Power OCR: private document reading for your company

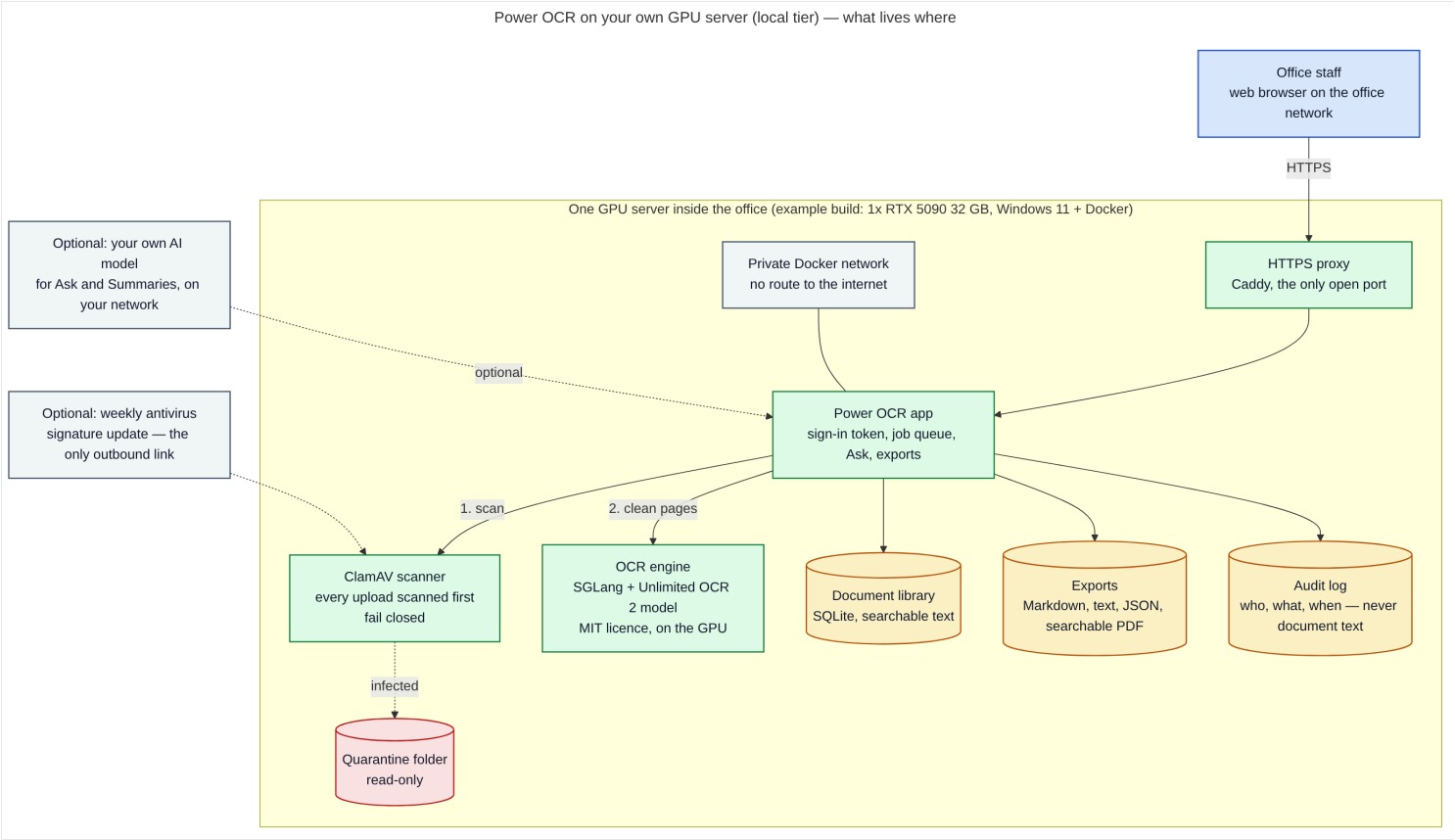
For: IT, security and operations buyers · AI Playground LLC · Nov 2026 · example build: one RTX 5090 workstation, measured Nov 12, 2026

> **What you control.** 1. **Where documents are processed:** your building, your cloud account, a server rented under your contract, or our pilot workstation. 2. **How long anything is kept:** nothing, by default. 3. **Which AI answers questions about your documents:** none, or your own model.
>
> The objective: turn scanned paper into searchable, exportable text, fast, without your documents leaving your control.

