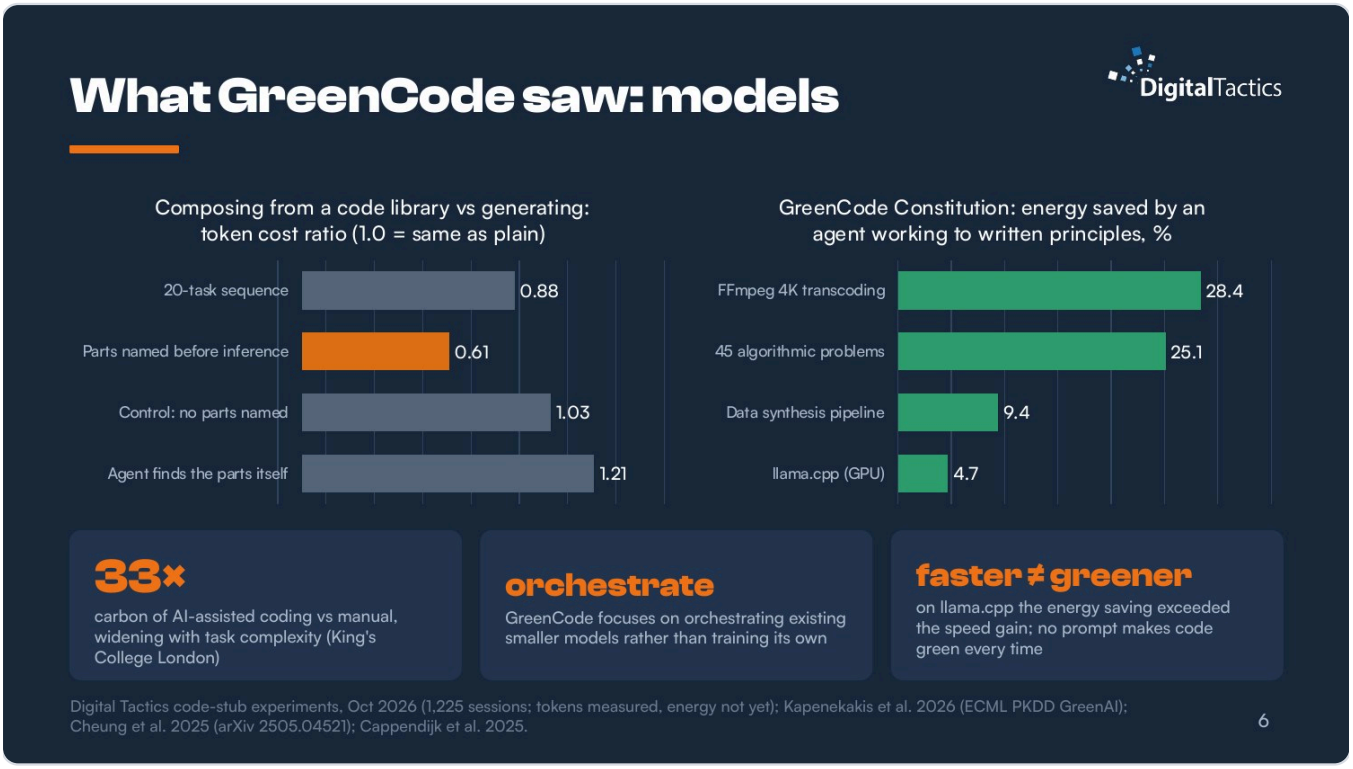
Sovereignty. The Center for a New American Security counts a hundred and eighty-four sovereign AI projects in sixty-seven countries, forty-one of them launched in the first half of this year. More in six months than in all of 2024. Most of them are building compute. Europe opened its call for up to seven AI gigafactories in July. The UK launched a five-hundred-million-pound Sovereign AI Unit in April, with a million GPU hours on Isambard-AI for the companies it backs. Every country wants its own capacity. Which means every country wants its own electricity for it.

