

Guide de l'IA Locale pour les Entreprises

Matériel, modèles, outils et premiers pas : tout ce qu'il faut pour introduire l'IA dans votre activité — sans cloud et sans partager les données. Avec des prix réels et des commandes prêtes à copier.

Sommaire

1. Qu'est-ce que l'IA locale (et ce que ce n'est pas)
2. Les trois questions à poser avant d'acheter
3. Matériel : comment lire les spécifications
4. Modèles et quantification : comment choisir
5. Outils : Ollama, Open WebUI, vLLM
6. Premiers pas : 10 minutes, deux commandes
7. Les limites : quand l'IA locale ne suffit pas
8. Où aller plus loin

À qui s'adresse ce guide : à celles et ceux qui ont une activité, une équipe ou un projet et veulent savoir si l'IA locale leur convient — sans budget IT dédié et sans jargon inexpliqué. Si vous cherchez l'installation pas à pas de chaque outil, les liens en fin de guide pointent vers la documentation complète.

Qu'est-ce que l'IA locale (et ce que ce n'est pas)

L'IA locale, c'est un modèle de langage qui tourne sur **votre** ordinateur ou votre serveur — pas sur celui d'un autre. Au lieu d'envoyer votre question vers un cloud, le calcul a lieu sur votre GPU (ou la mémoire de votre Mac), et la réponse revient à vous sans qu'aucun octet ne quitte la pièce.

Une image : la différence entre louer et acheter

Utiliser un service cloud, c'est **louer** : vous payez à chaque utilisation, le prix peut changer, et le contrat stipule que vos données passent entre les mains du propriétaire. L'IA locale, c'est **acheter** : le coût, c'est le matériel (une seule fois) et les données restent à vous. Qui a un budget illimité peut préférer la location ; qui a des coûts récurrents ou des données sensibles, presque toujours l'inverse.

Les trois avantages concrets

- **Contrôle des données.** Clients, factures, contrats : rien ne quitte votre infrastructure. Pas de log fournisseur sur lequel se poser, pas de données qui finissent dans l'entraînement d'autrui.
- **Coût prévisible.** Après l'achat du matériel, l'inférence coûte le courant. Un modèle au service de 10 personnes par jour sur un GPU à 800 euros coûte, en pratique, zéro par requête.
- **Disponibilité.** Ça marche quand internet tombe, au bureau, en usine, sur le chantier. Pas de limite de requêtes, pas de pic de prix, pas de dépendance à un fournisseur qui ferme son service.

La partie honnête : les modèles locaux n'égale pas les frontier models du cloud sur les tâches les plus difficiles (longs raisonnements, tâches scientifiques complexes). Et ils coûtent plus au départ. La bonne question n'est pas « quoi de mieux ? » mais « pour mon usage précis, qu'est-ce qui me coûte le moins et me convient le mieux ? ». C'est ce à quoi nous répondons ensemble dans les pages suivantes.

Les trois questions à poser avant d'acheter

Avant de dépenser un euro, répondez à trois questions. Ce sont celles que le planificateur de setup d'iaLocale utilise, et elles déterminent 90 % du choix.

Question 1 — Quelle tâche doit-elle accomplir ?

Résumés de documents, brouillons d'e-mails, recherche dans une base de documents, un agent qui appelle des outils : ce sont des tâches aux exigences différentes. Les résumés et le chat en demandent peu ; un agent qui raisonne en plusieurs étapes avec un long contexte en demande bien plus. La tâche définit la *taille du modèle*.

Question 2 — À quelle vitesse ?

Une réponse en 3 secondes et une en 15 secondes, c'est deux expériences de travail différentes. La vitesse dépend de la *bande de la mémoire* du matériel, pas de la puissance de calcul — c'est le point qui surprend tout le monde, et il est expliqué dans la section matériel.

Question 3 — Combien d'utilisateurs, en parallèle ?

Un modèle sert un utilisateur à la fois. Si cinq personnes écrivent en même temps, ou s'il faut le long contexte de toute la conversation, il faut plus de mémoire — ou une infrastructure qui gère la file.

