

Rapport — Gestion des modèles, architecture d'exécution et agents

- **Date** : 2026-09-12
- **Périmètre** : poste local (Kali One) · kvm8 (72.62.26.63) · kvm4 (187.124.39.137)
- **Objet** : qui fait quoi, où tournent les tâches 24/7, comment sont gérés les LLM et les agents

1. Résumé exécutif

- **OmniRoute (kvm8)** est le **registre unique des modèles** : providers, clés, quotas, aliases, cache. Tous les clients (dsh, chat-multi, minisite, n8n...) passent par `https://dsh.assemblee-citoyenne.net/omni-api/v1`.
- **kvm8** porte la **passerelle + l'IA lourde** (dsh, éthique locale, audio, vidéo).
- **kvm4** porte le **web public** (SPIP, boutique) + **les LLM locaux Ollama** + un **proxy éthique Go**.
- **Le poste local ne doit porter aucune tâche 24/7** : il sert au dev, aux tests et aux sessions.
- **Règle de délégation** : toute tâche continue ou planifiée est déléguée à kvm8 ou kvm4 via `systemd` OU `cron`.

2. Cartographie des machines

2.1 Poste local — Kali One (dev, non 24/7)

Service	Rôle
<code>ollama.service</code>	modèles de code locaux (qwen-coder-7b/14b)
OmniRoute local (127.0.0.1:20128)	providers <code>openai</code> , <code>b.ai</code> (x2), <code>searxng-search</code> — usage dev/test

2.2 kvm8 — Passerelle IA centrale + services lourds (Ubuntu 24.04, 8 vCPU, 31 Go)

Service	Port	Rôle
<code>omniroute</code>	20128	gateway central (providers/clés/quotas/catalogue)
<code>dsh-web</code> (user)	3080	UI DeepSeek Harness, branchée sur OmniRoute (<code>ds/deepseek-flash</code>)
<code>chat-multi</code>	8090	MultiChat OmniRoute (115 modèles)
<code>webttyd</code> / <code>ttyd</code>	8091 / 7681	terminal web
<code>deepseek-v4-ethique</code>	8081	LLM local GGUF DeepSeek-V4-Flash abliterated, API OpenAI, 3 lois
<code>minisite-audio</code>	8110	génération de chansons (DiffRhythm2 , CPU)
<code>minisite-video</code>	8111	avatars chantants (Wav2Lip)
<code>ollama</code>	11434	modèles locaux
<code>nginx</code>	80/443	façade <code>dsh.assemblee-citoyenne.net</code>
Docker	—	Loomio + PostgreSQL
<code>cloudflared</code> , <code>fail2ban</code>	—	tunnel / anti-bruteforce

Service	Port	Rôle
Caddy 2.11.4	—	installé, désactivé (préparation sous-domaines)

2.3 kvm4 — Web public + LLM locaux (Kali, 4 vCPU, 15 Go)

Service	Rôle
Apache2 + MariaDB	SPIP assemblees-citoyennes.org, boutique Thelia , cooperatives, copycoop, moret...
Minisite musique-42962	chat.php, test-llm.php, api-test-llm.php → appelle OmniRoute kvm8
omniroute	instance locale (non utilisée comme passerelle de référence)
ollama	gemma2, qwen2.5, qwen3.5-ethic, qwen3.5-deepseek-v4-ethic, deepseek-coder-v2, deepseek-r1, hermes3
deep-ethical	proxy éthique Go (:11436 → Ollama), 3 lois + mémoire (ethical_memory.json)
tor, fail2ban, clamav, Docker	sécurité / services

3. Gestion des modèles

3.1 Registre central = OmniRoute (kvm8)

- **Providers actifs** : deepseek (officiel), bai (gratuit), elevenlabs (audio). Inactifs : openai, opencode.
- **Nommage** : provider/modèle → ds/deepseek-flash, ds/deepseek-v4-pro, bai/glm-5.3-flash...
- **Cycle de vie** :
 - synchroniser le catalogue : POST /api/providers/{id}/sync-models
 - ajouter un modèle **persistant** (survit aux synchros) : POST /api/provider-models → source:"manual"
 - supprimer un modèle : DELETE /api/provider-models?provider=...&model=...
- **Décision actuelle** : **une seule lignée Flash** = ds/deepseek-flash (**V4.1 Flash, texte + image**) ; anciennes versions retirées.

