The Local AI Playbook

A business owner's guide to private, cost-effective AI infrastructure.

You already use AI. This guide helps you decide which parts should never leave your building.

Bykov-Brett Enterprises

Turn Intent Into Impact.

THE OPPORTUNITY

The Shift: From Renting AI to Owning It.

You pay per prompt. Your data leaves your premises. Your AI stops working when the provider has an outage. And your costs scale with every new use case.

Local AI changes the equation. Run models on your own hardware, keep data in-house, and pay once instead of forever.

CLOUD AI (RENTING)

Pay per prompt, costs scale with usage

Data leaves your premises every call

Dependent on provider uptime

LOCAL AI (OWNING)

Fixed cost, unlimited usage

Nothing leaves your building

Runs on your hardware, your schedule

DECISION FRAMEWORK

When to Run Local, When to Stay in the Cloud.

This is not an either/or decision. The smartest businesses use both, intentionally.

Run Locally

Sensitive client data

High-volume repetitive tasks

24/7 internal automation

Anything you wouldn't paste into ChatGPT

Stay in the Cloud

Complex creative writing

Frontier reasoning tasks

Web research and live data

Image generation

Hybrid (Best of Both)

Route by data sensitivity

Route by task complexity

Route by cost threshold

This is how production AI actually works.

GETTING STARTED

What You Can Run Today.

The RAM rule: a model's parameter count roughly equals the GB of RAM you need. 8B model = 8GB RAM. That's it.

REPLACE TODAY

Haiku / GPT-4o Mini Tier

Cloud: Claude Haiku, GPT-4o Mini

Local: Qwen 3.5 9B, Gemma 4, Phi 4 Mini

8-16 GB RAM

Most people use AI at this tier and don't realise it.

REPLACE SOON

Sonnet / GPT-5 Tier

Cloud: Claude Sonnet, GPT-5, Gemini 2.5 Pro

Local: Kimi 2.6, DeepSeek V4, Qwen 3.5 35B

32-64 GB RAM

Getting close. Viable for many business workflows now.

NOT YET

Opus / GPT-5.5 Tier

Cloud: Claude Opus 4.7, GPT-5.5, Gemini Ultra

Local: No equivalent yet.

6-9 months away

But most tasks in your pipeline can be routed to a smaller model. That's the hybrid play.

FIVE MINUTES TO FREE AI

Download an App. Pick a Model. Done.

No terminal. No code. No account required.
Five minutes from now, you'll have private AI running on your own hardware.

LM Studio

Mac / Windows / Linux

ChatGPT-like interface.
Built-in model browser.
Search, click, download.

Best for most people.

Locally

iPhone / iPad

AI in your pocket.
Runs 1–4B models on modern iPhones.
Fully offline.

Private AI on the go.

Ollama

Mac / Windows / Linux

Command-line power tool.
API server built in.
What we use in production at BBE.

For technical teams.

PICK YOUR MODEL

Match your RAM to a model.

8 GB	Gemma 4, Phi 4 Mini
16 GB	Qwen 3.5 9B, Gemma 4
32 GB	Qwen 3.5 35B, DeepSeek
64 GB+	Kimi 2.6, Llama 70B

Start with Gemma 4. Swap later once you know what you need.

THE BUSINESS CASE

Cost, Control, Continuity.

£0

per token, once hardware is in place

100%

of your data stays on-premises

24/7

availability, no rate limits or outages

CLOUD API COSTS

Costs scale linearly with usage. Every new automation, every new team member, every new use case adds to the monthly bill.

A mid-size firm running 50,000 API calls/month can spend £2,000–5,000+ on cloud AI alone.

Plus: vendor price changes, rate limits halting operations, and concentration risk.

LOCAL INFRASTRUCTURE

One-time hardware investment. Run it 10 times or 10 million times, same cost. Your bill doesn't grow with adoption.

A dedicated AI server pays for itself within 3–6 months at moderate usage volumes.

Scale your AI usage without scaling your costs.

DATA SOVEREIGNTY

Every prompt you send to a cloud API is data you no longer control.

05

Client briefs. Financial reports. Internal strategy documents. Legal correspondence. HR records. Competitive intelligence.

GDPR & UK Data Law

Sending personal data to third-party servers creates compliance liability.

Client Confidentiality

Professional services firms owe duty of care over client data.

Vendor Concentration

Single-provider dependence is a business continuity risk.

THE OBJECTION

"But My Team Isn't Technical."

They don't need to be. You don't need a mechanic to drive a car. You need a reliable vehicle and someone to service it.

Owner / Sponsor

Sets direction, approves use cases, owns the business case.

Technical skill: None

Process Lead

Identifies workflows, measures outcomes, manages adoption.

Technical skill: Minimal

AI Operator

Uses the tools daily. Prompts, reviews output, feeds back quality.

Technical skill: Basic

Specialist Partner

Builds infrastructure, selects models, handles the technical layer.

That's us.

IN PRODUCTION

What It Looks Like in Practice.

This isn't theory. We run 20+ local AI models in production for real business automation, every day.

Local Layer

Ollama on NVIDIA GB10

128GB unified memory

20+ models serving simultaneously

Utility tasks, classification, monitoring, internal ops

Zero per-token cost

Agentic Layer

Claude SDK

Complex multi-step tasks

Tool-using agents with business context

Content creation, client delivery, strategy

Creative Layer

Anthropic API (Opus)

High-quality creative writing

Newsletter editorial, client communications

Frontier model quality where it matters most

Worth paying for

USE CASES

Where Local AI Creates Immediate Value.

Operations

Internal reporting, process monitoring, scheduling, task triage

LOCAL

Client Delivery

Document summarisation, briefing packs, sensitive data extraction

LOCAL

Finance

Invoice processing, receipt classification, expense categorisation

LOCAL

Content

Drafting, reformatting, translation, repurposing across channels

HYBRID

Reporting

Data cleaning, pattern recognition, narrative generation from metrics

LOCAL

Knowledge

Internal Q&A, onboarding assistant, policy lookup, SOPs on demand

LOCAL — HIGHEST VALUE

IMPLEMENTATION

Your 90-Day Adoption Path.

Not a 25-minute setup. A structured rollout with gates, metrics, and clear decision points.

01

Audit

WEEKS 1-2

Map AI workflows. Identify high-volume, low-risk candidates for local deployment.

02

Pilot

WEEKS 3-6

Deploy one use case. Measure quality, speed, and cost against cloud baseline.

03

Integrate

WEEKS 7-10

Connect to business tools. Automate routing. Train the team on daily workflows.

04

Scale

WEEKS 11-12

Expand model library. Add use cases. Measure ROI against initial investment.

Decision gate: scale, pause, or adjust.

Ready to Own Your AI Infrastructure?

We build production-grade local AI infrastructure for businesses. Hardware selection, model deployment, hybrid routing, and ongoing support.

Book a Free AI Infrastructure Audit

Hybrid Stack Design Session

jamie@bykovbrett.net
bykovbrett.net

Bykov-Brett Enterprises

Turn Intent Into Impact.