Ollama

Overview

Ollama is the local large language model (LLM) inference backend for the homelab. It serves models via an OpenAI-compatible REST API and is consumed by Open Web UI, PaperlessAI, and any other service that needs LLM inference without sending data to external providers.

Ollama runs with full NVIDIA RTX 5080 acceleration via the `nvidia` container runtime. All model weights are stored on the local NVMe system drive.

Access

Type	URL	Notes
Internal	`https://ollama.home.local`	Traefik-proxied HTTPS
Direct	`http://192.168.1.85:11434`	Raw API (no TLS)

No external (internet-facing) route — LAN and Tailscale access only.

Configuration

Image: `ollama/ollama:latest` **Compose project:** `ai-stack` **Runtime:** `nvidia`

Ports

Port	Protocol	Purpose
`11434`	TCP	Ollama REST API (host-bound)

Traefik Labels

```
traefik.enable: "true"
traefik.docker.network: traefik-net
traefik.http.routers.ollama.rule: Host(`ollama.home.local`)
traefik.http.routers.ollama.entrypoints: websecure
traefik.http.routers.ollama.tls.certresolver: step-ca
traefik.http.services.ollama.loadbalancer.server.port: 11434
```

Internal-only route, no authentication middleware — API access is unrestricted on the LAN. Callers must be on the LAN or Tailscale.

Environment Variables

Variable	Value	Purpose
`OLLAMA_HOST`	`0.0.0.0`	Listen on all interfaces
`NVIDIA_VISIBLE_DEVICES`	`all`	Expose all NVIDIA GPUs to container
`NVIDIA_DRIVER_CAPABILITIES`	`compute,utility`	Required NVIDIA driver caps
`OLLAMA_NUM_GPU`	`999`	Use all available GPU layers
`no_proxy`	`localhost,127.0.0.1`	Bypass proxy for local calls

GPU Acceleration

Ollama uses the `nvidia` container runtime. The RTX 5080 provides 16 GB of VRAM, allowing large models (7B–27B parameter range) to run fully in VRAM without CPU offloading.

```
runtime: nvidia
environment:
  - NVIDIA_VISIBLE_DEVICES=all
  - NVIDIA_DRIVER_CAPABILITIES=compute,utility
```

Volumes / Bind Mounts

Host Path	Container Path	Purpose
`/home/jeeves/docker/ollama/`	`/root/.ollama`	Model weights and config

Model files are stored under `/home/jeeves/docker/ollama/models/` on the local NVMe system drive (1.8 TB). Large models can consume significant space.

Networks

Network	Purpose
`ai-stack_ai-internal`	Internal communication with Open Web UI, PaperlessAI
`traefik-net`	Exposes Ollama API via Traefik

Dependencies

- NVIDIA container runtime (`nvidia`) on the Docker daemon
- RTX 5080 connected via OcuLink (must be recognised as a CUDA device)

Notes / Gotchas

- `OLLAMA_NUM_GPU=999` is the conventional way to tell Ollama to use as many GPU layers as possible. It does not literally use 999 GPUs.
- Models are downloaded via `ollama pull <model>` or via the Open Web UI admin panel. Downloaded models persist in the bind-mounted `/root/.ollama` directory.
- If the OcuLink connection drops or the GPU is not recognised, Ollama falls back to CPU inference — responses will be significantly slower. Check with:

  ```
  docker exec -it ollama ollama ps
  ```

- The local file at `/home/jeeves/docker/ai-stack/ollama-intel-arc/docker-compose.yml` is the legacy IPEX-LLM config and is **no longer in use**. The current stack is managed via Portainer and uses `ollama/ollama:latest` with CUDA.

Last Updated: 2026-06-16

Revision #4
Created 2026-06-17 04:43:10 UTC by Admin
Updated 2026-06-17 13:41:44 UTC by Admin