

PRIVATE ENTERPRISE AI APPLIANCE



Nexus One AI

Technical buyer datasheet for private AI deployment, infrastructure review, and platform evaluation.

Product Datasheet | July 2026

Install Model Bootable ISO or scripted installer for Ubuntu 22.04	Runtime Stack Ollama, Open WebUI, ChromaDB, FastAPI, observability, admin controls	Evaluation Fit Private AI buyers assessing security, operability, and infrastructure readiness
---	--	--

System Architecture

Nexus One AI is packaged as a deployable private AI stack that combines local model serving, document retrieval, user access controls, management APIs, and system monitoring on customer-controlled infrastructure.

Layer	Primary Components	Technical Role
Inference	Ollama, local model catalog	Runs LLM inference on customer CPU/GPU resources without requiring public chat endpoints for core operation.
User Access	Portal UI, Open WebUI, auth layer	Provides browser access, prompt workflows, user sessions, and governed entry to AI tools.
Knowledge Retrieval	ChromaDB, upload pipeline, RAG paths	Indexes internal content and supports grounded question-answering over local documents.
Application & Admin	FastAPI backend, settings, audit, backup logic	Exposes management, branding, licensing, audit, backup, and service-control functions.
Observability	Grafana, Prometheus, GPU/system metrics	Supports operational monitoring, readiness checks, and infrastructure health review.

Deployment and Runtime Model

- Deploys on customer-controlled Ubuntu 22.04 infrastructure as an appliance-style stack rather than a hosted SaaS tenant.
- Provisioning options include a bootable ISO for zero-touch bring-up and a scripted installer for fresh server builds.
- Designed to run fully within the customer network after installation, including environments that restrict or avoid public internet access.
- Supports CPU-only baseline operation and scales upward with NVIDIA GPU capacity for stronger local inference and multi-user responsiveness.

Control and Risk Surface

- Prompt traffic, uploaded documents, application data, and model execution stay within the customer-managed environment.
- Authentication, user administration, audit logging, backup workflows, and service status are part of the packaged control plane.
- Observability components help buyers evaluate operational readiness rather than treating monitoring as a separate post-sale project.
- Final production posture should still be validated against customer IAM, backup, network segmentation, and compliance requirements.

Technical Fit Profiles

Profile	Infrastructure Shape	Expected Fit	Practical Notes
CPU Core	8+ CPU cores, 32 GB RAM	Pilot or low-concurrency evaluation	Good for portal, admin, document workflows, and limited local model usage.
GPU Starter	Single NVIDIA GPU, 24-48 GB VRAM class	Departmental production fit	Improves model responsiveness, document Q&A, and sustained internal use.
GPU Pro	Higher VRAM or multi-GPU server	Multi-team production rollout	Better suited for larger workloads, advanced routing, and stronger concurrency.
GPU Max	Large multi-GPU enterprise host	Custom enterprise package	Supports more demanding private AI programs and specialized deployment sizing.

Buyer Evaluation Checklist

The table below is structured for infrastructure, security, and procurement review teams evaluating whether the platform fits their internal AI deployment criteria.

Evaluation Area	What To Verify	Current Platform Position
Security Boundary	Whether prompts, documents, and model inference remain inside the customer environment.	Designed for on-premises use with local inference, local storage, and internal-network deployment patterns.
Deployment Method	How quickly the stack can be provisioned and reproduced across test, demo, and customer environments.	Supports both a bootable ISO path and an automated installer path for Ubuntu 22.04 systems.
Operations	Whether admins can monitor services, capacity, and readiness without separate tooling projects.	Includes system status, metrics, backup controls, audit surfaces, and appliance operations pages.
Extensibility	Whether internal teams can build or expose custom flows on top of the base stack.	Includes FastAPI, JupyterLab, and service building blocks for extension and internal integration work.
Commercial Fit	How packaging scales from departmental pilot to broader multi-team rollout.	Tiered packaging supports starter through enterprise-scale deployments with increasing controls and features.

Commercial Packaging View

Packaging is designed to align commercial scope with technical breadth, so buyers can evaluate a smaller controlled rollout first and expand to broader operational use later.

Tier	Users	Deployment Intent	Included Scope
Starter	Up to 10	Department pilot or contained team rollout	Private chat, portal, RAG basics, backup and audit foundations.
Basic	Up to 25	Single-department production deployment	Adds meeting assistant, broader knowledge features, and limited connector scope.
Pro	Up to 100	Multi-team production program	Advanced RAG, connectors, model routing, stronger guardrails, and workflow automation.

Max	Custom	Large enterprise or specialized infrastructure fit	Custom GPU sizing, advanced tuning paths, and expanded enterprise control scope.
-----	--------	--	--

Deployment Requirements

Operating Base	Ubuntu 22.04 server environment with internal network access and sufficient local storage for models, logs, and indexed content.
Minimum Evaluation Build	8+ CPU cores, 32 GB RAM, 100+ GB free storage. Suitable for controlled evaluation and lower-concurrency use.
Recommended Production Build	NVIDIA GPU-backed server for stronger local inference, larger document workloads, and multi-user responsiveness.

Selection Summary

- Best suited for buyers who want an on-premises AI platform rather than a thin front end on top of external hosted models.
- Stronger than a basic demo stack when the evaluation criteria include admin controls, packaged monitoring, deployment repeatability, and operational visibility.
- Most compelling where procurement, IT, and security teams need a concrete private-AI deployment story they can inspect and score.

Buyer note: Final sizing, model catalog, connector scope, identity integration, and backup posture should be confirmed against the buyer’s target workload and internal standards.