Meilleurs outils d’IA locale pour la productivité en 2026

Comparez sept outils d’IA locale pour la productivité en 2026 : IA de réunion sur l’appareil, dictée, chat sur ordinateur, notes et questions-réponses sur des documents, avec des limites cloud clairement définies.

Publié par Julian Pscheid · 1 juillet 2026
Lire cet article en ligne: https://www.hedy.ai/fr/post/best-local-ai-tools-for-productivity/

Homme parlant pendant une réunion virtuelle depuis son bureau à domicile, avec l’appel vidéo sur un moniteur externe, des outils de productivité sur l’écran de son ordinateur portable et un iPhone posé face contre table qui émet une douce lueur à côté du clavier

Réponse rapide Les outils d’IA locale exécutent des modèles d’IA sur votre propre appareil plutôt que dans le cloud d’un fournisseur. Le meilleur dépend de la tâche. Pour les réunions, Hedy exécute la reconnaissance vocale sur l’appareil sur toutes les plateformes, et les machines capables d’exécuter de puissants modèles locaux peuvent aussi faire tourner l’ensemble du pipeline IA sur l’appareil. Pour la dictée, VoiceInk transcrit entièrement sur votre Mac. Pour le chat sur ordinateur, Jan exécute des modèles ouverts hors ligne. Pour les notes, Obsidian conserve tout dans des fichiers locaux. Chacun des outils ci-dessous prend en charge une tâche, et ensemble ils offrent à chaque moment d’une journée de travail une solution locale et privée.

Qu’est-ce qui peut être considéré comme de l’IA locale ?

L’IA locale signifie que le modèle s’exécute sur votre matériel : votre ordinateur portable, votre téléphone, votre serveur. Vos données sont traitées là où elles se trouvent déjà au lieu d’être téléversées dans le cloud d’un fournisseur.

En pratique, il s’agit d’un spectre, et prétendre le contraire est une façon pour les outils de vous induire en erreur. Il existe trois positions transparentes :

• Entièrement local. Le modèle s’exécute sur votre appareil et l’outil fonctionne sans réseau. Jan, Ollama, PocketPal AI et le mode par défaut de VoiceInk appartiennent à cette catégorie.

• Local par défaut, cloud en option. La voie privée est celle par défaut ; vous pouvez connecter un fournisseur cloud si vous le souhaitez. La transcription cloud optionnelle de VoiceInk fonctionne ainsi.

• Hybride avec des garanties locales. Une partie du pipeline est locale par conception, une autre utilise le cloud, et l’outil indique clairement laquelle est laquelle. Hedy appartient à ce groupe : la reconnaissance vocale s’exécute sur l’appareil par défaut sur toutes les plateformes, et les machines capables d’exécuter de puissants modèles locaux peuvent faire tourner l’ensemble du pipeline sur l’appareil.