Models. The ATOM report, from Nathan Lambert and Florian Brand, tracked fifteen hundred open models. Chinese models overtook American ones in downloads in July 2025 and the gap has widened: Qwen alone has nearly twice Llama's downloads. Seventy per cent of new derivative models are built on Chinese bases. Europe's share is four per cent. If you run a model locally in 2026, the odds are it's Chinese.

And personal. NVIDIA sells a box it calls a personal AI supercomputer: a hundred and twenty-eight gigabytes of memory, a petaflop on paper, two-hundred-billion-parameter models on your desk. AMD is pitching desk-side AI chips the same way. The pitch is privacy and control, and it's a fair pitch.

But think about where the energy goes. A hyperscaler at least has a PUE and a sustainability report. A thousand desk-side boxes in offices and homes have neither. The demand disperses into places nobody measures.

06 What GreenCode saw: models



Now what we've actually seen. GreenCode is twenty-one partners, and on green AI our question changed twice.

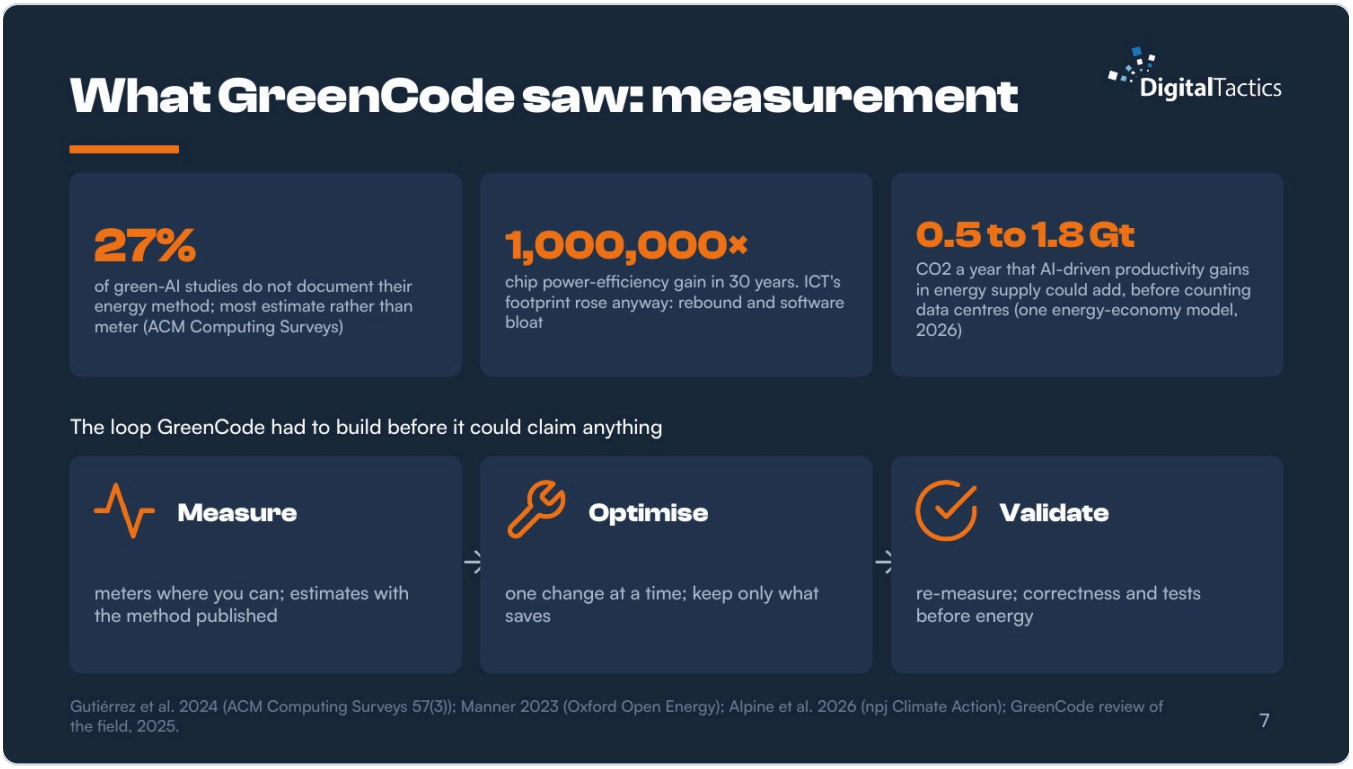
We started out wanting to build small, efficient models. But by the time a research project builds a careful model, someone with a far bigger budget has released a better one. So our focus is on orchestrating existing smaller models rather than training our own.

