GUIDE PRATIQUE

L'IA en local

Faire tourner une intelligence artificielle
sur son propre ordinateur, sans internet

Pour débutants | Pour techniciens | Procédure pas à pas

Par Jawed Tahir — Mai 2026

javed.fr

Document de référence — mai 2026

Avantages · Inconvénients · Limites · Installation

Sommaire

COMMENT LIRE CE DOCUMENT

Le texte courant est rédigé pour **tout le monde**, sans prérequis. Les encadrés bleus marqués « Niveau technique » apportent des précisions destinées aux lecteurs plus avancés : vous pouvez les **sauter entièrement** sans rien perdre de l'essentiel. Les encadrés verts sont des conseils pratiques, les encadrés orange des points de vigilance.

1. Comprendre l'« IA offline »

> « IA offline », « IA locale », « IA en local » désignent toutes la même chose : une intelligence artificielle qui fonctionne **directement sur votre appareil**, sans passer par internet ni par les serveurs d'une entreprise.

1.1 IA dans le cloud contre IA locale

Quand vous utilisez ChatGPT, Claude ou Gemini par leur site web, vous utilisez une IA **dans le cloud**. Votre question part par internet vers un centre de données distant, y est traitée par des ordinateurs très puissants, et la réponse vous revient. Vous ne possédez rien : vous vous connectez à un service.

Une IA **locale** fonctionne sur le principe inverse. Le « cerveau » de l'IA — un fichier appelé *modèle* — est téléchargé une fois pour toutes sur votre ordinateur. Ensuite, tout se passe chez vous : votre question est traitée par votre propre machine, la réponse est produite sur place. Une fois le modèle téléchargé, **vous pouvez débrancher internet** et l'IA continue de fonctionner.

L'image la plus parlante : l'IA cloud, c'est aller au restaurant ; l'IA locale, c'est cuisiner chez soi. Le restaurant offre une carte plus vaste et zéro effort, mais vous dépendez de ses horaires, de ses prix et de sa cuisine. Cuisiner chez soi demande un équipement et un apprentissage, mais vous maîtrisez tout, à toute heure, et personne ne sait ce que vous mangez.

1.2 Le vocabulaire essentiel

Quelques mots reviennent en permanence dès qu'on parle d'IA locale. Les voici, expliqués simplement :

Terme	Ce que ça veut dire
Modèle	Le fichier qui contient l'intelligence de l'IA. C'est ce que l'on télécharge. Différents modèles ont des « personnalités » et des compétences différentes.
Paramètres	L'unité qui mesure la « taille du cerveau » d'un modèle, comptée en milliards (le « B » de l'anglais *billion*). Un modèle « 7B » a 7 milliards de paramètres. Plus il y en a, plus le modèle est capable… mais plus il est lourd et lent.
Poids	Synonyme courant du contenu du modèle. Un modèle « à poids ouverts » (*open-weight*) est un modèle dont le fichier est librement téléchargeable.
Inférence	Le mot savant pour « l'IA réfléchit et produit une réponse ». Quand on dit qu'un modèle est lent, c'est l'inférence qui est lente.
Token	L'unité de base que l'IA manipule : un petit morceau de mot. La vitesse d'une IA locale se mesure en « tokens par seconde ».
Quantification	Une technique de compression qui allège un modèle pour qu'il tienne sur un ordinateur ordinaire, en échange d'une légère perte de qualité. Voir l'encadré ci-dessous.

NIVEAU TECHNIQUE — LA QUANTIFICATION EN DÉTAIL

Un modèle est, mathématiquement, une immense collection de nombres. À pleine précision, chaque paramètre occupe 16 bits (format `FP16`). La quantification réduit cette précision : `Q8` (8 bits), `Q5`, `Q4` … Un modèle 7B en `FP16` pèse environ 14 Go ; le même en `Q4_K_M` tombe autour de 4 Go, soit un facteur ~3,5.

