Below the surface

A deep dive into full-trust AI for SMEs.

IN THIS 14-PAGE GUIDE

Real AI adoption needs sensitive data, and commercial LLMs can't be fully trusted with it.

The path to a full-trust AI stack is easier and more affordable than most companies think.

We compare 6 models, explain 5 layers in the stack, and outline 4 steps to get you started.

beside

INTRODUCTION

Business AI doesn't go deep enough. It should.

Most SMEs have adopted AI at a superficial level. Drafting and summarising are useful, but the real value lies in the sensitive records AI is not allowed to touch. For a good reason: commercial AI was not built for trust.

This paper sets out what full-trust AI requires, which models deserve attention, what the overall stack looks like and how to approach it. To explore what this means for your organisation, book a free consultation on our website.

Giles and Theo
Your Beside Partners



1 Introduction
2 What we observed
3 Trust is everything
4 Why AI remains shallow
5 Limits of commercial AI
6 What gets missed
7 How we approach it
8 Benefits of an owned stack
9 Beyond the frontier
10 Stack architecture
11 Cost of ownership
12 Next steps

WHAT WE OBSERVED

Most AI projects stop where the data gets interesting.



Customers and regulators are asking harder questions.

Generated text has made competence cheap to imitate. A polished proposal takes a minute to produce, so polish no longer tells a buyer anything. The questions have moved to how a firm knows what it claims: who checked the work, what it drew on, and where the data went.

Regulation is moving the same way, and not on the timetable most assume. Transparency and literacy duties already apply; only the high-risk rules were deferred. Forward-thinking firms could start treating accreditation as a commercial asset rather than a compliance cost.

Are you ready for the EU AI Act?

Obligation	Applies from
Prohibited practices	Feb 2025
AI literacy duties	Feb 2025
General-purpose model rules	Aug 2025
Transparency, Article 50	Aug 2026
Marking synthetic content	Dec 2026
High-risk uses, Annex III	Dec 2027
High-risk, inside products	Aug 2028

Regulation (EU) 2026/1744 (Digital Omnibus on AI), amending Regulation (EU) 2024/1689. In force 27 July 2026.

Policies exclude the data that makes AI valuable.

Most AI in an SME runs on the green zone alone: material that was already public, where nothing is at risk and little is at stake. The amber zone is rarely ruled out explicitly. It is simply left undefined, so people avoid it. The red tier is excluded outright.

Value climbs as the lights change. Almost nothing in the green zone moves a number; nearly everything that would sits in amber and red. None of it is blocked by the technology. All of it is blocked by a classification made, usually quickly, when the policy was written.



Public data

- Product brochures
- Public research
- News and blog posts
- Team bios
- Help content



Internal data

- Meeting notes
- Support triage
- Compliance
- Complaints
- Project updates



Restricted data

- Client data
- Contracts
- Financial data
- IP documentation
- HR files

Commercial AI is built for exploring, not for trusting.

Commercial AI comes with serious limitations. None of them are accidental, and none are villainous. They follow from what the commercial frontier stack was designed to be: an open, general tool sold at enormous scale.

That design serves a great many purposes well. It serves poorly the narrow case of a mid-sized firm that needs a component it can govern, forecast and evidence.

Designed for business?

Business software	Commercial AI
Solve real challenges	Explore AI frontier
Minimise risk	Move fast, break things
Control over data	US residency (largely)
Business focus	Consumer focus
Reduce switching costs	Vendor lock-in
Predictable cost structure	Uncertain token pricing

WHAT GETS MISSED

Would you run your website on LinkedIn?

Running a business on a commercial LLM alone is a bit like publishing a company on social media. The reach is real, and the cost of entry is near zero. The lack of ownership can be costly.

The comparison is useful because the resolution is already known. Firms did not abandon social media when they understood the trade. They kept the reach and built a site alongside it.

The right tool for the job?

Social media	Owned content platform
Noisy discourse	Controlled environment
Standard user experience	Tailored experience
Tactical reach	Strategic projects
Short-term value	Long-term value
Limited IT governance	Full IT ownership
Simple publishing	Content workflow

HOW WE APPROACH IT

Owning the whole AI stack is easier than you might think.



