WOODHURST INSIGHTS · AUGUST 2026

You Can Have OpenAI or Claude and Keep Your Data in the UK

How UK financial services firms can deploy frontier AI models without customer data leaving the country, and what it really costs.

Woodhurst

The best models, the wrong postcode

Almost every lender, building society and bank we work with has landed in the same place. The models worth building on are the frontier ones, GPT from OpenAI and Claude from Anthropic. Shortly after reaching that conclusion, a significant compliance risk is realised: using frontier models means company data is sent outside the UK/EEA.

Both companies are US based. By default, their standard APIs process data on US infrastructure. OpenAI now offers data residency in a list of regions that includes the UK, and approved API customers can opt into European processing for some endpoints, but the UK option covers data stored at rest only; neither lab offers UK in-region processing. Strictly speaking, that isn't fatal: UK GDPR does permit transfers to the US with the right safeguards in place. But anyone who has taken a Data Protection Impact Assessment through a risk committee knows how that conversation tends to go. Between outsourcing rules, operational resilience expectations and the board's appetite for headlines, most firms end up with the same requirement. Customer data stays in the UK, or at a push, in Europe.

So the business case sits in a drawer. The capability is real, the risk position is reasonable, and nothing moves.

It doesn't need to be this way. There is a well-trodden route through, and it runs via Microsoft and Amazon rather than via OpenAI and Anthropic directly.

> **The models are already inside your existing cloud infrastructure, in regions your regulators and risk committees can live with.**

Why bother

It's worth checking, first, that the prize justifies the effort. If frontier models were just better chatbots, we would tell clients not to lose much sleep over residency.

They aren't. We see the value showing up in three places.

01

Language work

The familiar one: drafting complaint responses, summarising policy documents, producing management information on demand. Useful, well understood, quick to pilot.

02

Agents

An agent doesn't just answer, it acts. It retrieves the documents, checks the data against the rules, works through a multi-step process and hands the outcome to a human for approval. For lenders, this is where the operational value sits: bereavement casework, payments investigations, financial crime alert triage, drafting audit narratives.

03

AI-assisted capability delivery

The one that gets underestimated more than either of the others. Tools such as Anthropic's Claude Code and OpenAI's Codex let development teams hand over whole coding tasks, and the same generation of tools lets operations and risk teams build their own workflows, automations and reporting without waiting for a change queue. For a firm carrying twenty years of technical debt on its core systems, this may prove the most commercially significant use of the entire technology. It is also the one where your source code, system context and process knowledge flow through the model, so the residency question gets sharper here, not softer.

All three touch exactly the kind of data your DPIA cares about. Hence the compliance risk.

Buy the model from your cloud, not its maker

You don't have to buy these models from the companies that built them. That single fact changes the conversation more than any technical detail.

Microsoft hosts OpenAI's models inside its own Azure data centres through Microsoft Foundry (formerly Azure AI Foundry), and has added Claude to the same catalogue. Amazon does the equivalent for Claude through Amazon Bedrock, alongside a broad catalogue of others. In both cases the model runs on the cloud provider's infrastructure, under the cloud provider's contract, in regions you choose. Your prompts never reach OpenAI or Anthropic. Nothing is used to train their models. The whole arrangement sits inside the same Azure or AWS estate your firm almost certainly already runs on, under the same enterprise agreement, security controls and audit arrangements. Google runs the same pattern through Vertex AI, though UK in-region processing covers a narrower slice of its catalogue today; we focus on Microsoft and AWS because they are the estates most UK financial institutions already run.

That last point does more work than the technology. Microsoft and AWS are already approved suppliers at most UK financial institutions, already risk-assessed, already on the outsourcing register. Consuming a frontier model through them extends an existing supplier relationship. It is not a new third-country data transfer to a US AI lab, and your risk function will notice the difference. None of this is exotic engineering, and none of it is a concept waiting for a first mover. Institutions across the market are already running this arrangement in production, from tier-one high street banks to regional building societies. Adopting it is following a proven path, not running an experiment, and most firms can stand it up in weeks.

Where the data actually goes

Both platforms let you control where processing happens. The mechanics differ, and the detail matters, because this is precisely where most write-ups go vague.

Microsoft Foundry

OpenAI models · plus Claude

Global
Routed anywhere in Microsoft's estate. Cheapest, but no residency promise at all.

Data Zone
Kept within a defined geography, either the EU or the US. The EU zone excludes the UK, a quiet side-effect of Brexit.

Regional UK ✓
Pinned to a single Azure region such as UK South. The only route to strict UK residency, now with a catch on model choice.

Watch out: pay-per-use stops at the deprecated GPT-4o and GPT-4.1-mini; the GPT-5 family needs Provisioned Throughput and tops out at GPT-5.4; Claude here has no UK or EU residency yet (US data zone only, Europe slated for 2026).

Amazon Bedrock

Claude models · plus a broad catalogue

London region UK ✓
Claude Sonnet 4.6 served directly in eu-west-2 on pay-per-use, with inference pinned to London alone.

EU inference profile
Routes across European regions while data stays within the EU geography, at roughly 10% over global.

