08-minuspod.md

kstack: book: Centerpoint Home Lab chapter: Utilities page: MinusPod tags: [minuspod, podcasts, transcription, whisper, ai, cuda]

Overview

MinusPod is an AI-powered podcast manager and transcriber. It manages podcast feeds, downloads episodes, and transcribes audio using a local Whisper model. Summaries and notes are generated via an LLM (Qwen3:14b) through the Ollama API. The container runs with NVIDIA GPU access (RTX 5080) for accelerated Whisper inference. Internal-only access. Episode data is retained for 180 days.

Access

Type	URL	Auth
Internal	`https://minuspod.home.local`	MinusPod own auth (Step-CA TLS)

No external route — LAN access only.

Configuration

Image: `ttlequals0/minuspod:2.1.9`
Runtime: CUDA 12.9 / NVIDIA runtime (RTX 5080)

Key Environment Variables

Variable	Value / Notes
`BASE_URL`	`https://minuspod.home.local`
`WHISPER_BACKEND`	`local`
`WHISPER_DEVICE`	`cuda`
`WHISPER_MODEL`	`medium`
`LLM_PROVIDER`	`ollama`
`OPENAI_BASE_URL`	`https://ollama.home.local/v1`
`OPENAI_MODEL`	`qwen3:14b`
`OPENAI_API_KEY`	**REDACTED**
`RETENTION_PERIOD`	`4320` hours (180 days)
`MINUSPOD_TRUSTED_PROXY_COUNT`	`1`
`NVIDIA_VISIBLE_DEVICES`	`all`
`NVIDIA_DRIVER_CAPABILITIES`	`compute,utility`

Traefik Labels

```
traefik.http.routers.minuspod.rule: Host(`minuspod.home.local`)
traefik.http.routers.minuspod.entrypoints: websecure
traefik.http.routers.minuspod.tls.certresolver: step-ca
traefik.http.services.minuspod.loadbalancer.server.port: 8000
```

Volumes / Bind Mounts

Host Path	Container Path	Purpose
`/home/jeeves/dockers/minuspod/data`	`/app/data`	Episode database and downloads

> Note: The data directory is under `/home/jeeves/dockers/` (not `docker/`) — this is the actual bind mount path as confirmed by `docker inspect`.

AI Integration

MinusPod uses two AI components:

Component	Model	Transport	Purpose
Whisper	`medium`	Local CUDA inference (RTX 5080)	Audio transcription
Qwen3:14b	`qwen3:14b`	Ollama API at `ollama.home.local/v1`	Summaries and notes

`OPENAI_BASE_URL` points to the internal Ollama instance (which exposes an OpenAI-compatible API). The `OPENAI_MODEL` is `qwen3:14b` — this model must be present in Ollama. Verify with `docker exec ollama ollama list`.

Notes / Gotchas

- The Whisper `medium` model on CUDA provides a good balance of speed and accuracy. Upgrade to `large-v3` if accuracy is more important than speed (uses more VRAM). The RTX 5080 has 16GB GDDR7 — ample for either.
- `RETENTION_PERIOD=4320` hours = 180 days. Episodes older than this are automatically removed. Adjust if long-term episode storage is needed.
- `MINUSPOD_TRUSTED_PROXY_COUNT=1` tells MinusPod to trust one level of proxy headers (Traefik) for correct IP and protocol detection.
- The bind mount path is `/home/jeeves/dockers/minuspod/data` (note: `dockers`, not `docker`). This is different from the convention used by most other stacks.
- Ensure `qwen3:14b` is loaded in Ollama before enabling LLM summaries:

  ```
  docker exec ollama ollama pull qwen3:14b
  ```

Last Updated: 2026-06-17

Revision #2
Created 2026-06-17 04:46:04 UTC by Admin
Updated 2026-06-17 04:46:04 UTC by Admin