Le compromis recommandé pour la quasi-totalité des usages grand public est **`Q4_K_M`** : la perte de qualité est à peine perceptible, alors que le gain en mémoire et en vitesse est décisif. Les niveaux `Q5_K_M` et `Q6_K` sont un cran au-dessus en qualité si votre matériel le permet ; en dessous de `Q3`, la dégradation devient nette. Les outils grand public choisissent `Q4_K_M` par défaut — vous n'avez en général rien à régler.

2. Pourquoi le faire — les avantages

> L'IA locale n'est pas un gadget de passionnés. Elle répond à des besoins concrets que le cloud ne peut pas satisfaire.

2.1 Confidentialité totale

C'est l'avantage numéro un. Avec une IA locale, **vos données ne quittent jamais votre machine**. Vos questions, vos documents, vos brouillons, vos informations sensibles ne sont transmis à personne, ne sont stockés sur aucun serveur, ne servent à entraîner aucun futur modèle. Pour un avocat, un médecin, un journaliste, un chercheur, une entreprise manipulant des secrets industriels, ce point peut à lui seul justifier toute la démarche.

2.2 Gratuité après l'installation

Les outils et la grande majorité des modèles sont **gratuits**. Pas d'abonnement mensuel, pas de facturation à l'usage. Le seul coût réel est celui de l'électricité consommée et, éventuellement, du matériel si vous décidez de l'améliorer. Vous pouvez poser un million de questions sans voir une facture grimper.

2.3 Fonctionnement sans connexion

Une fois le modèle téléchargé, l'IA fonctionne **hors ligne** : dans un train, un avion, une zone mal couverte, un site isolé, ou simplement en cas de panne internet. Elle ne dépend ni de votre fournisseur d'accès ni de la disponibilité d'un service distant.

2.4 Contrôle, stabilité et pérennité

Le modèle que vous avez téléchargé est **à vous**. Il ne changera pas du jour au lendemain, ne sera pas « mis à jour » d'une façon qui modifie ses réponses, ne sera pas retiré du service. Aucun fournisseur ne peut suspendre votre accès, modifier ses tarifs ou ses conditions d'utilisation. Vous décidez quand et si vous changez de modèle.

2.5 Liberté d'usage et personnalisation

Les modèles locaux sont en général moins bridés que les services grand public et acceptent une plus grande variété de tâches légitimes. Surtout, ils sont **personnalisables** : on peut ajuster leur ton, leurs instructions par défaut, les brancher sur ses propres documents, voire les spécialiser pour un métier.

> **NIVEAU TECHNIQUE — POURQUOI LES ENTREPRISES S'Y INTÉRESSENT**
>
> Au-delà de l'usage individuel, l'IA locale (ou auto-hébergée sur un serveur maîtrisé) répond à des contraintes de conformité : RGPD, secret professionnel, hébergement de données de santé, propriété intellectuelle. Elle évite la *fuite de données vers un tiers* et la dépendance à un fournisseur unique (*vendor lock-in*). L'API d'outils comme Ollama étant compatible avec celle d'OpenAI, une organisation peut migrer une application du cloud vers le local en ne changeant souvent que l'adresse du serveur, sans réécrire son code.

3. Les inconvénients, limites et idées fausses

> Un guide honnête doit présenter les contreparties. L'IA locale a de vraies limites — les connaître évite la déception et permet de bien choisir.

3.1 Une qualité en deçà des meilleurs modèles cloud

C'est la limite la plus importante à accepter. Un modèle qui tourne sur un ordinateur personnel est, par construction, bien plus petit que les modèles géants du cloud. Pour des raisonnements complexes, de la rédaction très fine ou des tâches difficiles, **la différence de qualité est réelle et perceptible**. Pour des usages courants — résumer, reformuler, répondre à des questions générales, brouillonner, traduire, aider au code simple — un bon modèle local est tout à fait satisfaisant.

3.2 Des exigences matérielles

Une IA locale sollicite fortement l'ordinateur, en particulier sa mémoire vive (RAM). Une machine ancienne ou peu dotée ne pourra faire tourner que de petits modèles, plus limités. La partie 4 détaille comment savoir ce dont vous disposez.