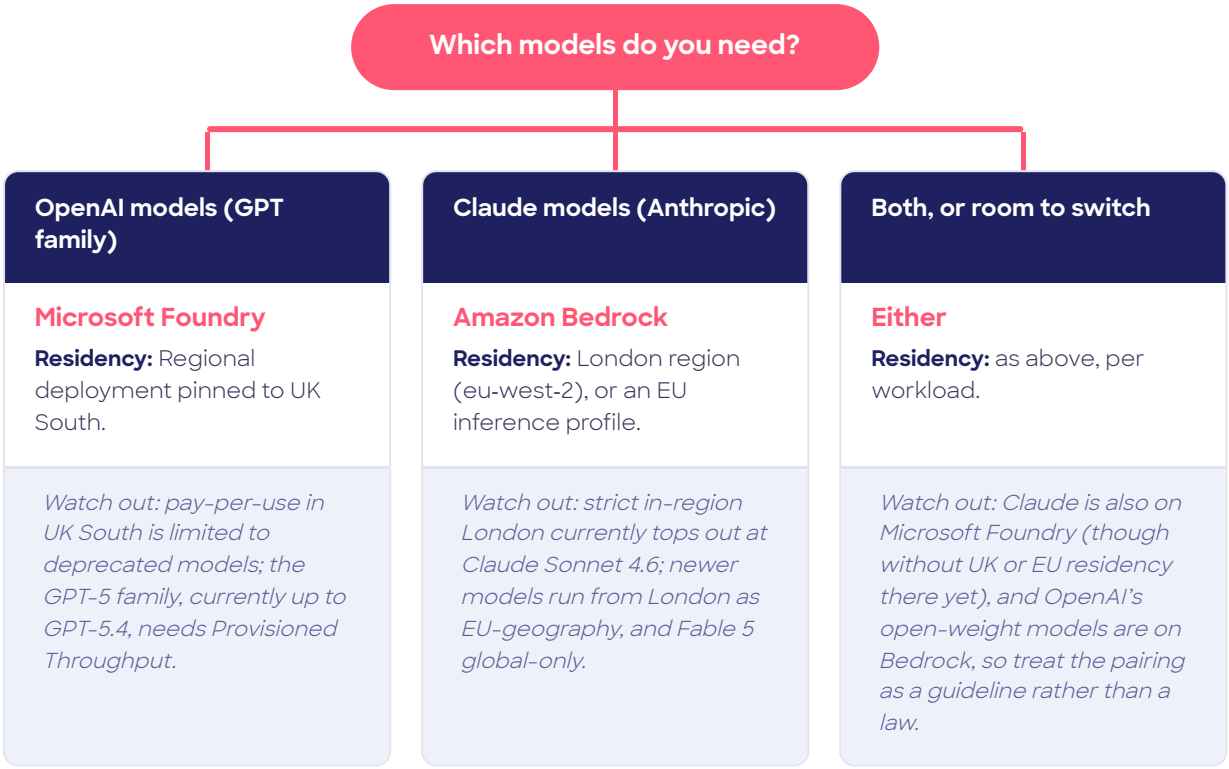
Model catalogue
Claude sits alongside many others, including OpenAI's open-weight models.

Watch out: Sonnet 5, Opus 4.8 and the new Opus 5 run from London as EU-geography only, and Fable 5 is global-only from Europe.

Today, the London story for Claude is simply cleaner than the UK South story for OpenAI models: a current Claude model on pay-per-use in-region, against deprecated models or reserved-capacity commitments in UK South. That gap is worth re-checking at the point of build, because availability shifts monthly.

Neither route asks you to give up the enterprise fundamentals. Private networking, your own encryption keys, no training on your data and full audit logging come as standard on both.

Which cloud provider for which model



Your cloud strategy and your model strategy have become the same conversation. A Microsoft-first estate points naturally at Azure and the GPT family; a firm comfortable on AWS gets first-class access to Claude in London. Run both and you have genuine optionality, plus a stronger negotiating position than firms locked to either.

What it costs

The useful reframe here: this is not a subscription decision. Copilot charges a fixed fee per user per month whether anyone uses it or not. Models consumed through Foundry or Bedrock are pay-per-use. You are charged per token processed, roughly per word in and per word out. A drafted complaint response costs a few pence; a heavyweight document-analysis agent run might cost tens of pence. Production spend tracks the value being generated rather than the size of the licence.

Headline per-token rates through the hyperscalers broadly match what OpenAI and Anthropic charge directly. Residency now splits the picture into three tiers. Routing globally is cheapest; staying inside the EU boundary carries a modest premium, roughly 10% over global on both platforms, but that boundary excludes the UK; strict UK-only processing carries that same premium and costs you model recency, plus on Foundry a reserved-capacity commitment to reach current models. Pilots stay cheap everywhere except UK-pinned Foundry workloads, where that commitment changes the entry cost.

TIER ONE	TIER TWO	TIER THREE
Global routing	**EU boundary**	**UK only** UK ✓
Cheapest per token. No promise about where your data is processed.	Roughly 10% over global: Bedrock EU profile and Foundry EU Data Zone. Excludes the UK.	Costs model recency and the same roughly 10% premium: current Claude on Bedrock London; deprecated models or reserved capacity on Foundry.

Two practical consequences follow. Because you pay per use, the governance question stops being “can we afford the subscription?” and becomes “which processes justify their token spend?”, which is a far healthier discipline. And reserved capacity now cuts both ways: it lowers unit costs at genuine scale, and on Foundry it is currently the only strict-UK route to the newest OpenAI models. Where you have the choice, start pay-as-you-go, measure, then commit. That opens a wider conversation about token optimisation, where clear best practices are emerging; we will explore them in a follow-up article.

The bottom line

Data residency is the objection we hear most often when frontier AI programmes stall in UK financial services. It is also the one with the clearest fix. The models are already inside the clouds you already trust, in regions your regulators and risk committees can live with. The trade-off is only marginally about price; it is mostly about how new a model you can run inside your chosen boundary, and that picture shifts monthly.

The firms making progress never found a way around the residency requirement. They noticed it had already been solved, and spent their energy on the harder questions instead: which processes, what governance, and how fast.

WRITTEN BY

Mokhammed Sagaev

Charlie Patterson

woodhurst.com

Woodhurst