RÉPARLAB / SAMEDI 22 NOVEMBRE 2026 / 14H-17H / SALON DE THÉ « DIVERS SI THÉ », CUNLHAT — GRATUIT

ATELIER · DÉMONSTRATION · ON DÉBRANCHE LE WI-FI

L'IA libre, sur place ou en livraison

Handy · LFM3-VL · DeepSeek · Mistral · OpenCode
Trois façons de se servir, sans dépendre d'un géant.

Des modèles libres, deux endroits pour les faire tourner, et une seule question qui compte : **où vont tes données ?** On teste tout en vrai, avec un portable et une connexion qu'on peut débrancher.

1 Handy

LA DICTÉE QUI RESTE À TABLE

Un raccourci, tu parles, tu relâches : le texte s'écrit dans le champ où tu es. Aucun son ne quitte la machine, même sans Wi-Fi.

`MIT` `v0.10.1` `Parakeet Ultra` `CPU`

2 LFM3-VL

L'ŒIL DU JOUR

Deux milliards de paramètres qui regardent une image et répondent en français. Il tourne ici, dans l'interface maison, sur mon PC.

`2B` `Q4_K_M` `128k` `17 langues`

3 DeepSeek + Mistral

LA LIVRAISON, DEUX CONTINENTS

DeepSeek V4.1 Flash — 552 Md de paramètres, MIT, ~0,30 $/M.

Mistral Large 4 — 1 050 Md, licence ouverte (MIT modifiée), 0,70 $/M, entraîné en Europe.

`1M ctx` `Ollama Cloud` `US · FR`

Le comptoir : OpenCode

Un seul terminal pour les deux : le petit modèle en local sur **127.0.0.1:11434**, le gros du cloud sur **ollama.com**. Même conversation, on change juste de fournisseur — le local regarde la capture d'écran, celui du cloud écrit le correctif.

```
ollama launch opencode \
--model deepseek-v4.1-flash:cloud
```

Libre ≠ local (et c'est très bien)

La licence dit ce que tu peux faire du modèle ; l'hébergement dit où vont tes données. Deux choix séparés : un modèle libre peut tourner chez toi **ou** chez un hébergeur. DeepSeek V4.1 Flash est MIT et tourne chez Ollama ; LFM3-VL tourne ici, poids ouverts. Les deux sont libres — aucun n'est « le vrai ». Et son petit frère, **Ministral 3B**, est Apache 2.0 pour 3 milliards de paramètres.

Concrètement : ta photo de famille reste chez toi ; ta question de code peut partir chez Ollama.

⚠ `ollama run modele:cloud` envoie la requête aux États-Unis sans prévenir. `OLLAMA_NO_CLOUD=1` coupe net.

	CHEZ TOI	CLOUD LIBRE	CLOUD FERMÉ
Où ça tourne	ton PC	serveurs Ollama	ChatGPT & co
Tes données	**restent**	partent*	partent
Coût	**0 €**	~0,30 $/M	abonnement
Sans Internet	**oui**	non	non
Licence	MIT · poids ouverts	MIT	propriétaire



1 on charge une image · 2 on demande en français · 3 on lit la réponse et le temps que ça prend. Ici : `34,3 s` pour décrire l'affiche et `12 tok/s` en décodage, sur un `1,7B Q4_K_M`. *Capture : interface maison, 19/11/2026. *zéro rétention annoncée par Ollama.*

Ce soir, chez vous

1. Installe Handy — 5 minutes.
2. Un modèle de 2 Go, via **Ollama**.
3. Coupe le Wi-Fi et essaie.

LA SÉANCE, MINUTE PAR MINUTE

Cinq manches, 30 minutes

Six minutes par manche, montre en main. Seul le duel a besoin du réseau — c'est justement l'expérience.

16h30 — L'ŒIL

L'image mystère

Une photo que personne dans la salle n'a vue. Que voit-il ? Que rate-t-il ? On compare avec vos réponses, à main levée.

16h36 — L'ŒIL

Lire une étiquette

Boîte de médicament, ordonnance, facture : il lit à voix haute, on vérifie chaque chiffre. C'est là qu'on voit les limites.

16h42 — LE DUEL

Le duel, au comptoir

Même capture d'écran, même question. OpenCode interroge le 2B en local, puis bascule sur les 552 Md du cloud. On chronomètre, on compare, on regarde la facture.

16h48 — LA VOIX

Handy, en direct

Je dicte la fin de cette fiche pendant que vous la lisez. Ctrl + Alt + Espace, on maintient, on parle, on relâche — le texte arrive.

16h54 — LES RATÉS

Les ratés, assumés

Compter, épeler un nom propre, inventer sans prévenir : on regarde les erreurs en face. OpenCode demande 64k de contexte minimum — c'est là que le matériel compte.

Le filet de sécurité

Modèle préchargé avant la séance, captures d'écran de secours pour chaque manche. Si ça plante, on ouvre le terminal ensemble : le libre, c'est aussi pouvoir regarder sous le capot.

Deux endpoints, un seul comptoir

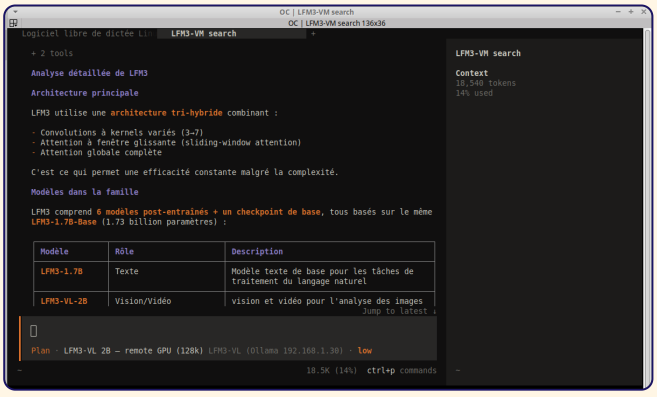
`local → 127.0.0.1:11434`

`cloud → ollama.com`

On change de fournisseur, pas de conversation.



Handy — Ctrl + Alt + Espace, mode « Maintenir », Parakeet Ultra en français. Rien ne sort de la machine.



OpenCode — même terminal, deux fournisseurs. Ici : LFM3-VL servi par Ollama en local, 18 540 tokens en contexte.

~275×
plus de paramètres dans le nuage que dans le portable

≈ 4,7 %
d'erreur en français avec Parakeet Ultra, ~13× le temps réel sur un CPU de portable

0
octet d'audio envoyé au cloud par Handy

0 €
la démo locale, hors électricité

handy.computer · opencode.ai · docs.liquid.ai · ollama.com
chiffres : benchmarks FLEURS Q8_0 publiés par Handy · tarifs Ollama Cloud au 19/11/2026

contact@libratoi.fr · libratoi.fr