3.3 Des connaissances figées dans le temps

Un modèle local a été entraîné à une certaine date et **ne se met pas à jour tout seul**. Il ignore tout de ce qui s'est passé après son entraînement. Sans configuration supplémentaire, il **n'a pas accès à internet** : il ne peut pas vérifier une information, consulter l'actualité ni chercher une page web.

3.4 Pas d'outils intégrés

Les services cloud intègrent de nombreuses fonctions : recherche web, lecture de fichiers, génération d'images, exécution de code, mémoire entre conversations. Une IA locale, dans sa configuration de base, ne fait **qu'une chose : converser par texte**. Les autres capacités existent mais demandent une installation supplémentaire (voir partie 7).

3.5 Lenteur possible et effort initial

Selon le matériel et la taille du modèle, les réponses peuvent s'afficher plus lentement que dans le cloud — parfois mot à mot. Et il faut accepter une **mise en place initiale** : installer un logiciel, télécharger plusieurs gigaoctets, choisir un modèle. Rien d'insurmontable, mais ce n'est pas aussi immédiat que d'ouvrir un site web.

IDÉES FAUSSES À CORRIGER

« L'IA locale, c'est forcément moins bien. » Faux dans l'absolu : pour beaucoup de tâches quotidiennes, l'écart est mineur. C'est sur les tâches difficiles qu'il se creuse.

« Il faut être informaticien. » Plus vrai en 2026 : des outils comme LM Studio s'installent et s'utilisent comme une application ordinaire, sans la moindre ligne de commande.

« Il faut une carte graphique très chère. » Non : une carte graphique accélère, mais de petits modèles tournent correctement sur un ordinateur récent sans carte dédiée, et les Mac à puce Apple sont particulièrement à l'aise.

NIVEAU TECHNIQUE — LA NUANCE « CONNAISSANCES FIGÉES »

L'absence de mise à jour et d'accès web se contourne par une technique nommée **RAG** (*Retrieval-Augmented Generation*) : on fournit au modèle, au moment de la question, des documents pertinents qu'il utilise comme source. Le modèle ne « sait » toujours pas plus de choses, mais il peut *raisonner sur des documents à jour* qu'on lui donne. Des outils comme AnythingLLM ou Open WebUI intègrent le RAG clé en main (partie 7). De même, un module de recherche web peut être ajouté — mais attention : l'activer signifie que certaines requêtes repartent vers internet, ce qui entame l'avantage de confidentialité.

4. De quoi ai-je besoin ? Les prérequis matériels

> La question décisive est : « mon ordinateur en est-il capable, et jusqu'à quel point ? » Voici comment y répondre sans être technicien.

4.1 Le facteur numéro un : la mémoire vive (RAM)

La **RAM** est la mémoire de travail de l'ordinateur. C'est elle qui détermine la taille du modèle que vous pourrez faire tourner — bien plus que la puissance du processeur. Règle simple : le modèle, une fois chargé, doit tenir dans la RAM, en laissant de la place au reste.

Comment connaître ma RAM ?

- **Windows :** clic droit sur le menu Démarrer → *Système* → ligne « Mémoire RAM installée ».
- **Mac :** menu Pomme () → *À propos de ce Mac* → ligne « Mémoire ».
- **Linux :** commande `free -h` dans un terminal, ligne « Mem ».

4.2 Tableau de correspondance : RAM et modèles

Ce tableau donne les ordres de grandeur. Les tailles indiquées concernent des modèles quantifiés en `Q4`, le format standard.

