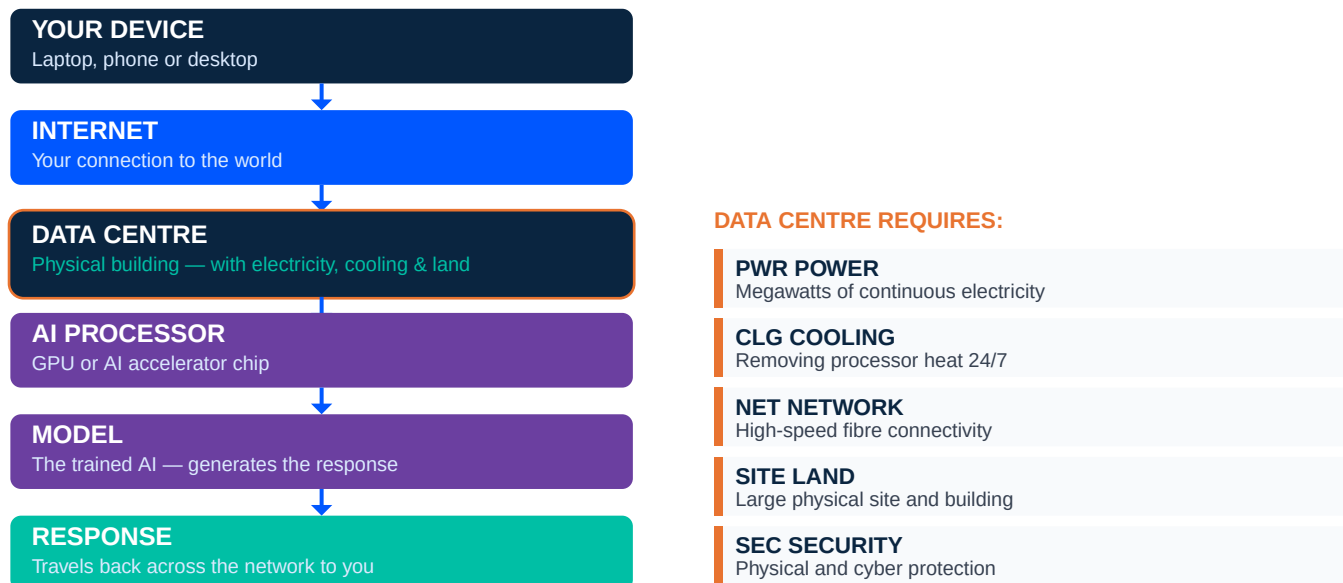


YOU TYPE A PROMPT. WHAT HAPPENS NEXT?

Where Does AI Actually Run?

The physical infrastructure behind every AI response

AI may appear inside a browser or app, but the computing normally happens in large physical data centres — not in 'the cloud'. Every prompt you submit follows a physical chain before a response arrives.



Training vs Inference

TRAINING

Building the AI model. Very large compute for weeks or months. Happens infrequently.

INFERENCE

Running the model for each user request. Continuous at scale — millions of queries daily.

AI FEELS DIGITAL. ITS INFRASTRUCTURE IS PHYSICAL.

Every AI answer has a physical infrastructure cost somewhere, even when the user never sees it.

What This Means for Your Business

AI and cloud pricing is influenced by infrastructure economics — electricity, hardware, land.

Data residency — where your data physically sits — may matter for compliance. Check which region your AI provider uses.

Supplier concentration: most AI runs on a small number of hyperscalers. Consider your dependency.

Infrastructure constraints affect service availability and resilience.

Questions Worth Asking Your AI Provider

- ☐ Which cloud region or data centre handles my requests?
- ☐ What is your commitment on service availability and resilience?
- ☐ Where does my data sit? Can I restrict it to a specific geography?
- ☐ What are your sustainability commitments for your data centre infrastructure?