BENEFITS OF AN OWNED STACK

AI adds value when you can trust it.

An owned stack is not an end in itself. It removes the one condition that caps every business case: the data the work depends on can finally be used. Customer records come into scope, and AI moves to where the work already happens.

What follows is control. The experience is designed rather than inherited, spending becomes a line a firm can forecast, and residency, audit trails and access control turn into things that can be shown to an assessor rather than described.



Support customers
Customer records are in scope



Team workflows
AI sits where the work happens



Unique experience
Where AI appears and how it is labelled



Control spending
Predictable token and operations costs



Governance and compliance
Residency, audits and access control

BEYOND THE FRONTIER

Six models, many reasons to switch

Name	Hosting and licence	Best for	Capability (frontier = 100%)
Mistral France	EU cloud · Licence not yet published	Maximum flexibility, European throughout	66%
Lumo Switzerland / EU	Hosted only, EU data centres · Proprietary	Confidentiality built-in, no maintenance	Varies
Qwen China	A single GPU · Apache 2.0	Self-hosting at a lower cost	59%
DeepSeek China	Multi-GPU server or private cloud · Open weights	When cost per token is the binding constraint	67%
GLM China	Multi-GPU server or private cloud · MIT	Strongest self-hosted model. Few licence constraints.	78%
Gemma United States	Local or Cloud · Google's Gemma terms	Start with zero infrastructure at a low cost	29%

Derived from Artificial Analysis Intelligence Index, retrieved 8 October 2026 · Epoch Capabilities Index, May 2026. Versions compared: Mistral Large 4, Qwen3.8 27B, DeepSeek V4.1 Flash, GLM-5.3 and Gemma 4 26B.

STACK ARCHITECTURE

The model is only one part of the stack.

Most conversations about AI adoption are about which model to pick. That is the one decision a firm can reverse cheaply. The layers around it decide what matters: where the keys live, who sees the logs, and which jurisdiction does the processing.

The gateway is worth taking first. It holds the keys, caps the budgets and records every request, and it sits in front of whatever a firm already uses. Each layer can come in-house as the work justifies it.

Interface Open WebUI, LibreChat, bespoke app
Branded user experience

Gateway Otari, LiteLLM, OpenRouter
Holds keys, budgets, routing and usage logs

Inference Ollama, vLLM
Directs queries and runs models

Model Claude, Mistral, Qwen
Reasoning, generation and agentic AI

Hosting Azure, AWS, own hardware
Physical data storage and processing

COST OF OWNERSHIP

Trust and control require some upfront investment.

Renting is cheap to start and expensive to leave. Owning reverses both. The investment goes in at the front — people, review, a test harness — and buys the freedom to change model, host or vendor without rebuilding the work on top.

That freedom compounds. The second use case costs less than the first, the third less again, and the capability stays in the business rather than on someone else's roadmap.

Where the resources go

Resource	Investment now	Benefit tomorrow
Compute	Cloud or server	Lower token cost
People	Hire or train	Build expertise
Model selection	Evaluation process	No lock-in
Security	External testing	Verified compliance
Support	Support desk	Faster resolution
Implementation	Development cost	Roadmap control

NEXT STEPS

Start your journey to full-trust AI. We can help.

1. Discovery

You need to ask the right questions from the right people.

We can help uncover valuable workloads that the current AI policy rules out.

2. Select vendors

You need independent and impartial system comparison.

We only look after your interests, and receive no referral fees from anyone assessed.

3. Roadmap

You need a step-by-step approach to solve the challenge.

We will design an architecture that builds on what is already in place.

4. Handover

Your team will need training and change management.

We put in place an operating model that keeps running once the project is complete.

BOOK A FREE CONSULTATION

We can help you navigate AI adoption in your organisation.

Giles Thurston and Theo Paraskevopoulos, your Beside Partners



© 2026 Beside Partners Limited • besidepartners.com