Second: generating is not automatically efficient. A study at King's College London measured AI-assisted coding at about thirty-three times the carbon of writing the code by hand, and the gap widens with task complexity. And prompting a model to write green code only works for some prompts, some models, some tasks. There is no magic prompt.

Third, our own experiment. We gave an AI coding agent a library of pre-written, energy-checked code to compose from instead of generating. When the agent had to find the parts itself, it cost twenty per cent more tokens, not less. When the parts were named before the model started, it cost thirty-nine per cent less. And over a long sequence of tasks the saving faded, because the model reuses its own code. Twelve hundred sessions, token counts measured, energy not yet. The lesson: the saving is in the selection, done outside the model's loop.

Fourth, and this one's encouraging. Our Aalborg partners gave an agent a written constitution: correctness first, then thirty-seven energy anti-patterns. It trimmed the GPU energy of llama.cpp, an already tuned inference engine, by nearly five per cent, and saved far more on video transcoding. The energy saving exceeded the speed gain. Faster is not the same as greener.

07 What GreenCode saw: measurement

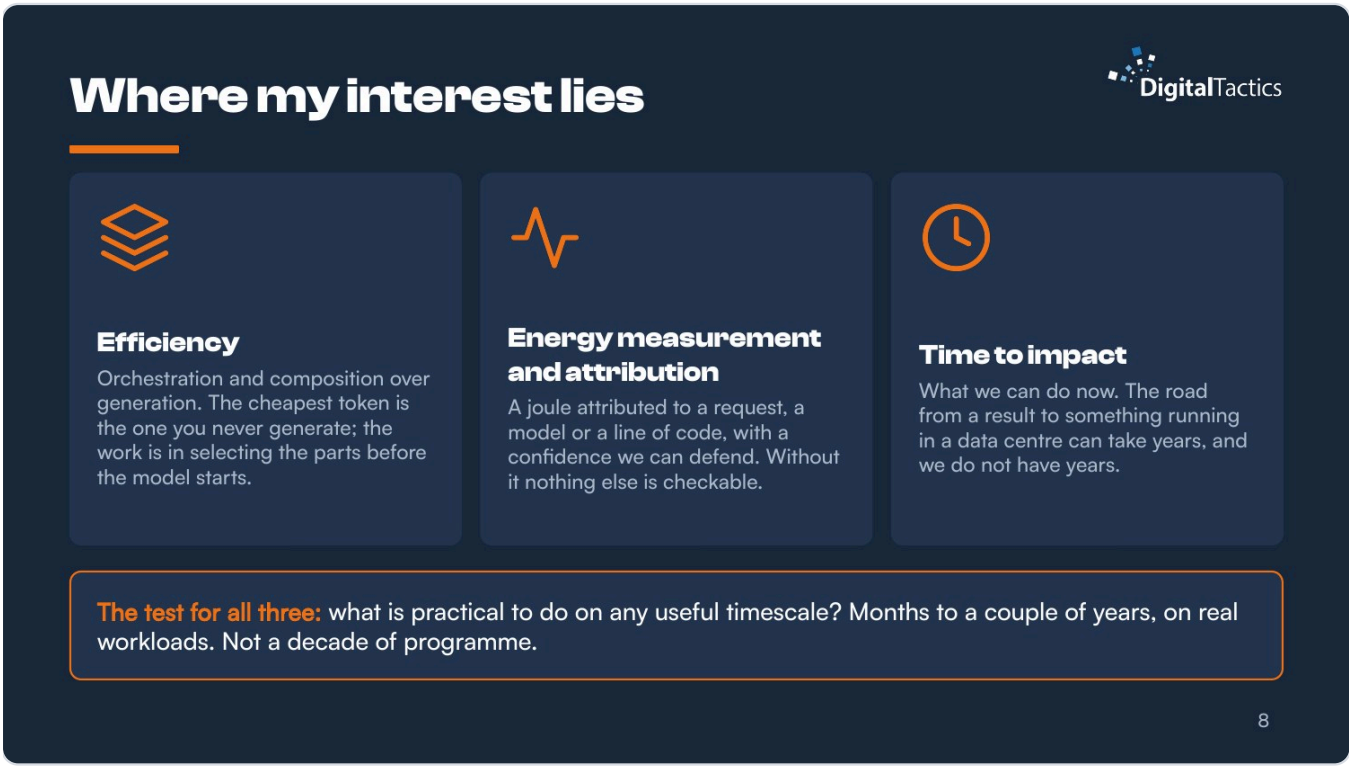