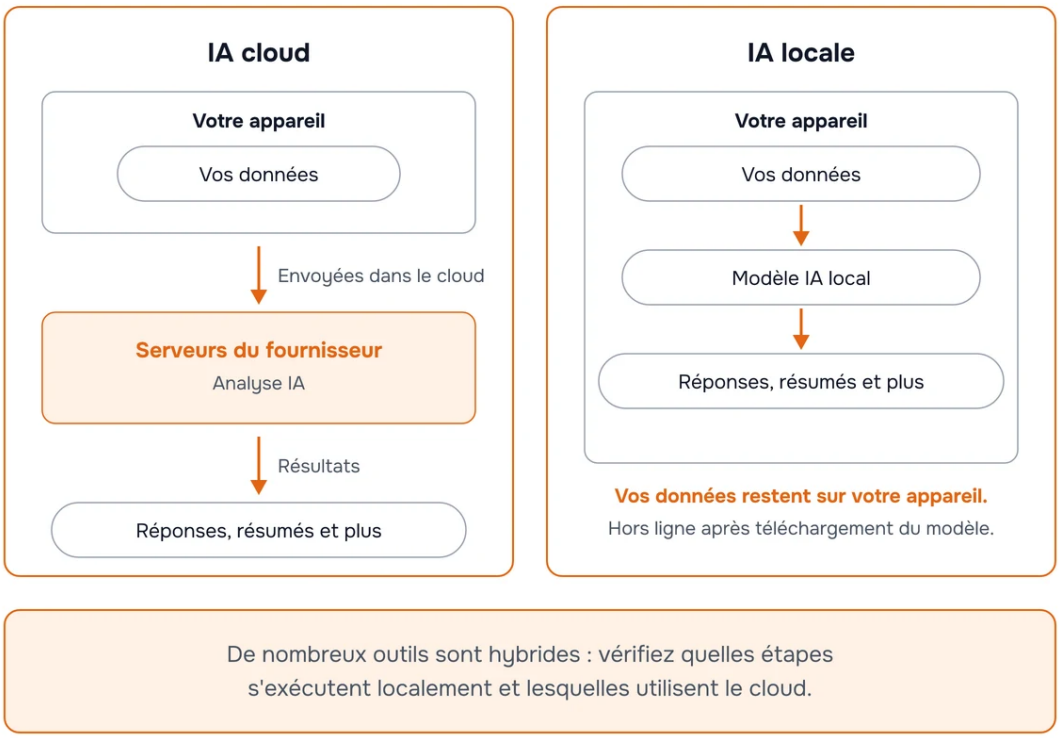

Schéma comparant l’IA cloud, où vos données sont envoyées aux serveurs du fournisseur pour être traitées, à l’IA locale, où vos données, le modèle IA et les résultats restent tous sur votre appareil

Utiliser le cloud ne disqualifie pas un outil de cette liste. Rester vague à ce sujet, si. Chaque entrée ci-dessous indique quelles étapes s’exécutent sur votre machine.

Pourquoi l’IA locale connaît-elle un tel essor ?

Deux courbes se sont croisées. Les appareils deviennent toujours plus puissants, tandis que les modèles deviennent à la fois plus petits et plus intelligents. À un moment donné l’année dernière, les modèles qui tiennent sur un ordinateur portable ou un téléphone récent sont devenus suffisamment performants pour accomplir un vrai travail : transcrire une conversation, résumer un document, rédiger une réponse. Nous avons assisté directement à ce point de bascule en construisant le pipeline local de Hedy et avons détaillé le processus dans notre analyse technique approfondie de l’IA locale (/fr/post/local-ai-

engineering-deep-dive-hedy-3-2/) .

Les inquiétudes concernant la destination des données confiées à l'IA ont augmenté au cours de la même période. Pew Research a constaté en juin 2026 (https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/) que 71% des adultes américains s'attendent à ce que l'utilisation accrue de l'IA rende leurs informations personnelles moins sûres ; seuls 3% s'attendent au contraire. Cette inquiétude n'empêche pas les gens d'utiliser l'IA : ils continuent de lui fournir du contenu sensible. Cyberhaven a mesuré (https://info.cyberhaven.com/hubfs/Content%20PDF/Cyberhaven%20Q2%202024%20AI%20Adoption%20and%20Risk%20Report%20052024.pdf) que 27.4% des données saisies par les employés dans des outils d'IA au Q2 2024 étaient sensibles, contre 10.7% un an plus tôt. L'exemple emblématique reste l'incident Samsung de 2023 (https://www.engadget.com/three-samsung-employees-reportedly-leaked-sensitive-data-to-chatgpt-190221114.html) , lorsqu'un employé a téléversé l'enregistrement d'une réunion interne dans ChatGPT pour générer un compte rendu.

Voici notre pari, qui guide cet article : l'IA locale est une direction, pas une niche. Le cloud conservera longtemps l'avantage pour les capacités de pointe et, pour la plupart des gens, il restera le meilleur choix par défaut pour les raisonnements complexes. Mais le niveau minimal du local s'élève chaque année, le matériel se trouve déjà dans votre poche et sur votre bureau (Gartner prévoit (https://www.gartner.com/en/newsroom/press-releases/2025-08-28-gartner-says-artificial-intelligence-pcs-will-represent-31-percent-of-worldwide-pc-market-by-the-end-of-2025) que les PC IA représenteront 55% du marché en 2026), et dès qu'une tâche peut être bien accomplie sur l'appareil, rien ne justifie d'envoyer ces données ailleurs. « Où cela s'exécute-t-il ? » devient une question standard à poser pour chaque outil d'IA, comme « Est-ce chiffré ? » l'est devenue pour la messagerie. Les outils ci-dessous couvrent les tâches pour lesquelles la réponse peut déjà être « ici même ».

Comment avons-nous choisi ces outils ?

