Air-gapped edge AI

Language, vision, speech and document AI on hardware you own, in places with no connectivity. For defence, government and regulated enterprises that cannot use a cloud API.

AT A GLANCE

DEPLOYMENT
Under 5 minutes
Case open to first inference

CONNECTIVITY
None required
Air-gapped by design

COMPUTE
Up to three nodes
NVIDIA enterprise GPUs

DATA
Never leaves site
Zero egress, no telemetry

SPECIFICATION

Compute

Compute nodes	One, two or three (two standard)
Acceleration	NVIDIA enterprise GPU
Memory	128GB unified LPDDR5x per node
Storage	4TB NVMe per node
Node interconnect	200Gb/s QSFP, direct-attach copper
Site network	10GbE RJ45

Power

Input	110-240V AC / 12-24V DC
Battery	1000Wh LiFePO4
Solar	400W MPPT input
Runtime	8 hours (typical load)

Physical

Case	Mil-spec polymer
Rating	IP67 (closed)
Weight	45kg / 99lbs
Temperature	-20°C to +60°C

Assurance

Frameworks	Aligned with NIST SP 800-53 Rev. 5, NCSC CAF v4.0, MOD Secure by Design
Data egress	None: no telemetry, no licence server
Encryption at rest	AES-256
Updates	Signed offline bundles, verified before install, with rollback

INSTALLED CAPABILITIES

Local LLM Inference
High-throughput local text generation on the node GPU.

Text Translator
Neural translation for operational documents and intercepted material.

Audio Transcription
Speech to text across a wide range of languages, long recordings included.

Speaker ID
Voice profiling and speaker classification in multi-speaker audio.

Image Analysis
A vision-language model that answers questions about an image.

Omni-LLM
One endpoint for chat, image understanding and short audio input.

Facial Recognition
Face embedding and identity verification with liveness detection.

OCR Service
Scans and PDFs to structured Markdown, headings and tables included.

RAG Storage & Retrieval
A vector store and chunking pipeline for retrieval-augmented answers.

Image Generation
On-node diffusion for concept work and synthetic training data.

MEASURED PERFORMANCE

September 2026. One RAPTOR compute node. Figures are per node. A unit with two or three nodes can run more capabilities at once, but it does not make one model faster.

CONCURRENT USERS	TOTAL OUTPUT	PER USER	FIRST TOKEN P50	FIRST TOKEN P95
1	50 tok/s	51 tok/s	110 ms	114 ms
8	243 tok/s	32 tok/s	469 ms	557 ms
16	367 tok/s	24 tok/s	594 ms	968 ms
32	526 tok/s	17 tok/s	727 ms	1,514 ms
64	721 tok/s	12 tok/s	839 ms	2,856 ms

Document OCR	**26 pages / min**
Image analysis	**325 ms to first token**
Transcription	**3.8× realtime**
Image generation	**19.7 images / min**

> Figures are measured, with caching disabled and unique prompts. Method and the runs we discarded are published at raptorai.uk/performance/

ANALYSIS MODULES

Audio and language

- Multi-lingual audio transcription
- Text translation
- Multi-lingual document OCR
- Speaker and voiceprint identification
- Keyword alerting

Computer vision

- Image analysis
- Facial recognition
- Image object detection
- Multi-lingual image OCR

Local AI

- Local large language models
- Retrieval augmented generation
- Repository-aware coding assistant
- Code generation with no external calls

PRICING

Trial deployment	£5,000	A RAPTOR unit on your site, run against your own work before you commit.
Base standard loadout	£20,000	Two compute nodes, the standard loadout; Full App Store, core model catalogue; Perpetual software licence, with 1 year of updates and new models; 3-year warranty and support
Pro enhanced loadout	On request	Three compute nodes, the full loadout; 1000Wh battery, 8 hours at typical load; 400W solar input for off-grid running; Full model catalogue, 24/7 support
Ultra bespoke loadout	On request	Your own models integrated and benchmarked; On-site specialist for installation; Priority operational support; 5-year warranty

> Prices in pounds sterling, excluding VAT. The software licence for a delivered version is perpetual.

PROCUREMENT

RAPTOR unit

The platform on hardware we build, test and benchmark. **From £20,000.**

- Hardware, software and models delivered as one unit
- Benchmarked before shipping, so capacity is known before you buy
- Deploy from closed case to first inference in under five minutes
- Support and update bundles sized to your security process

Platform licence

The App Store and model catalogue on your own hardware. **Per host, on request.**

- Licensed per physical host that serves models, not per call or per token
- Runs on x86 or ARM with an NVIDIA GPU; we confirm your configuration first
- Same offline install path: signed bundles, no connectivity required
- Bring your own models alongside the shipped catalogue

> What happens when a licence term ends? It keeps running. The licence for your delivered version is perpetual, and the term only covers support, updates and new models. If it lapses, those stop arriving and nothing on the unit is disabled.