RAM	Ce que vous pouvez viser	À quoi s'attendre
8 Go	Petits modèles : 1 à 4 milliards de paramètres (« 3B »)	Le minimum praticable. Bien pour résumer, reformuler, des questions simples. Réponses parfois approximatives.
16 Go	Modèles moyens : 7 à 9 milliards (« 7B », « 8B »)	Le bon point d'équilibre pour la plupart des gens. Qualité satisfaisante pour un usage quotidien varié.
32 Go	Grands modèles : 14 à 32 milliards	Qualité nettement meilleure, raisonnement plus solide. Confortable.
64 Go et +	Très grands modèles : 70 milliards et au-delà	On approche, sur certaines tâches, de la qualité des services cloud. Réservé aux machines bien équipées.

4.3 La carte graphique (GPU) : utile, pas indispensable

Une **carte graphique** dédiée accélère considérablement l'IA — les réponses arrivent beaucoup plus vite. Mais elle n'est **pas obligatoire** : sans elle, le modèle tourne sur le processeur, simplement plus lentement. Pour débuter, ce n'est pas un achat nécessaire.

4.4 Le cas des trois systèmes

Système	Situation pour l'IA locale
Mac (Apple Silicon — M1, M2, M3, M4)	Très favorable. L'architecture de ces Mac partage la mémoire entre processeur et partie graphique, ce qui les rend efficaces et silencieux pour l'IA, même sans carte dédiée.
Windows	Parfaitement supporté. Idéal avec une carte graphique de jeu récente ; fonctionne aussi sans, sur le processeur.
Linux	Pleinement supporté, souvent le choix des utilisateurs avancés et des serveurs. Installation par une simple commande.

NIVEAU TECHNIQUE — VRAM, MÉMOIRE UNIFIÉE ET ESTIMATION FINE

Sur PC avec GPU dédié, le facteur limitant n'est pas la RAM système mais la **VRAM** (mémoire de la carte graphique) : pour des performances optimales, le modèle doit y tenir entièrement. Une carte de 8 Go de VRAM vise des modèles 7B–8B en `Q4` ; 16 Go ouvrent les 14B confortablement ; 24 Go visent les 32B. Au-delà de la VRAM, l'*offloading* vers la RAM système fonctionne mais réduit fortement la vitesse.

Sur Apple Silicon, la **mémoire unifiée** est partagée : un Mac 32 Go peut allouer l'essentiel de cette mémoire au modèle, d'où son efficacité. Estimation rapide de l'empreinte : *nombre de milliards de paramètres* × *0,6 à 0,75 Go* en `Q4`, plus une marge pour le « contexte » (l'historique de conversation). Exemple : un 13B en `Q4` ≈ 8–9 Go, à quoi il faut ajouter le système d'exploitation et les applications ouvertes.

5. Procédure pas à pas : votre première IA locale

> Voici le cœur du guide. Deux voies sont proposées : la **Voie A**, tout en interface graphique, recommandée si vous débutez ; la **Voie B**, par la ligne de commande, plus rapide pour les habitués. Choisissez-en une seule. Comptez 15 à 30 minutes, surtout du temps de téléchargement.

Avant de commencer — préparation commune

- Vérifiez votre RAM (partie 4) et notez la taille de modèle visée.
- Assurez-vous d'avoir **au moins 10 à 20 Go d'espace disque libre** : les modèles sont volumineux.
- Prévoyez une connexion internet correcte **pour le téléchargement initial uniquement**. Ensuite, vous pourrez la couper.
- Branchez les ordinateurs portables sur secteur : le premier lancement sollicite la machine.

Voie A — La plus simple : LM Studio (interface graphique)

LM Studio est une application gratuite qui ressemble à une messagerie. Aucune ligne de commande, tout se fait à la souris. C'est la voie recommandée pour une première fois.

1 Télécharger LM Studio

Rendez-vous sur le site officiel `lmstudio.ai` avec votre navigateur. Le site reconnaît votre système (Windows, Mac ou Linux) et propose le bon fichier. Cliquez sur le bouton de téléchargement.

> **VIGILANCE**
>
> Téléchargez toujours depuis le site officiel. Évitez les liens trouvés via une publicité ou un site tiers : un logiciel d'IA détourné pourrait être malveillant.