La méthode : prenez un cas d'usage réel (le résumé d'un procès-verbal, le brouillon d'une offre, la recherche dans les manuels produit) et exprimez-le en chiffres : combien de texte entre, combien de mots sortent, en quel temps, combien de personnes l'utilisent. Avec ces chiffres, les sections 3 et 4 deviennent un choix, pas une hypothèse.

Un exemple concret

Une agence de 8 personnes qui veut : brouillons d'e-mails et de devis, recherche dans les contrats, un assistant en français. Profil : modèle 8–14B en quantification Q4, 16 Go de VRAM, bande 800+ Go/s. Résultat : un GPU RTX 4070 Ti Super (16 Go, ~800 €) couvre le cas avec de la marge. Le même budget en cloud serait un abonnement récurrent qui, à ces volumes, coûte plus que le matériel la première année.

Matériel : comment lire les spécifications

Pour l'IA locale, **deux chiffres** du matériel comptent : la VRAM (combien de mémoire a la carte) et la bande (à quelle vitesse la mémoire se lit). Tout le reste est secondaire.

VRAM : décide QUANT de modèle entre

Règle pratique : les poids d'un modèle occupent environ $\text{paramètres} \times \text{bits} / 8$ octets. Un modèle de 8 milliards de paramètres en Q4_K_M pèse ~4,8 Go de poids, plus la mémoire de travail (le contexte, la KV cache). Pour travailler avec de la marge : **modèle 8B → 12–16 Go ; modèle 14B → 16–24 Go ; modèle 30B → 24–32 Go**.

Bande : décide COMME C'EST RAPIDE

Les mots sortent à une vitesse proportionnelle à la bande de la mémoire, divisée par les octets du modèle. Une RTX 3060 (360 Go/s) sur un 8B Q4 fait ~18–22 mots par seconde : lisible. Une 4090 (1008 Go/s) en fait ~50 : très fluide. Deux cartes identiques en puissance mais avec une bande différente peuvent avoir des performances *différentes de moitié*.

Les trois profils, avec les prix de octobre 2026

Profil	Matériel	Prix indicatif	Ce que ça fait bien
Point de départ	RTX 4060 Ti 16 Go (ou RTX 3060 12 Go, 455 €)	455–930 €	Modèles jusqu'à 8B en Q4, 1–2 utilisateurs, tâches de résumés/chat
Bureau	RTX 4070 Ti Super 16 Go	~3.225 €	8–14B avec long contexte, plusieurs utilisateurs, vitesse fluide
Équipe	RTX 4090 24 Go / RTX 5090 32 Go	3.330–6.130 €	30B+ et agents en parallèle, très longs contextes, multi-utilisateurs

Et le Mac ? Un Mac avec puce Apple Silicon de 36 Go (M4 Max) héberge des modèles jusqu'à 32B avec élégance — la bande de la mémoire unifiée est élevée et la consommation modérée. Pour beaucoup d'équipes, c'est la voie la plus simple, sans assembler un PC. Les prix changent d'un mois à l'autre : le [planificateur de setup](#) met à jour les chiffres chaque jour.

CPU seul : si le budget est très serré, un CPU moderne avec 64 Go de RAM (bande ~40–90 Go/s) fait tourner un 8B à 5–8 mots par seconde : lent, mais ça marche pour les résumés non en temps réel. C'est le plan B, pas le plan A.

Modèles et quantification : comment choisir

Familles à connaître

- **Qwen** — la famille open avec le meilleur support multilingue aujourd'hui, français compris. C'est notre point de départ recommandé pour les contenus en français.
- **Llama** (Meta) — la famille généraliste la plus répandue ; 8B et 70B sont les références de taille.
- **Mistral** — efficaces et compacts, excellents quand la machine est petite.
- **Qwen3-30B-A3B** — modèle « mixte » (30B au total, 3B actifs) : qualité de grand modèle, vitesse de petit modèle. Le choix le plus intéressant pour le profil « équipe ».

