ENTERPRISE AI PLATFORM · WORKSTATION & SERVER

Nexus One AI

Technical buyer datasheet for Nexus One AI Workstation and Server deployment, infrastructure review, and platform evaluation.

Product Datasheet | July 2026

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

Product Structure

Nexus One AI is an enterprise AI platform delivered in two deployment models. Both share the same software platform, licensing model, and management experience — only the deployment form factor changes.

Nexus One AI Workstation	Nexus One AI Server
Personal AI appliance for individual developers, data scientists, researchers, universities, innovation labs, and CXOs. 1-3 active users, 5 max. Not a shared server — no HA or server-redundancy expectations.	Shared enterprise AI platform for teams, departments, and organizations. Four tiers — Server S, Server M, Server L, Server Max — scaling from small-team pilots to 100+ user enterprise deployments.

This separates deployment form factor from enterprise capability, allowing the platform to scale from a single workstation to multi-node enterprise clusters without changing the user experience.

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	Departmental production	Improves model

	GB VRAM class	fit	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 a single developer workstation to broader multi-team rollout.	Nexus One AI Workstation for individual use, plus Server S through Server Max for department-to-enterprise-scale deployments with increasing controls and features.

Server 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. This table covers the Nexus One AI Server line; Nexus One AI Workstation is a separate, standalone category (see “Product Structure” above) and is not part of this tier ladder.

Server Tier	Users	Deployment Intent	Included Scope
Server S	Up to 10	Department pilot or contained team rollout	AI Workspace, portal, RAG basics, backup and Governance & Audit foundations.
Server M	Up to 25	Single-department production deployment	Adds meeting assistant, broader knowledge features, and limited MCP & Enterprise Connectors scope.
Server L	Up to 100	Multi-team production program	Advanced RAG, connectors, model routing, stronger guardrails, and Workflow Designer automation.
Server 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.

Platform Roadmap

Near-term enterprise capabilities planned for Nexus One AI, building directly on today's platform. Direction, not dates.

Capability	Scope
Enterprise Authentication	LDAP, Active Directory, SAML, OAuth, OIDC.
MCP & Enterprise Connectors	SharePoint, OneDrive, SAP, Oracle, SQL, REST, GitHub, Jira, ServiceNow via MCP.
AI Monitoring	GPU/CPU/RAM/VRAM, latency, tokens/sec, queue depth, model usage, active sessions, health.
Governance & Audit	Prompt history, model usage, compliance, PII detection, policy violations, usage reports.
Model Catalog	Browse, install, update, rollback, versioning, recommended models.
Deployment Management	Node management, cluster health, storage, model synchronization, upgrade status, licensing overview.
Secrets Management	Secure storage for API keys, passwords, certificates, tokens, external credentials.

Beyond these near-term items, the same additive pattern used for Workstation — a standalone deployment model alongside Server, not inside its tier ladder — is the intended template for future Cluster, Edge, and Hybrid Cloud deployment models running the same software platform.

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.