Cinq règles ont guidé cette liste :

• Un outil par tâche. Une pile de spécialistes vaut mieux qu'un ensemble d'applications qui se chevauchent et conservent chacune une copie de vos données. Nous avons retenu une option axée sur le local pour chaque tâche de productivité : réunions, dictée, chat, exécution de modèles, notes, documents et mobile.

• Le local doit être réel. Chaque outil s'exécute soit entièrement sur l'appareil, soit documente précisément les étapes qui le font. Un marketing vague axé sur la « confidentialité » sans détails d'architecture n'était pas suffisant.

• Actif en 2026. Nous avons vérifié les tarifs, les fonctions et l'activité du projet de chaque outil sur les pages et dépôts des fournisseurs en juillet 2026.

• Les hybrides sont acceptés, mais signalés. Les outils qui utilisent le cloud pour certaines étapes sont inclus lorsque la limite est explicite et sous le contrôle de l'utilisateur.

• Transparence totale : Hedy est notre produit. Il s'agit d'une sélection publiée par son propriétaire, pas d'une étude comparative indépendante, et Hedy occupe volontairement la place des réunions. Les autres outils de réunion concurrents (Granola, Fireflies, le mode réunion de Superwhisper) ne sont pas classés ici, car nous ne sommes pas un arbitre neutre dans cette catégorie. Nous avons rédigé des comparatifs distincts et sourcés, comme nos guides sur une alternative à Granola (/fr/post/granola-alternative/) et une alternative à Fireflies (/fr/post/fireflies-alternative-hedy/) , pour les lecteurs qui évaluent ces outils.

Quel outil d’IA locale convient le mieux à chaque tâche ?

Comparatif en un coup d’œil

#	Outil	Idéal pour	Fonctionnement local	Plateformes	Prix
1	Hedy (https://www.hedy.ai/)	Réunions et conversations en direct	Reconnaissance vocale sur l’appareil par défaut ; pipeline entièrement local en option sur le matériel compatible	iOS, Android, Mac, Windows, web	Gratuit (5 h/mois) ; $12.99/mo ou $99.99/an
2	VoiceInk (https://tryvoiceink.com/)	Dictée	Entièrement sur l’appareil par défaut ; fournisseurs cloud en option	macOS (application iOS disponible)	$25 à $49 en achat unique
3	Jan (https://jan.ai/)	Chat IA sur ordinateur	Entièrement local par défaut ; API cloud en option	Windows, Mac, Linux	Gratuit
4	Ollama (https://ollama.com/)	Exécution locale de modèles	Entièrement local ; offre cloud en option	Mac, Windows, Linux	Gratuit ; Pro $20/mo
5	Obsidian (https://obsidian.md/)	Notes et base de connaissances	Fichiers Markdown locaux ; les plugins peuvent utiliser des modèles locaux	Mac, Windows, Linux, iOS, Android	Gratuit ; Sync à partir de $4/mo
6	PrivateGPT (https://github.com/zylon-ai/private-gpt)	Questions-réponses sur des documents privés	Auto-hébergé ; vous choisissez le serveur d’inférence	Auto-hébergé (Mac, Linux, Windows)	Gratuit
7	PocketPal AI (https://pocketpal.dev/)	Chat IA sur votre téléphone	Entièrement sur l’appareil après le téléchargement du modèle	iOS, Android	Gratuit

Sources : tarifs de Hedy (/fr/pricing/) , tarifs de VoiceInk (https://tryvoiceink.com/pricing) , documentation de Jan (https://jan.ai/docs) , tarifs d’Ollama (https://ollama.com/pricing) , tarifs d’Obsidian (https://obsidian.md/pricing) , dépôt de PrivateGPT (https://github.com/zylon-ai/private-gpt) , PocketPal AI (https://pocketpal.dev/) . Prix vérifiés sur les pages des fournisseurs en juillet 2026 et susceptibles de changer souvent ; vérifiez à nouveau avant d’acheter.

1. Hedy : réunions et conversations en direct

Les réunions sont le lieu où les données professionnelles les plus sensibles sont prononcées à voix haute : noms de clients, conditions commerciales, informations de santé, stratégie. C’est aussi un flux de travail où les outils d’IA populaires ont tendance à être les moins respectueux de la vie privée, car de nombreux assistants de réunion fonctionnent comme un bot qui rejoint votre appel et envoie l’audio à un serveur.