What Power OCR does

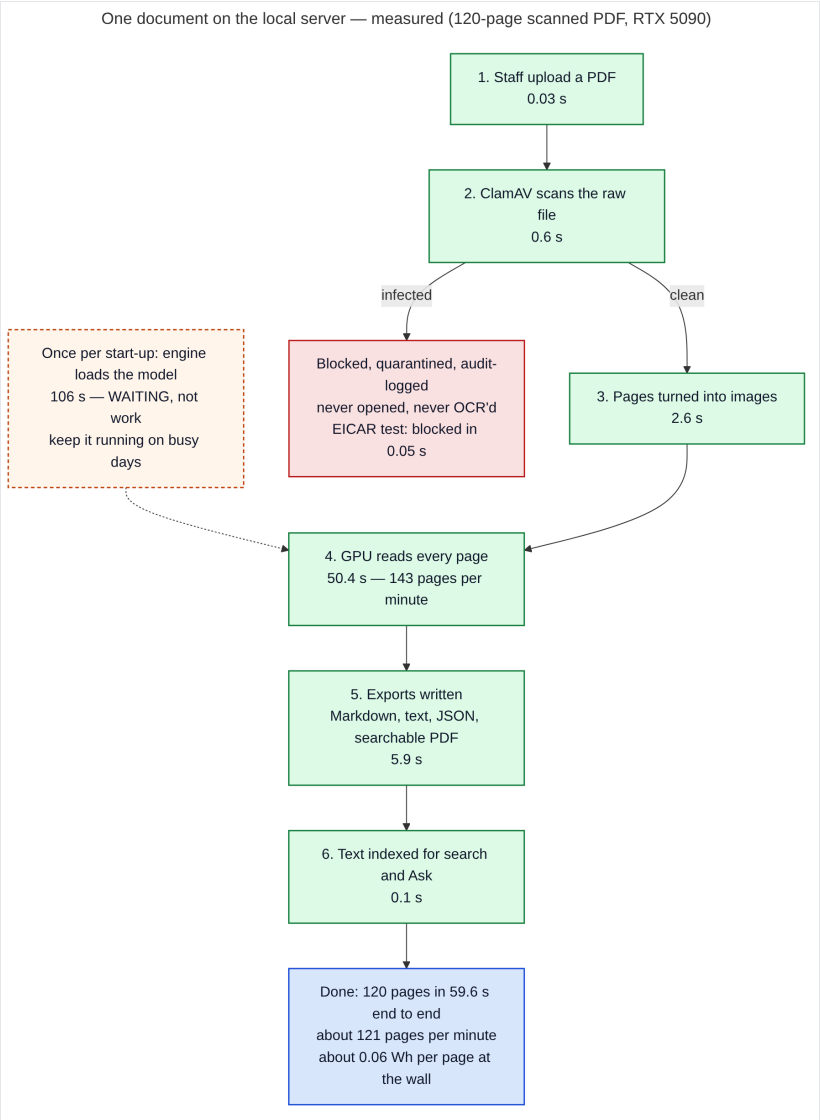
Upload PDFs or images and get back Markdown, plain text, JSON and a **searchable PDF**. Each document is malware-scanned before anything opens it. It can then be indexed so staff can search it and ask questions, with answers that cite the page they came from. Power OCR runs on one NVIDIA GPU, needs no internet, and uses an MIT-licensed OCR model with the call-home code removed.

What lives where (example build)



- **One open door.** An HTTPS proxy is the only open port; the app, scanner and engine sit on a private network with no route to the internet.
- **Scan first, fail closed.** ClamAV checks every upload before any PDF or image code touches it. If the scanner is down, uploads are refused.
- **The audit log** records who, what and when. It never holds document text.