Quantification : le compromis qualité/poids

La quantification compresse les poids du modèle de 16 bits (F16) à quelques bits, réduisant la VRAM et la taille avec une perte de qualité contrôlée. En pratique :

Quantification	Bits/poids	Qualité	Quand l'utiliser
Q4_K_M	4,8	≈ 85%	Le défaut : le meilleur compromis pour presque tous les cas
Q5_K_M	5,7	≈ 92%	Si la VRAM le permet et que la qualité compte
Q6_K	6,6	≈ 95%	Petits modèles sur gros matériel
Q2_K	3,5	≈ 55%	Seulement si ça ne rentre pas autrement : la qualité s'effondre

La méthode de choix en trois gestes

- **1 Partez de la VRAM disponible.** Retirez ~2–4 Go de mémoire de travail (selon le contexte) et divisez : c'est le poids maximal du modèle qui rentre.
- **2 Descendez de famille, pas de qualité.** Mieux vaut un 8B en Q5 qu'un 14B en Q2 : la qualité de la quantification pèse autant que la taille.
- **3 Testez avec votre tâche réelle.** Le benchmark général ne suffit pas : prenez 10 prompts de votre travail et comparez. C'est le seul test qui compte.

Outils : Ollama, Open WebUI, vLLM

Le modèle, ce ne sont que des poids. Pour le faire tourner il faut un moteur d'inférence, et pour lui faire parler une interface. Trois outils couvrent 95 % des cas.

Ollama — le moteur pour commencer

Une seule ligne d'installation, des modèles qui se téléchargent avec une commande, une API locale prête. C'est le point de départ : si vous devez « faire tourner quelque chose aujourd'hui », c'est Ollama. Il sert un modèle par utilisateur à la fois, ce qui est exactement ce qu'il faut pour 1–5 personnes.

Open WebUI — l'interface

Une interface web de type chat au-dessus d'Ollama (ou d'autres API) : conversations enregistrées, plusieurs utilisateurs, import de documents pour le RAG (poser des questions à vos documents). En 10 minutes de setup vous avez votre « ChatGPT maison », sur votre serveur.

vLLM — quand c'est de la production

Quand les modèles servent beaucoup d'utilisateurs en parallèle avec de longs contextes, vLLM fait du batching des requêtes et utilise la VRAM plus efficacement. C'est l'outil « serveur » : plus complexe à mettre en place, mais c'est la réponse quand Ollama commence à faire la queue.

La progression naturelle : Ollama (aujourd'hui) → Open WebUI (cette semaine) → vLLM (quand l'équipe grandit). Ne partez pas de vLLM : la plupart des activités ne le rencontrent jamais, et celles qui le rencontrent le font avec de l'expérience.

Outil	Setup	Utilisateurs	Quand
Ollama	~5 min	1–5	Toujours le point de départ
Open WebUI	~10 min	2–15	Quand il faut une interface d'équipe
vLLM	heures	10+	Quand les requêtes s'accumulent en file

Premiers pas : 10 minutes, deux commandes

Sur un Mac ou un Linux (et sur Windows, via WSL2), le parcours de zéro à la première réponse :

```
# 1. Installez Ollama
curl -fsSL https://ollama.com/install.sh | sh

# 2. Téléchargez un modèle 8B (Q4_K_M, ~4,9 Go)
ollama pull qwen2.5:7b

# 3. Parlez
ollama run qwen2.5:7b "Résume en 3 phrases : <votre texte>"
```

Sur Mac, l'installation est un .dmg ; le modèle se choisit avec la même syntaxe et tourne sur la mémoire unifiée. Sur un PC NVIDIA, Ollama détecte le GPU tout seul.

Ensuite, dans cet ordre

- **1 Définissez le cas d'usage.** Un seul. Écrivez les 10 prompts réels que vous lui donneriez.
- **2 Mesurez.** Vitesse (mots/seconde), qualité (sur ces 10 prompts), mémoire occupée. Notez les chiffres.
- **3 Comparez avec le planificateur.** Sur le [planificateur de setup](#), saisissez activité, utilisateurs et vitesse : il vous donne le matériel minimum et les prix actuels, et les compare à vos chiffres.
- **4 Ajoutez Open WebUI** quand le second collègue veut essayer. De là, c'est une équipe.