2 Installer l'application

Ouvrez le fichier téléchargé et suivez l'installation, exactement comme pour n'importe quelle application. Sur Windows, lancez l'installateur ; sur Mac, glissez l'icône dans le dossier Applications. Puis ouvrez LM Studio.

3 Choisir et télécharger un premier modèle

Dans LM Studio, ouvrez la section de recherche de modèles (icône en forme de loupe). Tapez le nom d'un modèle adapté à votre matériel — voir la partie 6 pour bien choisir. Un bon premier choix sur une machine de 16 Go de RAM est un modèle de la famille **Llama** ou **Qwen** en taille **7B ou 8B**.

L'application indique généralement si le modèle est compatible avec votre ordinateur (mentions du type « bon ajustement »). Choisissez une version `Q4_K_M` si le choix vous est proposé, puis cliquez sur *Download*. Le téléchargement fait plusieurs gigaoctets : patientez.

4 Charger le modèle et lancer la conversation

Une fois le téléchargement terminé, ouvrez la section *Chat* et sélectionnez le modèle en haut de la fenêtre. LM Studio le charge en mémoire (quelques secondes à une minute). Une zone de saisie apparaît : écrivez votre première question et validez. La réponse s'affiche, produite entièrement par votre ordinateur.

5 Vérifier le fonctionnement hors ligne

Pour constater que tout se passe en local : coupez le Wi-Fi ou débranchez le câble réseau, puis posez une nouvelle question. L'IA répond toujours. C'est la preuve concrète que vous disposez désormais d'une IA offline.

> **CONSEIL**
>
> Si les réponses sont trop lentes ou que l'ordinateur peine, revenez à l'étape 3 et choisissez un modèle d'une taille inférieure (par exemple 3B au lieu de 7B). Mieux vaut un modèle un peu plus modeste qui répond avec fluidité.

Voie B — Par la ligne de commande : Ollama

Ollama pilote l'IA locale depuis un terminal. Plus sobre, très rapide à mettre en place, c'est l'outil favori des utilisateurs techniques et la base de nombreuses applications avancées.

1 Installer Ollama

Sur le site officiel `ollama.com`, téléchargez l'application pour Windows ou Mac et installez-la normalement. Sur Linux, une seule commande d'installation est fournie sur le site. Une fois installé, Ollama tourne discrètement en arrière-plan.

2 Ouvrir un terminal

Le *terminal* est la fenêtre où l'on tape des commandes. Sur Windows : « Terminal » ou « PowerShell » depuis le menu Démarrer. Sur Mac : application « Terminal ». Sur Linux : votre émulateur de terminal habituel.

3 Télécharger et lancer un modèle en une commande

Tapez la commande suivante, puis validez avec Entrée :

```
ollama run llama3.2
```

Ollama télécharge le modèle s'il n'est pas déjà présent, puis le lance. Remplacez `llama3.2` par le nom du modèle voulu (partie 6). Pour un petit modèle sur une machine modeste : `ollama run llama3.2:3b`.

4 Converser

Quand l'invite `>>>` apparaît, écrivez votre question et validez. La réponse s'affiche dans le terminal. Pour quitter la conversation, tapez `/bye`. Pour reprendre plus tard, il suffit de relancer la même commande `ollama run` — le modèle est déjà téléchargé.

CONSEIL — LE CONFORT D'UNE VRAIE INTERFACE AVEC OLLAMA

Le terminal n'est pas agréable pour de longues conversations. On peut conserver Ollama comme moteur tout en ajoutant une interface graphique élégante par-dessus : **Open WebUI** recrée une expérience proche de ChatGPT dans votre navigateur, et **Jan** est une application de bureau autonome. C'est ce que font la plupart des utilisateurs réguliers.