Where the time goes (measured from the job log)



Document	Pages	End to end	OCR stage	Pages/min (end to end)
Typed letter	2	1.5 s	1.2 s	79
Real court scan	30	17.2 s	12.5 s	105
Same scan ×4	120	59.6 s	50.4 s	**121**

Where the time goes	What it is	Who fixes it
Reading the pages on the GPU (85%)	The OCR model at 143 pages/min. This is the work.	Nobody; it is already fast. A faster or second GPU scales it.
Engine start-up: 106 s, once	Loading the model onto the GPU. Waiting, not work.	You choose: keep it warm during working hours, off overnight.
Exports: 5.9 s per 120 pages	Building the searchable PDF and the other formats.	Us (later). It is small and measured.
Scan, render, index: about 3.3 s	ClamAV 0.6 s, page images 2.6 s, index 0.1 s	Nobody; it is already fast.

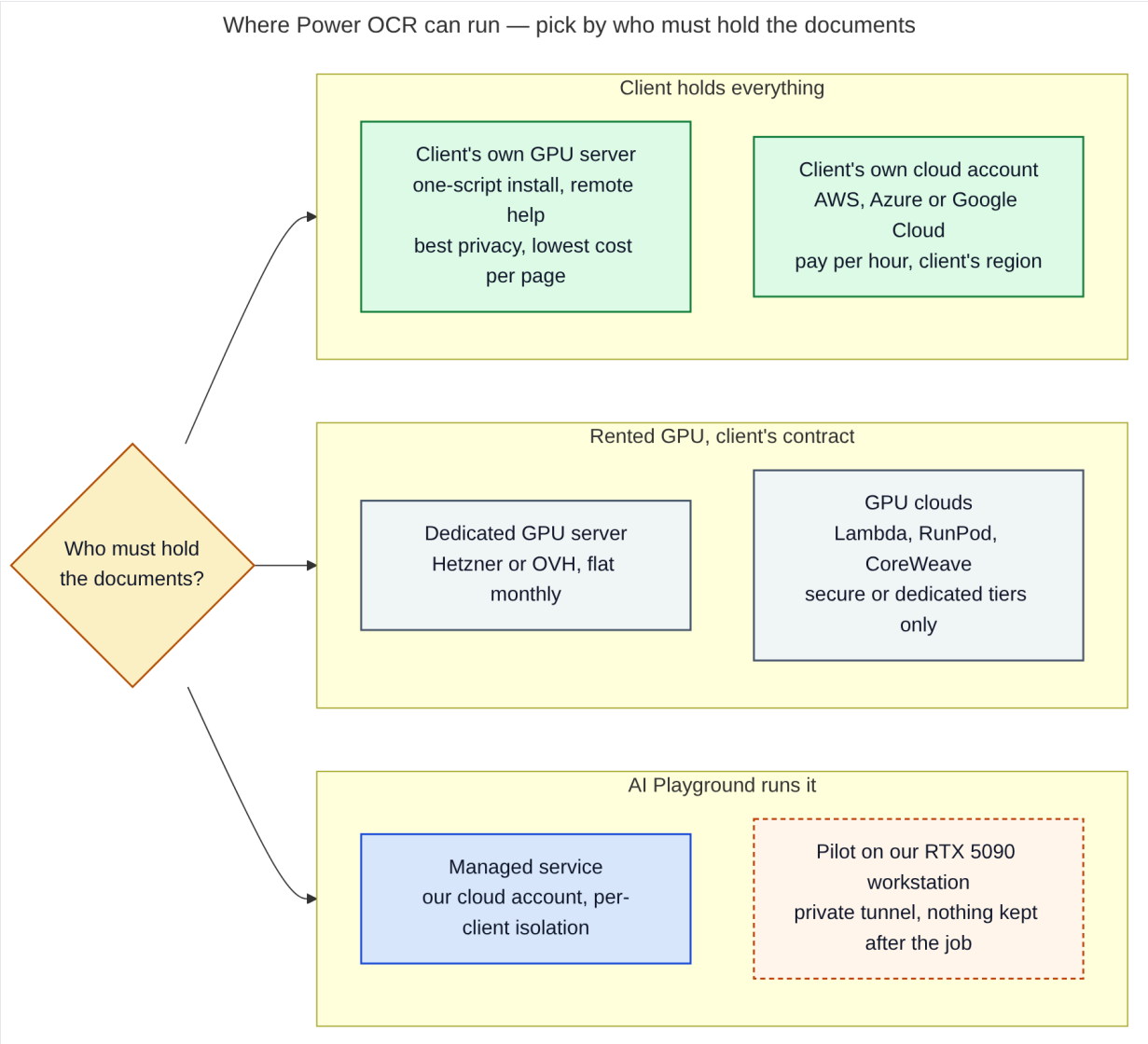
Safety check, same run: a standard antivirus test file renamed `.pdf` was **blocked in 0.05 s**, quarantined, audit-logged, and never reached the OCR engine. Requests without a sign-in token were refused.

