Private LLM Setup Guide

Terminal Velocity AI · terminalvelocityai.tech

A step-by-step technical guide to deploying Ollama and open-source models on your own infrastructure. Keep your data on-prem, eliminate per-token costs, and maintain full control over your AI stack.

1. Hardware Requirements

Minimum (dev/testing)

- 16 GB RAM, modern CPU (8+ cores)
- 50 GB free disk space for model weights
- Works with: Llama 3.2 3B, Phi-3 Mini, Mistral 7B (quantised)

Recommended (production)

- NVIDIA GPU with 16–24 GB VRAM (RTX 3090, A4000, or better)
- 64 GB RAM, NVMe SSD for fast model loading
- Works with: Llama 3.1 70B (Q4), Mixtral 8x7B, CodeLlama 34B

2. Install Ollama (Ubuntu / macOS)

Ubuntu / Debian:

1. curl -fsSL https://ollama.com/install.sh | sh
2. sudo systemctl enable --now ollama
3. ollama --version # verify installation

macOS:

1. Download Ollama.app from https://ollama.com
2. Move to /Applications and launch
3. ollama --version # verify in Terminal

3. Deploy Your First Model

Pull and run a model with a single command:

ollama pull llama3.2 # 2 GB download, great starting point

ollama pull mistral # 4 GB, strong reasoning

ollama pull codellama # 4 GB, optimised for code

ollama run llama3.2 # starts interactive chat

Tip: Use quantised models (e.g. llama3.1:8b-instruct-q4_K_M) to halve memory usage with minimal quality loss.

4. Expose an API Endpoint for Your Team

Ollama exposes an OpenAI-compatible REST API on port 11434 by default.

```
# Allow network access (set before starting Ollama)
export OLLAMA_HOST=0.0.0.0

# Test the API
curl http://localhost:11434/api/generate -d '{"model":"llama3.2","prompt":"Hello"}'

# Use with OpenAI SDK (Python)
from openai import OpenAI
client = OpenAI(base_url="http://YOUR_SERVER:11434/v1", api_key="ollama")
```

5. Security Hardening

- ■ Run Ollama behind a reverse proxy (nginx/Caddy) — never expose port 11434 directly
- ■ Add API key authentication at the proxy layer
- ■ Restrict firewall: only allow internal VPN/subnet IPs to reach the LLM server
- ■ Disable model pulling on production server (OLLAMA_NOPRUNE=1)
- ■ Run Ollama as a non-root system user with minimal permissions
- ■ Enable TLS termination at the proxy for all client connections
- ■ Audit prompt logs periodically — do not log sensitive business data

6. Monitoring & Observability

- • Metrics: expose /metrics endpoint and scrape with Prometheus
- • Dashboards: import the Ollama Grafana dashboard (ID: 20571)
- • Alerts: set up alerts for GPU memory > 90%, response latency > 5s
- • Logs: pipe ollama logs to your central log aggregator (ELK, Loki, Datadog)
- • Model versioning: pin model tags in deployment scripts for reproducibility

---

Want a fully managed private LLM deployment? Terminal Velocity AI handles the full stack — hardware selection, deployment, security, and team training. terminalvelocityai.tech