



Remoder · OLLAMA-CLI

Essentials for Modern Engineers



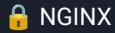
Remoder.com



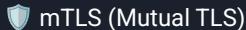
Ollama Best Practices (From an Engineering Perspective)



1. Never expose Ollama directly to the internet



A high-performance reverse proxy and load balancer that accelerates and protects your apps. Ideal for routing traffic, caching, SSL termination, and edge-level control.



Provides strong authentication by verifying both the client and server certificates. Ensures encrypted traffic and identity validation across all connections.



Centralized control point for routing, rate-limiting, authentication, and monitoring API traffic. It standardizes how services communicate and protects your backend from noise.



Creates a secure, encrypted tunnel into your infrastructure. Ideal for remote access to internal services or isolating sensitive workloads from the public internet.



Adds identity-based access at the gateway level. Only authenticated and authorized users can reach protected applications – enforcing Zero Trust by default.



2. Use Docker or Kubernetes for isolation



Version Control

Each container image locks your Ollama version and dependencies, making rollbacks, upgrades, and multi-version testing effortless and predictable.



Reproducible Environments

Your entire setup – models, configs, runtime – is captured in code. Every engineer gets the same environment, eliminating “works on my machine” problems.



Easy GPU Pass-Through

Containers make GPU access simple using NVIDIA runtimes. You can assign GPU profiles, isolate workloads, and ensure Ollama runs with full hardware acceleration.



Better Scaling Patterns

Once containerized, Ollama instances can scale horizontally using orchestrators like Docker Swarm or Kubernetes. Perfect for handling heavier LLM traffic or multi-agent workloads.



3. Protect your models & internal data

🚫 Disable Remote Model Pulls in PROD

Prevent production systems from downloading models from the internet. This protects you from supply-chain attacks and ensures only approved, vetted models are used.

🔒 Store Models in Private Registries

Keep your .gguf or model artifacts in secure, internal registries. This guarantees controlled access, versioning, and auditing of every model your agents rely on.

📜 Log & Monitor All Requests

Track every inference, prompt, and API call. This provides visibility, detects anomalies, and helps meet compliance and incident-response requirements.

⚙️ Use Rate Limiting & Auth Everywhere

Enforce authentication on all AI endpoints and throttle incoming requests. This protects your systems from overload, brute-force attempts, and unauthorized use.





4. Move to GPU-backed infra when scaling



A fast, cost-efficient GPU playground ideal for development, testing, and running AI agents with full GPU acceleration. Spin up, experiment, and tear down in minutes.



Great for scalable, production-grade AI workloads. Flexible instance sizes, strong networking, and seamless integration with AWS tooling make it perfect for enterprise agents.



Optimized for GPU-heavy inference and training with tight Azure ecosystem integration. Excellent choice for teams already leveraging Azure's monitoring, security, and networking stack.



Top-tier performance for large LLMs, vector workloads, and multi-agent inference. A2 and H100 boxes deliver massive compute for high-throughput, low-latency AI deployments.





5. Treat AI services like any production microservice



Dashboards, metrics, and alerting give full visibility into model performance, latency, GPU usage, and agent behavior – essential for real production workloads.



Health Checks

Automated liveness and readiness checks ensure your Ollama containers self-heal, restart on failure, and stay responsive under load.



Environment Isolation

Separate dev, test, and prod environments prevent cross-contamination of models, configs, and secrets – keeping deployments predictable and secure.



Failover & Autoscaling

Load spikes and node failures are handled automatically through replicas, horizontal scaling, and restart policies – keeping AI agents always available.



Ollama is simple – but production isn't.

And this is exactly where DevOps, SRE, Cloud, and Platform engineers shine – transforming lightweight tools into reliable, scalable, secure AI platforms.





Why This Matters

- Engineers don't just *run* models – we build the systems that make AI reliable, secure, and scalable.
- That is the real gap in the AI world today.
- And that's exactly what we teach inside **Remoder**.



 *Re-modernizing Engineering for the AI Era – where human brilliance meets machine intelligence.* 

Questions? Reach out to us

 Website: <https://remoder.com>

 LinkedIn (Company): <https://www.linkedin.com/company/remoder>

 LinkedIn (Sanjars): <https://www.linkedin.com/in/sanjars/>

 YouTube: <https://www.youtube.com/@remoder-inc>

 Contact: hello@remoder.com OR Directly Message Me: <https://www.linkedin.com/in/sanjars/>

 Want the Full Walkthrough?

This entire project – step-by-step, production-ready, with diagrams, videos and code – is covered inside our **Master AI Deployment** course.

 Got questions or want to join the next cohort?

Reach out anytime – always happy to help engineers level up! 