NIVEAU TECHNIQUE — AU-DELÀ DE LA CONVERSATION

Ollama expose un serveur local (par défaut sur `localhost:11434`) dont l'API est **compatible avec celle d'OpenAI**. Concrètement, toute application ou bibliothèque conçue pour l'API d'OpenAI peut être pointée vers Ollama en changeant l'URL de base — utile pour faire tourner en local des outils existants. La commande `ollama list` affiche les modèles installés, `ollama pull` télécharge sans lancer, `ollama rm` supprime un modèle pour libérer de l'espace disque. Un fichier `Modelfile` permet de figer une instruction système et des paramètres dans un modèle personnalisé.

6. Bien choisir son modèle

> Le paysage des modèles évolue très vite. Plutôt que de retenir des noms qui changeront, retenez la **méthode** de choix : elle, restera valable.

6.1 Les trois questions à se poser

1. **Quelle taille mon matériel supporte-t-il ?** C'est la contrainte première. Reportez-vous au tableau de la partie 4 et ne visez pas au-dessus de vos moyens : un modèle trop gros ne se lancera pas ou rendra la machine inutilisable.
2. **Pour quel usage ?** Conversation générale et rédaction, aide au code, ou tâche spécialisée ? Certaines familles de modèles sont plus douées pour le code, d'autres pour le multilingue.
3. **Quelle langue ?** Si vous travaillez surtout en français, privilégiez un modèle reconnu pour ses bonnes performances multilingues.

6.2 Les grandes familles de modèles (mi-2026)

À titre indicatif, voici les familles les plus répandues pour un usage local. Les noms et numéros de version changeront ; l'esprit de chaque famille est plus durable.

Famille	Origine	Réputation pour l'usage local
Llama	Meta	Le plus polyvalent. Bon choix par défaut pour la conversation générale, décliné en plusieurs tailles.
Mistral	Mistral AI (France)	Réputé efficace et bon en multilingue, dont le français. Très bon rapport qualité / légèreté.
Qwen	Alibaba	Apprécié pour le code et les mathématiques, disponible dans une très large gamme de tailles.
Gemma	Google	Modèles compacts et soignés, adaptés aux machines modestes.
Phi	Microsoft	Petits modèles étonnamment capables, pensés pour le matériel limité.
DeepSeek	DeepSeek	Variantes très orientées code et raisonnement, dont des versions spécialisées pour la programmation.

6.3 Recommandation de départ

Si vous voulez une réponse simple sans vous perdre dans les comparaisons : sur une machine de **16 Go de RAM**, partez d'un modèle **Llama ou Mistral en taille 7B/8B, quantification** `Q4_K_M`. C'est l'équilibre que recommandent la plupart des guides : assez capable pour un usage quotidien, assez léger pour rester fluide. Vous affinerez ensuite selon votre ressenti.

CONSEIL — ESSAYEZ AVANT DE VOUS FIXER

Rien ne vous oblige à un seul modèle. Les outils permettent d'en télécharger plusieurs et de basculer de l'un à l'autre. Une pratique courante : utiliser LM Studio pour *tester* rapidement plusieurs modèles, puis garder celui qui vous convient. Pensez seulement à supprimer ceux que vous n'utilisez pas — ils occupent beaucoup d'espace disque.

NIVEAU TECHNIQUE — AU-DELÀ DE LA TAILLE BRUTE

Le nombre de paramètres ne fait pas tout. À taille égale, deux modèles peuvent différer nettement selon la qualité de leur entraînement, leur ajustement aux instructions (*instruct tuning*) et leur date de sortie. Un bon 8B récent peut dépasser un 13B plus ancien. Surveillez aussi la **fenêtre de contexte** (le volume de texte que le modèle peut prendre en compte d'un coup, exprimé en tokens) : elle conditionne votre capacité à traiter de longs documents. Enfin, pour des choix éclairés, consultez des classements comparatifs publics plutôt que de vous fier au seul nombre de paramètres — mais gardez un œil critique, ces classements ne reflètent pas toujours votre usage réel.

7. Aller plus loin (usages avancés)

> Une fois la conversation de base maîtrisée, plusieurs extensions augmentent nettement l'utilité de votre IA locale. Cette partie est un survol : chacune mérite son propre approfondissement.

