# Written evidence submitted by Dr Harri Renney (LEC0005)

Dear Dame Chi,

I welcome the opportunity to submit evidence to this inquiry. I am Dr Harri Renney, Head of Research at Kaze Consulting and Associate Lecturer at the University of the West of England. My work spans applied research in AI, multi-agent systems, data science, and cyber security, with particular experience in defence and private sector applications. I am lead author of the paper *Cloud to Edge: Benchmarking LLM Inference on Hardware-Accelerated Single-Board Computers* (April 2026), conducted in collaboration with researchers at the University of Exeter, the Lebanese American University, and Carnegie Mellon University. The paper measures the energy cost of running AI on small, **low-power devices** using **hardware accelerators**. The findings relate to this committee's questions.

## Key points

1. Low-energy AI computing is not just a future possibility. It is happening now. Our benchmarking shows that small AI models running on affordable edge devices with hardware accelerators use as little as 1.4 watts, compared with 300–1,200 watts for a single data-centre GPU, reducing energy draw by at least 99%.
2. Adding a neural processing unit (**NPU**) to a small single-board computer can improve energy efficiency by up to 40 times compared with running the same AI task on the CPU alone.
3. These devices cost under £350, are smaller than a paperback book, and can run useful AI tasks without any internet connection, making them suitable for defence, critical infrastructure, satellite operations, and other sensitive environments where data cannot leave a controlled network.
4. The UK has no sovereign capability to manufacture NPUs. The hardware we tested comes from Israel, Taiwan, China, and the United States. This is a supply chain vulnerability the committee should consider alongside the energy question.

## Question 1: Can neuromorphic computing and silicon photonics provide a solution to the sustainability challenge of increasing AI-related energy demands?

5. Yes, but significant energy savings do not depend solely on future breakthroughs. Our research shows substantial gains are available now through existing NPU-based edge accelerators combined with model compression techniques such as quantisation.
6. On a Raspberry Pi 5, adding a Hailo NPU dropped energy consumption from 27–77 MJ/Mtok[1] on CPU to 0.88–5.5 MJ/Mtok with the NPU, improving energy consumption efficiency by up to 40×.
7. Edge devices in our study operate at 1.4 to 13 watts. Centralised GPUs can require 300–1,200 watts each. It is a fundamentally different operating model in which AI runs locally rather than in energy-intensive data centres.
8. Low-energy computing should be understood as a spectrum. Neuromorphic and photonic approaches represent the longer-term frontier. NPU-accelerated edge AI is deployable today and offers immediate relief. The two are complementary: NPU architectures already draw on neuromorphic design principles, and photonic interconnects may enhance edge performance further in future.

## Question 2: What other opportunities does this create, and how well placed is the UK to take advantage?

9. Through my work in defence and cyber security, I see particular value in four areas:
   - 9.1. **Resilient critical infrastructure**, where distributed edge nodes reduce single points of failure.
   - 9.2. **Operational contexts** where mission success requires federated compute independence and reducing dependence on connectivity.
   - 9.3. **Energy-constrained autonomous systems** such as drones, where onboard power budgets are extremely limited.
   - 9.4. **Distributed satellite ground stations**, where we demonstrated useful AI inference on Raspberry Pi platforms.

[1] megajoules per million tokens

10. Edge AI also addresses data sovereignty. In defence and operational technology, it is imperative for data to remain within controlled networks. Running AI locally removes the need for cloud connectivity, reducing both energy use and security exposure.
11. The Digital Catapult / Innovate UK report *The UK at the AI Frontier* (April 2026) independently identifies edge AI and small language models as a key opportunity across all five sectors it examined. It also flags barriers: limited compute access, short-term funding cycles, NVIDIA CUDA lock-in, and a disconnect between academia and industry.

### Question 3: Where does the UK sit in terms of scientific research, and who are our main competitors?

12. The UK has genuine research strengths in neuromorphic computing, silicon photonics, and AI hardware co-design, with notable capabilities in circuit design, analogue engineering, and photonics. Our own research, conducted across UK and international institutions, contributes to the emerging field of edge AI benchmarking.
13. In edge AI hardware, however, leading NPU manufacturers are all based abroad: Hailo (Israel, fabricated in Taiwan), Axera (China), and NVIDIA (United States). NVIDIA's dominance is reinforced by its CUDA software ecosystem, which creates significant lock-in. Hence, the UK's main competitors in this space are the United States, China, Israel, and Taiwan.

### Question 4: To what extent has the UK developed sovereign capabilities in this area?

14. Every hardware platform in our study was designed and manufactured outside the UK:

    14.1 The Hailo NPU is Israeli-designed, TSMC-fabricated.

    14.2 The AX630C is Chinese.

    14.3 The NVIDIA Jetson is American.

    14.4 Even the Raspberry Pi uses a Broadcom chip fabricated abroad (but assembled in **Wales**).
15. The Sovereign AI Unit (£500 million) and ARIA (£800 million) are welcome initiatives, but focus primarily on AI models and software rather than the hardware supply chain and AI adoption methodologies/support. A UK strategy relying entirely on imported edge AI hardware is exposed to geopolitical disruption and export controls.
16. The UK's strengths in photonics and compound semiconductors could provide a route to domestically designed low-energy AI accelerators. Supporting co-design of AI models with novel UK hardware would play to existing national strengths.

### Question 5: What objectives should future government policy interventions seek to deliver?

17. Fund research bridging benchmarking and real-world deployment. Our work introduces practical metrics for **operational hardware selection**. A national programme extending this across more hardware and domains would have significant value.
18. Invest in **UK-based NPU** or **low-energy accelerator design capability**: not fabrication, but the design talent and IP to shape next-generation edge AI hardware.
19. Support **open-source**, **hardware-agnostic** AI software frameworks to reduce dependence on NVIDIA CUDA and allow UK-designed hardware to compete.
20. Align funding timescales with hardware realities. A single semiconductor design cycle takes at least twelve months, and meaningful research requires multiple iterations. **Short-term funding windows do not support this.**

### Further information

21. The full publication is available at: https://arxiv.org/pdf/2604.24785. All supporting data at: https://osf.io/5r9t4
22. I would welcome the opportunity to give oral evidence to the committee should that be helpful.

Head of Research, Kaze Consulting
Associate Lecturer, University of the West of England
*30 April 2026*