The real money

Item	Cost	What it buys
Electricity on your own GPU server	**about $0.01 per 1,000 pages** (measured GPU power plus an estimate for the rest of the PC, at 15.9¢/kWh)	the work itself
Your own cloud account, 1x L4 GPU	about **$0.49-0.56 per 1,000 pages** while busy; GPU auto-stops when idle	pay only for hours used, your region, your keys
Dedicated GPU server (EU)	from **€214/month** flat	steady volume, single tenant
Keeping the engine warm all day	about $0.46/day on a workstation	removes the 106 s start-up wait

For scale: 100,000 pages is about 14 GPU-hours on the example build and about $0.90 of electricity. The waiting costs minutes; the work costs cents.

How to host it



Option	Best for	Data stays
Your own GPU server (one-script install, remote help)	regulated or "never leaves" rules	in your building
Your AWS, Azure or Google Cloud account	cloud-first IT, single sign-on	in your account and region
Dedicated rented server (e.g. EU)	steady volume, EU residency	on a named single-tenant machine
Managed by AI Playground	no IT time to spare	our isolated per-client stack, nothing kept
Pilot on our workstation	trying it on your own documents first	kept on our GPU server only for the demo, then deleted when it ends (or sooner, per job); you get a deletion record per job

What you have to answer

Q1. Where should production run?

a) Your own GPU server. b) Your cloud account. c) A rented dedicated server. d) Managed by us.

Answer: _____

Why it matters: this decides who holds the documents and which contract covers them.

Q2. Keep a searchable library, or nothing at all?

a) Nothing kept: download the results and we delete everything. b) Keep a searchable library on your side, with Ask and citations.

Answer: _____

Q3. Run a pilot on your sample documents first?

a) Yes: a booked window with a private tunnel, NDA and pilot agreement, and a deletion record per job. b) No: go straight to an install.

Answer: _____

If asked

- **Do you keep our documents?** No, not by default. The library is opt-in; cloud copies auto-delete.
- **Does it phone home?** No. The upstream call-home code was removed, and the containers have no internet route.
- **What GPU do we need?** An NVIDIA card with 12 GB minimum, 16 GB recommended, and 24-48 GB for many users. The engine used about 11 GB in our run.
- **Is the licensing OK for commercial use?** The OCR model is MIT and the engine (SGLang) is Apache-2.0, with credits kept. There is no AGPL code: PyMuPDF (AGPL-3.0) was removed from the engine build on 14 Nov 2026, and PDFs render with pypdfium2 (Apache-2.0 / BSD-3). A scan of every package in the images lists what is left to note: ClamAV (GPL-2.0) runs as a separate, unmodified container; two LGPL Python libraries are used unmodified inside the engine; and the NVIDIA CUDA libraries are under NVIDIA's licence. Full list: LICENSE_INVENTORY.md.
- **How do you prove pilot files were deleted?** Each job gets a hash-chained deletion record. It lists the SHA-256 of each file as received, every artifact deleted, database rows before and after (after is 0), the engine stopped, and GPU memory before and after. In our test, the data folder and database went back to their exact pre-test counts. When the demo ends, the files and working data are deleted and the model is unloaded from the GPU. It is a normal deletion, not a forensic disk erase.
- **How accurate is it?** 97.6% character accuracy on our ground-truth test set, and 100% on an invoice table in crop mode. We will score your samples in the pilot.
- **What happens if the GPU machine is down?** On your own server, your IT restarts it; it starts working again in about 2 minutes. In the cloud tiers, a new GPU host starts automatically.