

ON-PREMISES AI FOR MANUFACTURING

Everything in our boxes, and everything it was measured on.

What is in the box, what it reads, how fast it answers, and what it does on your network.

200k

documents, photos and voice notes on one unit

1.4 s

median synthesis answer

15/15

answers citing a source, by schema

0

outbound connections off the unit

01 · WHICH ONE

Mini is sized for one line or one small plant; Scale for a site. Where a figure differs it is printed in both columns; where it does not, the row is marked **BOTH**.

	INPLANT MINI	INPLANT SCALE
PEOPLE	Up to 20	A few hundred
SITES	One	One or two
THROUGHPUT	8 cores, 32 GB Answering is limited by memory bandwidth and indexing by cores, and Scale has more of both — so it answers quicker, indexes a first import quicker, transcribes quicker, and holds more people asking at once.	Bigger silicon, 64 GB unified
BUYING IT	Order it, we ship and invoice	We quote first, then ship
PRICE	\$299 / month + shipping.	\$2,640 / month + shipping.

02 · COMPUTE

	INPLANT MINI	INPLANT SCALE
PROCESSOR	AMD Ryzen™ 7 8845HS	AMD Ryzen™ AI Max+ 395
MEMORY	32 GB	64 GB unified

INFERENCE	BOTH	Entirely on this unit
CHAT MODEL	BOTH	qwen2.5:3b, constrained decoding
EMBEDDING MODEL	BOTH	nomic-embed-text, 768 dimensions

03 · STORAGE, COOLING AND POWER

	INPLANT MINI	INPLANT SCALE
PRIMARY DISK	512 GB NVMe	SanDisk SN7100 PCIe® 4.0 M.2 2280 · 4 TB
SECONDARY DISK	None	SanDisk SN7100 PCIe® 4.0 M.2 2280 · 1 TB
HOLDS	BOTH	Your documents, the index, and the log of what was asked
COOLING	Fan, in the chassis	Noctua NF-A12x25 HS-PWM, 120 mm
MOVING PARTS	BOTH	One, the fan. Storage is solid state.
POWER	19 V DC, external supply	AC cable, C13, US/CA
CONNECTIONS	BOTH	Power, one Ethernet cable.

04 · WHAT IT READS

Identical on both appliances. Everything becomes a citable chunk; only the locator differs: a page for documents, a timestamp range for audio, the filename for a photograph, a sheet and zone for a drawing.

TYPE	FORMATS	CITES AS
DOCUMENTS	.pdf .docx .pptx .xlsx .html .md .txt	manual.pdf, p. 47
PHOTOGRAPHS	.png .jpg .tif .bmp .webp	nameplate.jpg
VOICE	.wav .mp3 .m4a .flac .ogg .aac	handover.m4a, 2:15–3:00
CAD DRAWINGS	.dxf SHIPPING SOON	sheet and grid zone. A .dxf uploaded today is refused; a drawing exported to PDF works now.

VIDEO	Not supported currently	—
READS FROM	SMB share, SharePoint, or any folder Nothing has to be moved or re-filed first; re-reading the same directory updates in place rather than duplicating.	
SCANS AND PHOTOS	OCR, through the same pipeline A scanned manual becomes a one-page document with a text layer, so it chunks, cites and retrieves exactly like a born-digital PDF. OCR language is set per corpus from about fifty options, not guessed per file.	
SPOKEN LANGUAGE	Detected per recording — a handover note in Spanish and one in English sit in the same corpus.	
BEFORE YOU FEED IT	General engineering knowledge	

05 · MEASURED LATENCY

Measured over a corpus of machine manuals, shift-handover voice notes and nameplate photographs.

SOCIAL, META	LOOKUP	SYNTHESIS
Instant Canned, or read from index metadata. No model runs.	350–850 ms Verbatim from the index, no model. Observed 351–613 ms across repeat runs; the range is quoted rather than an average, because the spread is real.	1.4 s median 3B model with constrained decoding. Zero of fifteen questions over 5 s.

Why a 3B model with a response schema

MODEL	MEDIAN	OVER 5 S	GROUNDED	CITED	REFUSED CORRECTLY
4B reasoning, terse prompt	16.2 s	15/15	12/15	14/15	5/5
3B, terse prompt	3.4 s	2/15	12/15	4/15	5/5
3B, terse + schema · shipping	1.4 s	0/15	12/15	15/15	5/5

Identical grounding, twelve times faster, and the citation rate goes from 4 of 15 to 15 of 15. Asking for citations in the prompt produced them a quarter of the time; requiring a response schema with a non-empty citations array produced them every time.

06 · NETWORK BEHAVIOUR

Identical on both appliances.

INBOUND	A browser on your network — laptop, tablet at the line, or phone. Nothing installed on the client; the traffic does not leave your site.
OUTBOUND	Nothing off the unit — the model, the index and the documents are on the same box, so there is nowhere else for a request to go.
WHILE INDEXING	Still nothing — transcription is installed during the build; OCR is compiled into the software.
INTERNET ACCESS	Not required, not used — can be activated for updates and new versions.
UPDATES	Downloaded and applied by you — we publish software and model updates; you fetch them and upgrade when it suits you.
TELEMETRY	None — the query log stays on the unit and is readable in the admin panel.

07 · STATED LIMITS

It won't state a specification it didn't read.

Anything shaped like a number about your equipment — torque, clearance, RPM, tolerance, pressure, capacity — comes from your documents.

It won't pass general knowledge off as something it read.

Ask what R-8 means and it answers, labelled *general knowledge, not from your documents*, with no sources attached and none implied.

It won't let a note override a written procedure.

Safety questions are answered from published documents only, and the restriction is stated in the answer rather than applied quietly.

It won't read a scan a person couldn't read.

On a creased, oily photocopy it returns what is legible and cites the page so you can check it yourself.

It won't replace the person who knows the machine.

It makes what they already wrote down reachable in seconds — a smaller claim, and a true one.

When the documents don't cover it, it says so.

The question lands in the admin activity view marked *unanswered*, so you find out what your plant hasn't written down.

08 · COMMERCIAL TERMS

PRICE

\$299 / month

INPLANT MINI

\$2,640 / monthINPLANT SCALE

Plus shipping, hardware included. Month to month, no minimum term, no per-user fee, no query limit, no deployment charge. Thirty days to send the appliance back for a full refund.

THE SOFTWARE

All of it, on both appliances. No tier above, nothing metered.

SUPPORT

Remote, from the people who built it.

hello@inplant-ai.tech

inplant-ai.tech/datasheet