> TOPNOTCH.TECH

PRIVATE AI SPECS

PROCUREMENT BRIEF

Purpose: To outline the specific physical compute components, software layers, and project pricing required to assemble the private AI server for [REDACTED].

Last Updated: April 2026

PRIVATE AI SPECS

Target	[REDACTED]
Audit Date	October 2025
Scope	Local Private AI (Hardware & Software)
Outcome	APPROVED

CAPABILITIES

The deployment includes a 70-Billion parameter Large Language Model (LLM) and a vector database generated from [REDACTED]'s data. This infrastructure operates independently of external vendor APIs and prevents company data from leaving the local network.

The system provides:

- **Local Data Isolation:** Restricts all data reading, synthesis, and analysis to the physical premises. No information exits the local network.
- **System Independence:** Operates without external internet connectivity, removing reliance on third-party vendors. (Regular software updates will require connectivity, but are not mandatory for operation.) Offline capability prevents downtime caused by upstream API modifications, altered schemas, or model deprecations.
- **Company Data:** Indexes business documents and files into a local, private database, never shared with any AI company. This setup allows the model to generate contextual responses based on internal records rather than generic public data.
- **Uncapped Batch Processing:** Executes high-volume text parsing, including [REDACTED] processing. Processing is private (processed data is not shared with any AI company), and does not encounter external token rate limits or computing fees.

PERFORMANCE METRICS

- **System Latency:** Database queries and text answers in under 2,000 milliseconds.

- **Traffic Control:** Backend processing queue to manage simultaneous user requests smoothly during peak operational hours.
- **Remote Login:** Secure remote worker access through encrypted network authentication; requires two-factor authorization (2FA).
- **Consistency:** Sustains token output speeds matching standard reading speed, without degradation during continuous multi-user loads.

RECOMMENDED HARDWARE

Component	Specification & Rationale
GPU Array	**4x NVIDIA RTX 4090 (24GB) or 2x RTX 6000 Ada (48GB).** 96GB VRAM minimum required to keep the model weights in fast memory and prevent slowdowns.
CPU	**AMD Threadripper PRO or AMD EPYC.** Provides 128 PCIe lanes to eliminate data transit bottlenecks between GPUs.
System RAM	**256GB - 512GB DDR5 ECC.** Error-correcting memory recommended to guarantee 24/7 server uptime.
Primary Storage	**2x 2TB NVMe M.2 (RAID 1 Mirror).** High-speed SSD for fast container resets. RAID for drive failure protection.
Vector DB Array	**2x 4TB NVMe M.2.**
Power Supply	**2x 1600W ATX 3.0.** Handles power draws up to 1800 Watts under load. (Requires a dedicated 20-Amp circuit.) *While standard office wiring solutions are possible, for high-speed concurrent access intended for [REDACTED]'s application, this build is recommended. Standard wiring solutions require changes to the GPU Array.*
Networking	**Standard 1GbE/10GbE NIC.**
Enclosure	**Full-Tower Airflow Case.** Uses high-volume, low-RPM cooling fans. *(Recommend "Noctua" or "be quiet!")*
Peripherals	**Headless Operation & USB Crash Cart.** No permanent monitor needed. Local administrative laptops can plug directly into the tower for troubleshooting.

(Important Note: The 4-GPU configuration maximizes the available CPU PCIe lanes and chassis power delivery. It cannot be physically expanded. Scaling compute capacity in the future will require the independent hardware nodes.)

SOFTWARE

Infrastructure Layer	Configuration & Operational Rationale
Host Operating System	Ubuntu Server 22.04 LTS
Driver Layer	Version-Locked NVIDIA Drivers & CUDA Toolkit.
Containerization	Docker Engine with NVIDIA Passthrough. Keeps the text processing server, database, and front-end interface isolated; strict memory limits.
Inference Server	vLLM Framework. 70-Billion parameter model configured to operate offline.
User Interface (UI)	Open WebUI or LibreChat. Provides a browser-based chat window that matches the look and feel of ChatGPT.

COST & TIMELINE

Estimated Cost

$[REDACTED] to $[REDACTED]
(Varies based on component rates at the time of purchase)

Project Schedule

[REDACTED] weeks total.

Deploy Steps

1. Hardware Procurement
2. Physical Assembly
3. Software Orchestration
4. Pre-Transit Performance Testing
5. On-Site Deployment
6. Disaster Recovery Validation (2-10 weeks post deploy)

~ End of Document ~

Full deployment specifications for this system are available via The Backchannel. Access is subject to corporate domain verification.

https://topnotch.tech/backchannel/private-ai/