Two more things, and they're the uncomfortable ones.

Measurement is the bottleneck. A survey in ACM Computing Surveys found most green-AI studies use software estimates rather than meters, and more than a quarter don't document their method at all. In the public cloud the power interfaces are hidden, so attribution in a shared system is a statistical model, not a meter. We've spent a good part of GreenCode just on getting a trustworthy number. Without one, every green claim in this industry is a press release.

And rebound. Chip power efficiency has improved roughly a million-fold in thirty years. ICT's footprint went up anyway, because cheaper compute gets used more. One energy-economy model published this year finds that AI making energy supply more productive, fossil as well as renewable, could add half a gigatonne to nearly two gigatonnes of CO2 a year, and that is before counting the data centres. Efficiency is necessary. It's not sufficient.

08 Where my interest lies



Before the questions, let me say where my own interest lies, because it shapes what I'd ask of this room.

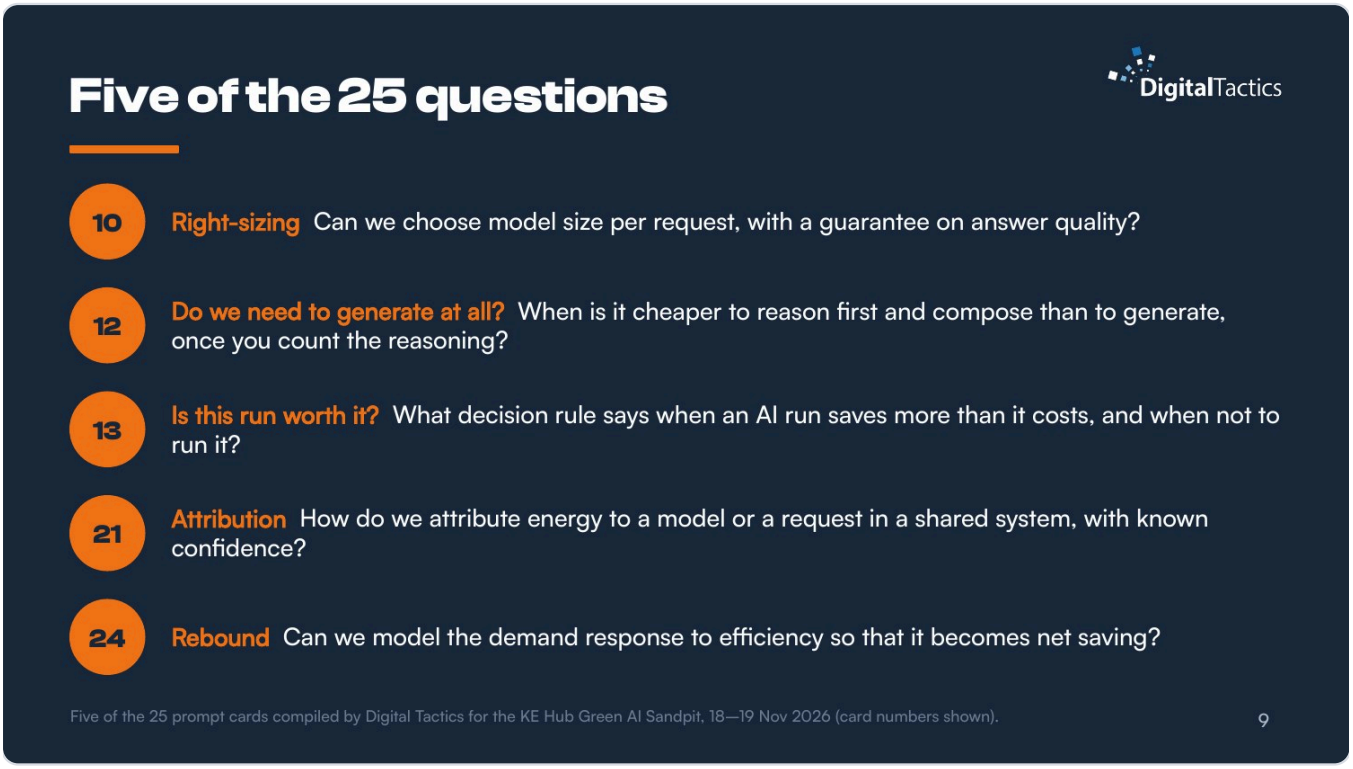
Efficiency, but specifically orchestration and composition over generation. The cheapest token is the one you never generate, and the work is in selecting the parts before the model starts.