Attention au contexte : si vous devez coller des documents entiers, vérifiez combien de tokens entrent (le 8B de Qwen2.5 arrive à 128K de contexte théorique, mais c'est la VRAM qui décide du *pratique*). Pour les documents : le RAG, sujet du numéro le plus récent de la newsletter d'iaLocale.

Les limites : quand l'IA locale ne suffit pas

Soyons honnêtes, c'est le seul moyen de bien choisir.

- **Écart frontier.** Pour les tâches de raisonnement les plus profondes, les plus grands modèles cloud restent devant. Si votre cas est celui-ci, la réponse est hybride : du local pour 90 % des volumes, du cloud pour les 10 % difficiles.
- **Le premier coût est réel.** 300–800 euros de matériel, ce n'est rien, mais ce n'est pas zéro. En dessous d'un certain volume d'usage, le cloud reste moins cher.
- **Les documents longs ont un prix.** Chaque mot de contexte occupe de la VRAM. Les très longs contextes (des centaines de pages) exigent du vrai matériel ou des techniques de RAG.
- **Les modèles changent.** Le meilleur 8B en français d'aujourd'hui peut être un autre dans 6 mois. La bonne nouvelle : avec une infrastructure locale, changer de modèle est une commande, pas un projet.

La règle pratique : si vos données sortent déjà du cloud sans problème et que les volumes sont faibles, partez avec un service managé et gardez ce guide comme feuille de route. Aucun fournisseur en particulier : le choix dépend du budget, de la confidentialité et du type de modèle — le [catalogue des services d'inférence](#) rassemble les principaux, avec les détails de chacun. Si les données sont sensibles, que les volumes grandissent ou que la disponibilité compte, l'ordre s'inverse : la voie locale n'est plus une option.

Et quand il faut plus ?

L'échelle existe : de la 4090 au serveur multi-GPU, en passant par les Mac 128 Go. Le planificateur de setup montre l'échelle complète avec les prix, pour que la croissance soit un plan, pas un pari.

Où aller plus loin

Ce guide, c'est le « quoi » et le « pourquoi ». Le « comment » dans le détail est déjà sur le site :

- **Le planificateur de setup** — saisissez activité, utilisateurs, vitesse : vous obtenez le matériel minimum, les prix mis à jour chaque jour et les liens vers les sources. </fr/strumenti/setup>
- **5 cas d'usage pratiques pour votre activité** — les cas concrets avec chiffres et flux (article en italien). </it/blog/ia-locale-attivita-5-casi-duso-pratici>
- **RAG : posez des questions à vos documents** — la technique pour les longs contextes (article en italien). </it/blog/rag-fai-domande-ai-tuoi-documenti>
- **GPU ou CPU : quand la CPU suffit** — la règle des 40 Go/s (article en italien). </it/blog/gpu-o-cpu-quando-la-cpu-basta>
- **La newsletter** — chaque numéro est un pas, avec l'article complet. </fr/newsletter> (archives : </fr/newsletter/archivio>)
- **La page entreprises** — le même raisonnement, avec tableaux et coûts à jour. </fr/aziende>

La méthode, en une phrase : partez d'un cas d'usage réel, mesurez avec vos prompts, comparez avec les chiffres du planificateur et ne montez en gamme que quand les chiffres l'exigent. C'est toute la philosophie d'iaLocale : l'IA locale n'est pas un choix idéologique, c'est un calcul — et les calculs se font.

Guide de l'IA Locale pour les Entreprises · 1re édition, septembre 2026 · iaLocale Team

Prix de référence : marché européen, octobre 2026 — ils varient dans le temps ; vérifiez toujours le planificateur de setup pour les valeurs courantes. Dernière mise à jour : 26 octobre 2026.

Document généré avec IA.