3.2 Couches de modèles

Couche	Emplacement	Exemples	Coût
Officiels DeepSeek	kvm8 OmniRoute	ds/deepseek-flash, ds/deepseek-v4-pro	payant
Gratuits b.ai	kvm8 OmniRoute	bai/glm-5.3-flash, qwen3.8-flash, hy3, mimo-v2.5	0
Éthique local	kvm8 :8081 + kvm4 :11436	GGUF DeepSeek-V4-Flash abliterated, qwen3.5-deepseek-v4-ethic	0 (CPU/GPU local)
Audio	kvm8 OmniRoute	ElevenLabs	payant
Audio/Vidéo local	kvm8 :8110 / :8111	DiffRhythm2, Wav2Lip	0
Code local	Kali One / kvm4 Ollama	qwen-coder-*, deepseek-coder-v2	0

4. Qui fait quoi (config → rôle)

Config	Machine	Ce qu'elle gouverne
/root/.omniroute/.env + storage.sqlite	kvm8	providers, clés, catalogue, quotas, cache
/root/.dsh/settings.yaml + .credentials.yaml	kvm8	modèle par défaut de dsh (omniroute / ds/ deepseek-flash)
~/.config/systemd/user/dsh-web.service	kvm8	UI dsh (port 3080)
/etc/systemd/system/*.service	kvm8	deepseek-v4-éthique, minisite-audio/video, chat-multi, webttyd
/root/Production-vps8/maintenance.sh	kvm8	backup Loomio + scan logs
/root/Production-vps8/night-agent.sh	kvm8	agent de maintenance nocturne (DSH headless)
/etc/caddy/Caddyfile.prod-draft	kvm8	façade sous-domaines (en attente)
/etc/nginx/sites-available/dsh	kvm8	façade actuelle basic-auth + routes
/root/.omniroute/ (local)	kvm4	gateway local (secondaire)
/opt/deep-kali-ia-pro/laws/	kvm4	proxy éthique 3 lois (Go)
/root/scripts/healthcheck.sh, backup-spip.sh	kvm4	santé + backup SPIP
/var/www/html/assemblees-citoyennes-clean/musique-42962/*.php	kvm4	minisite (chat, test LLM, API)
~/dsh-workspace/	local	workspace projet, docs, scripts de déploiement

5. Tâches 24/7 et délégation

Tâche	Machine	Déclencheur	Fréquence
Passerelle OmniRoute	kvm8	systemd	continu
UI dsh + agents	kvm8	systemd user	continu
LLM éthique local	kvm8	systemd	continu
Audio / Vidéo IA	kvm8	systemd	continu
Loomio + DB	kvm8	Docker	continu
Agent de nuit (rapport maintenance)	kvm8	cron 0 2 * * *	quotidien
Backup Loomio	kvm8	cron 30 3 * * *	quotidien
Site SPIP + boutique	kvm4	Apache/MariaDB	continu
Broadcast éthique Ollama	kvm4	systemd	continu
Healthcheck minisite	kvm4	cron */30	30 min
Backup SPIP	kvm4	cron 0 3 * * *	quotidien

Principe : kvm8 = **cerveau IA** ; kvm4 = **vitrine + LLM locaux** ; local = **atelier**. Aucune dépendance à la machine de bureau pour un service continu.