Energy measurement and attribution. If we cannot attribute a joule to a request, a model or a line of code, with a confidence we can defend, nothing else we say is checkable.

And time to impact: what we can do now. The road from a result to something running in a data centre can take years, and I don't think we have years.

Which gives me one test for everything that follows: what is practical to do on any useful timescale? Months to a couple of years, on real workloads. Not a decade of programme.

09 Five of the 25 questions



We wrote twenty-five of these questions for this sandpit, where mathematicians and industry sit down together for the next two days. Here are five.

Choosing the smallest adequate model per request, with a guarantee. Knowing when composing beats generating, counting the cost of the reasoning. A decision rule for whether a run is worth it at all. Attributing energy in a shared system with a confidence interval. And modelling the rebound so a saving is a saving.

None of these is a hardware problem. They're decision theory, statistics, information theory and economics. That's why we went to the mathematicians. The full set is available; ask me afterwards.

10 An industrial agenda



So, an industrial agenda. Five things, none of them exotic.

How can mathematics support industrial needs that lead to rapid adoption of Green AI? Some examples.

Measure inference and show it to the person who caused it. People change behaviour when they see the cost. Where you can't meter, estimate, and publish the method.

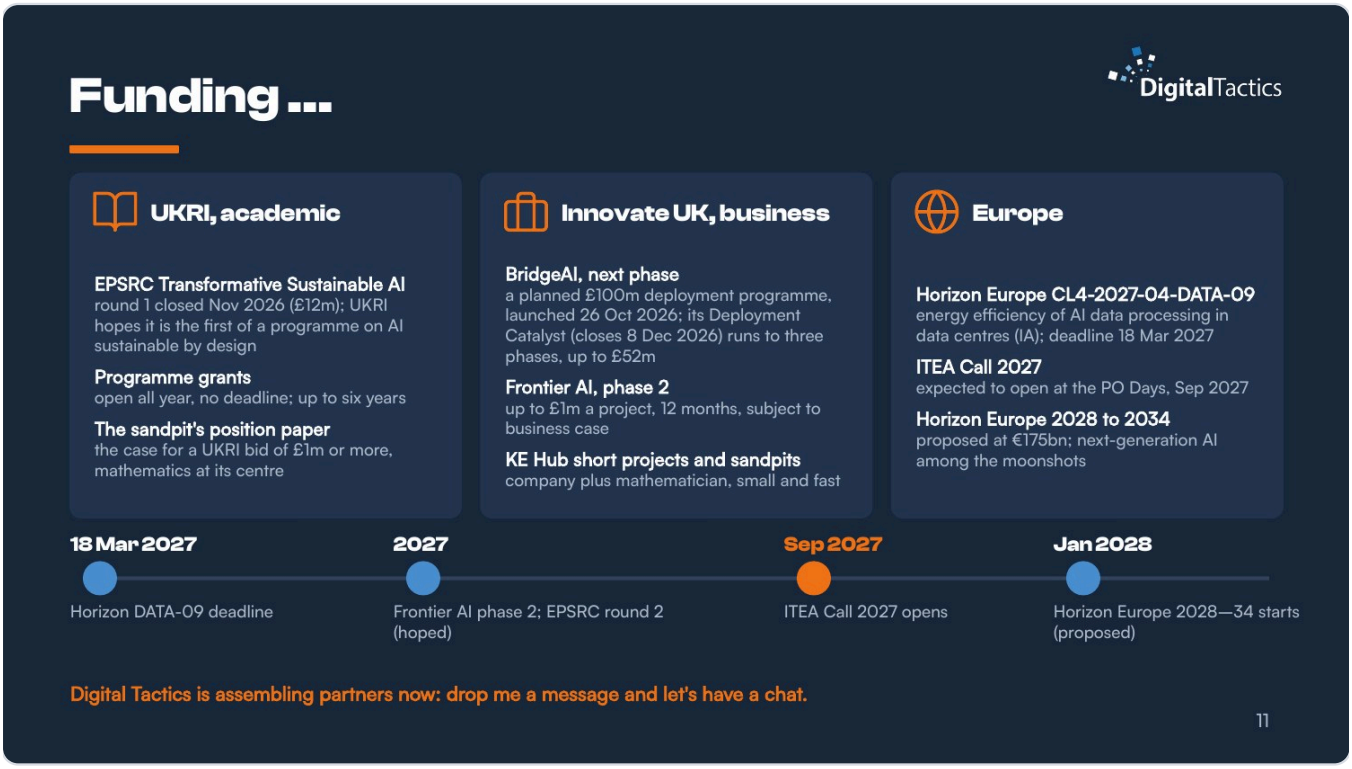
Pick the smallest adequate model, and pick it before the model starts. Generate only what's new.

Schedule the work that can wait, training, batch jobs, evaluation, against the grid forecast. Britain publishes carbon intensity ninety-six hours ahead.

Count the hardware. A desk-side AI box carries its manufacturing emissions whatever you run on it.

And last, the one I'd underline. Make the green result ride on something else. Cost, speed, sovereignty. In 2026 it will not carry a budget on its own.

11 Funding