7.1 Faire dialoguer l'IA avec vos propres documents

C'est sans doute l'extension la plus utile. Plutôt que de poser des questions générales, vous fournissez à l'IA vos fichiers — rapports, notes, contrats, cours — et vous l'interrogez **sur leur contenu**. L'IA cite vos documents pour répondre. Des outils comme **AnythingLLM** ou **Open WebUI** proposent cette fonction clé en main : on dépose ses fichiers, on pose ses questions. Tout reste local, donc même des documents confidentiels peuvent être traités sans risque de fuite.

7.2 Brancher l'IA sur d'autres logiciels

Une IA locale peut assister directement dans d'autres applications. L'exemple le plus courant est l'**aide au code dans un éditeur de programmation** : des extensions permettent à un modèle local de suggérer et d'expliquer du code sans que celui-ci ne quitte la machine — précieux pour un code source confidentiel.

7.3 Les modèles spécialisés

Au-delà des modèles généralistes, il existe des modèles affûtés pour un domaine : programmation, langues, raisonnement mathématique. Si vous avez un usage dominant et précis, un modèle spécialisé de taille modeste peut surpasser un généraliste plus gros sur ce terrain particulier.

7.4 Au-delà du texte

L'écosystème local ne se limite pas à la conversation écrite. Il existe des modèles locaux pour **générer des images**, pour **transcrire de la parole en texte**, ou des modèles dits « multimodaux » capables d'analyser une image que vous leur soumettez. Ces usages demandent souvent davantage de matériel et une installation distincte.

NIVEAU TECHNIQUE — PISTES POUR APPROFONDIR

RAG avancé : au-delà des outils clé en main, des cadres comme LlamaIndex ou LangChain permettent de construire des pipelines sur mesure (découpage des documents, base vectorielle, stratégie de récupération). **Agents :** des outils tels que des extensions d'édition de code en mode agent peuvent enchaîner des actions de façon autonome, entièrement hors ligne. **Fine-tuning :** il est possible de réentraîner partiellement un modèle sur vos propres données (techniques de type LoRA) pour le spécialiser — démarche exigeante en matériel et en savoir-faire. **Quantification fine et serveurs :** pour un déploiement multi-utilisateurs, on s'oriente vers des moteurs d'inférence optimisés et une gestion fine de la VRAM et du débit.

8. Dépannage et questions fréquentes

8.1 Les blocages les plus courants

Symptôme	Cause probable et solution
Les réponses sont très lentes, mot à mot	Le modèle est trop gros pour votre matériel, ou vous n'avez pas de carte graphique. Solution : choisir un modèle plus petit (3B au lieu de 7B), fermer les autres applications gourmandes.
Le modèle refuse de se charger / message de mémoire insuffisante	Le modèle dépasse la mémoire disponible. Solution : passer à une taille inférieure ou à une quantification plus légère.
L'ordinateur chauffe, le ventilateur s'emballe	Normal pendant l'inférence : la machine travaille intensément. Si c'est excessif, optez pour un modèle plus petit.
Le téléchargement du modèle échoue	Connexion interrompue ou espace disque insuffisant. Vérifiez l'espace libre et relancez ; les outils reprennent en général là où ils s'étaient arrêtés.
Les réponses sont décevantes ou erronées	Limite normale d'un petit modèle local. Essayez un modèle plus grand si le matériel le permet, formulez des questions plus précises, ou utilisez le RAG pour les questions factuelles.
L'IA invente des informations	Comportement commun à toutes les IA (« hallucination »), plus marqué sur les petits modèles. Ne jamais considérer une réponse comme une vérité vérifiée ; recouper les informations importantes.

8.2 Questions fréquentes

Est-ce vraiment gratuit ?

Les outils et la grande majorité des modèles sont gratuits. Les seuls coûts sont l'électricité et, si vous le souhaitez, du matériel plus performant. Aucun abonnement.

Mes données sont-elles vraiment privées ?