6. LLM spécialisés et agents

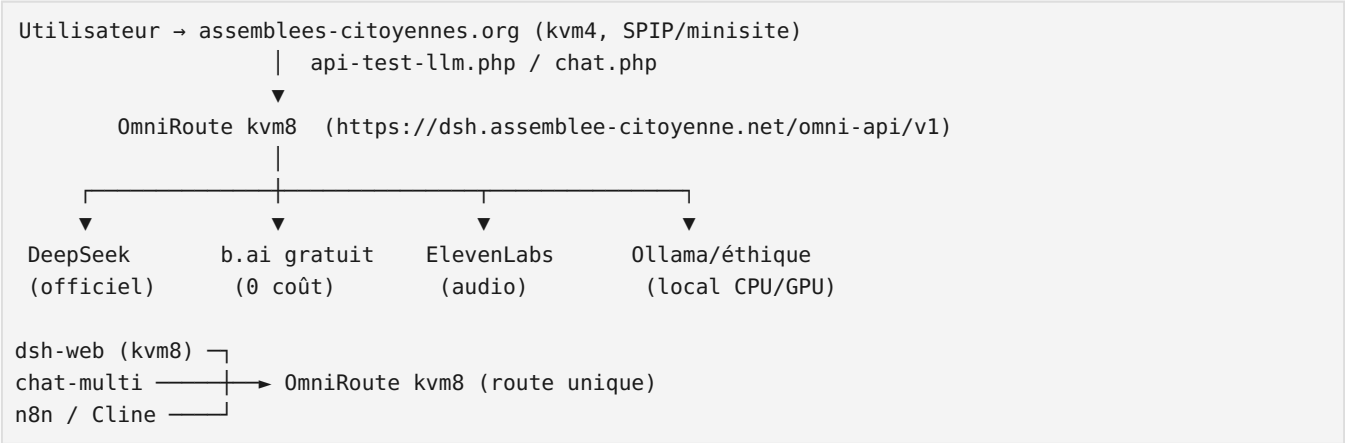
6.1 LLM spécialisés

Usage	Modèle	Où
Général, vision, rapide	ds/deepseek-flash (V4.1 Flash)	kvm8 OmniRoute
Raisonnement lourd / code	ds/deepseek-v4-pro	kvm8 OmniRoute
Gratuit polyvalent	bai/glm-5.3-flash, qwen3.8-flash, hy3, mimo-v2.5	kvm8 OmniRoute
Éthique 3 lois (API)	DeepSeek-V4-Flash abliterated (GGUF)	kvm8 :8081
Éthique 3 lois (proxy + mémoire)	qwen3.5-deepseek-v4-ethic	kvm4 :11436
Code local	qwen-coder-7b/14b	Kali One
Code / raisonnement local	deepseek-coder-v2, deepseek-r1, hermes3	kvm4 Ollama
Audio / voix	ElevenLabs	kvm8 OmniRoute
Chansons	DiffRhythm2	kvm8 :8110
Avatars chantants	Wav2Lip	kvm8 :8111

6.2 Agents

Agent	Machine	Rôle
Agent de nuit (night-agent.sh → dsh -- profile headless)	kvm8	vérifie services/disque/backups/ports, rédige un rapport, pré-vol éthique 3 lois
Maintenance (maintenance.sh)	kvm8	santé conteneurs Loomio, pg_dump, scan logs, purge, espace disque
Healthcheck minisite	kvm4	contrôle 11 pages + APIs coop
Backup SPIP	kvm4	dump MariaDB + fichiers
Pipeline clients (gestion-spip-clients.php)	kvm4	inscrits COP + commandes Thelia → comptes SPIP (textes via OmniRoute)
Services de génération (audio/vidéo)	kvm8	jobs à la demande déclenchés par le minisite
Sessions humaines (opencode/dsh)	local	dev, déploiement, audits

7. Flux



8. Écarts constatés / actions recommandées

1. `night-agent.sh` **utilise encore** `DEEPSEEK_API_KEY` (clé directe) alors que dsh route désormais via OmniRoute : le garde-fou sur cette variable est **obsolète**. → Aligner le script sur `OMNIROUTE_API_KEY` et retirer la clé directe.
 2. **kvm4 n'expose pas ses LLM locaux (Ollama + éthique) à OmniRoute kvm8** : ils restent consommables seulement en local. → Envisager un *provider node* OmniRoute vers kvm4 (11436/11434) pour centraliser.
 3. **DNS** : créer les A records (`omniroute`, `multi`, `terminal`, `api` → 72.62.26.63) avant la bascule Caddy.
 4. **Bascule nginx → Caddy** : préparée (`Caddyfile.prod-draft` validé), en attente de décision A/B (bloc par sous-domaine ou wildcard) et du token Hostinger.
 5. **Cycle de vie modèles** : documenté (`sync-models`, `provider-models`); garder **une seule lignée Flash**.
-

9. Emplacements

- Déploiement kvm8 : `/root/Production-vps8/`
- Workspace local : `/home/c3b/dsh-workspace/`
- Documentation pro : `DOC-INSTALLATION-PRO-OMNIROUTE-DEEPSEEK-V41-20260912.md`
- Bundle de duplication : `/root/Production-vps8/backups/omniroute-gateway-20260912-1553.tar.gz`