Which brings me to the ask, and I'll look past this year's deadlines to what opens next.

On the academic side, UKRI has already shown its hand. EPSRC ran a twelve-million-pound call called Transformative Sustainable AI Technologies this autumn, for energy-efficient algorithms, smaller task-specific models and hardware and software co-design. It closed in November, and UKRI says it hopes it is the first in a broader programme on AI that is sustainable by design. That's the signal. The programme-grant route stays open all year with no deadline, and the whole point of these two days is a position paper that makes the case for a UKRI investment of a million pounds or more with the mathematics at its centre.

On the business side, Innovate UK has started the next phase of BridgeAI: a planned hundred-million-pound programme for real-world deployment of advanced AI in the Industrial Strategy's priority sectors. Business Connect ran shaping round tables with industry in July, and it launched on the twenty-sixth of October. The first piece is the Deployment Catalyst, open until the eighth of December: three phases, up to fifty-two million pounds across them, pairing an AI SME with a deployer. The Frontier AI competitions carry a phase two, up to a million pounds a project, if the business case is approved.

In Europe, there is a Horizon topic that could have been written for this room: energy efficiency and sustainability of AI data processing in data centres, an innovation action, deadline the eighteenth of March. Real facilities, not papers. ITEA, the programme GreenCode came out of, opens its next call at the PO Days in September 2027, and I'd recommend it to anyone: businesses, universities and research organisations across Europe and beyond, brought together on real industrial problems. And beyond that, the next framework programme from 2028, proposed at a hundred and seventy-five billion euros, with next-generation AI among its moonshots.

We're assembling partners for all of these now. If you run AI at scale, build the software underneath it, or do the mathematics, drop me a message and let's have a chat.

Thank you.