Oui, tant que vous restez en configuration de base, sans module de recherche web ni connecteur externe. Dès que vous activez une fonction qui interroge internet, une partie des données peut sortir : c'est à vous de décider, fonction par fonction.

Puis-je supprimer un modèle qui ne me sert plus ?

Oui, et c'est recommandé : les modèles occupent beaucoup d'espace. LM Studio propose une gestion des modèles installés ; avec Ollama, la commande `ollama rm` suivie du nom du modèle s'en charge.

L'IA locale peut-elle remplacer ChatGPT ou Claude ?

Pour de nombreux usages quotidiens, oui. Pour les tâches les plus exigeantes, les meilleurs services cloud gardent l'avantage. Beaucoup d'utilisateurs adoptent une approche mixte : l'IA locale pour tout ce qui est courant ou confidentiel, le cloud pour les cas difficiles.

Faut-il une machine récente et chère ?

Non pour débuter. Un ordinateur des dernières années avec 16 Go de RAM fait très bien l'affaire avec des modèles moyens. Un investissement matériel ne se justifie que si vous visez les très grands modèles.

9. Glossaire et ressources

9.1 Glossaire

Terme	Définition
Cloud	Serveurs distants accessibles par internet. Une IA « cloud » s'exécute ailleurs que sur votre machine.
Contexte (fenêtre de)	Quantité de texte qu'un modèle peut prendre en compte simultanément, mesurée en tokens.
GPU / carte graphique	Composant qui accélère fortement l'IA. Utile mais non indispensable pour débuter.
Hallucination	Réponse fausse présentée avec assurance par une IA. Comportement commun à tous les modèles.
Inférence	Le travail de production d'une réponse par le modèle.
Modèle	Fichier contenant l'intelligence de l'IA, à télécharger une fois.
Modèle à poids ouverts	Modèle dont le fichier est librement téléchargeable et utilisable.
Multimodal	Se dit d'un modèle capable de traiter autre chose que du texte (images, son).
Paramètres (« B »)	Mesure de la taille d'un modèle, en milliards. Ex. « 7B » = 7 milliards.
Quantification	Compression d'un modèle pour l'alléger. Format standard : `Q4_K_M`.
RAG	Technique fournissant des documents au modèle pour qu'il réponde à partir de sources à jour.
RAM	Mémoire vive de travail de l'ordinateur. Facteur déterminant pour la taille de modèle possible.
Token	Petit fragment de texte, unité de base manipulée par l'IA.
VRAM	Mémoire dédiée de la carte graphique.

9.2 Outils cités dans ce guide

Outil	Rôle
LM Studio	Application graphique tout-en-un. Idéale pour débuter et tester des modèles.
Ollama	Moteur d'IA locale piloté en ligne de commande. Base de nombreux usages avancés.
Open WebUI	Interface web type messagerie, à poser au-dessus d'Ollama.
Jan	Application de bureau autonome pour l'IA locale.
AnythingLLM	Outil orienté dialogue avec ses propres documents (RAG clé en main).

9.3 Pour finir

L'IA locale n'est pas un substitut universel aux services cloud, mais un **complément précieux** dès que comptent la confidentialité, l'autonomie, la maîtrise des coûts ou le fonctionnement hors ligne. La meilleure façon de se faire une idée est d'essayer : une machine ordinaire, un outil gratuit, un modèle de taille moyenne, et une demi-heure devant soi suffisent pour disposer d'une IA qui vous appartient entièrement.

> **AVERTISSEMENT**
>
> Ce domaine évolue très rapidement : noms de modèles, numéros de version et fonctionnalités des outils changent de mois en mois. Les **méthodes** et **principes** exposés ici restent valables ; vérifiez en revanche les **noms précis et versions** sur les sites officiels au moment de votre installation. Document à jour en mai 2026.

Par Jawed Tahir — Mai 2026. Site personnel de l'auteur : javed.fr

Projet conçu et développé par **Virgule Studio**, studio marseillais de conception et de développement web et applications : virgule-studio.fr