Hedy (https://www.hedy.ai/) adopte l’approche inverse. La reconnaissance vocale s’exécute sur l’appareil par défaut sur toutes les plateformes (Whisper partout, ainsi que le moteur Nemotron avec identification des intervenants sur l’appareil). Avec les moteurs par défaut, votre audio ne quitte donc jamais votre appareil et aucun bot ne rejoint vos appels. En plus de la transcription, Hedy agit comme un coach de réunion en temps réel : il propose des suggestions, des réponses et des questions pertinentes pendant que la conversation est encore en cours, puis produit ensuite des résumés, des points clés et des tâches à accomplir.

Les réglages par défaut de Hedy reposent sur une séparation délibérée. La reconnaissance vocale s’exécute localement sur chaque appareil parce que cela est possible partout, avec un niveau de qualité élevé. L’analyse IA s’exécute dans le cloud par défaut, car nous voulons que Hedy fonctionne sur le plus grand nombre d’appareils possible avec les meilleurs résultats que nous puissions fournir, et les données envoyées ne sont jamais stockées ni utilisées pour entraîner des modèles. Mais le plafond local est bien plus élevé que le réglage par défaut : toute machine capable d’exécuter de puissants modèles locaux peut faire tourner l’ensemble du pipeline de réunion sur l’appareil grâce au Traitement IA local (/fr/help/local-ai-processing/) . Transcriptions, résumés, notes détaillées, réponses de chat et suggestions en direct sont tous produits sur la machine qui a capturé l’audio, hors ligne si vous le souhaitez, sur les

Macs Apple Silicon, les machines Windows équipées de GPU adaptés, ainsi que les iPhones et iPads récents. Désactivez Cloud Sync et la conversation n'existe que sur l'appareil qui l'a capturée, de bout en bout. Hedy 3.2 a rendu cela possible, comme nous l'expliquons dans L'IA locale pour les réunions (/fr/post/local-ai-meetings-hedy-3-2/) . Vous pouvez aussi désactiver l'analyse IA cloud (/fr/help/cloud-ai-analysis-privacy-control/) tout en conservant la transcription complète, choisir la résidence des données dans l'UE (/fr/post/eu-data-residency/) et examiner chaque réglage dans notre guide des paramètres de confidentialité (/fr/post/ai-meeting-privacy-settings-explained/) .

Le mode par défaut n'est pas entièrement local ; le mode entièrement local est disponible lorsque votre matériel et vos exigences de confidentialité le nécessitent, et dans les deux cas vous disposez d'une limite documentée que vous contrôlez. Hedy a aussi réalisé un examen indépendant SOC 2 Type I de ses contrôles de sécurité et une évaluation HIPAA indépendante en tant que Business Associate en avril 2026. Il est noté 4.8 sur 5 sur 500+ avis par 30,000+ utilisateurs.

Prix : gratuit pour 5 heures de sessions par mois, sans carte bancaire. Pro coûte $12.99/month ou $99.99/year ; une licence à vie à $299 couvre tout de façon permanente.

Idéal pour : toute personne dont les conversations constituent les données les plus sensibles qu'elle confierait autrement à un outil cloud : consultants, commerciaux, journalistes, patients et managers.

2. VoiceInk : dictée

La dictée est le gain de productivité discret de l'IA locale. La plupart des gens parlent beaucoup plus vite qu'ils ne tapent, et un bon modèle sur l'appareil transcrit désormais avec une précision suffisante pour qu'il n'y ait aucune raison de téléverser votre voix.

VoiceInk (https://tryvoiceink.com/) est le spécialiste de la dictée pour Mac. Il transcrit votre voix sur l'appareil à l'aide de modèles Whisper et Parakeet locaux, fonctionne hors ligne et conserve votre historique de transcription sur votre machine, avec suppression automatique configurable. Un dictionnaire personnel gère les noms et le jargon, tandis que des modes propres à chaque application adaptent le formatage aux e-mails, au chat ou au code. Il est open source (https://github.com/Beingpax/VoiceInk) , ce qui permet de vérifier ses affirmations en matière de confidentialité plutôt que de les croire sur parole ; le projet compte 5,500+ étoiles GitHub et le développeur indique plus de 200,000 téléchargements. Des fonctions d'amélioration cloud et des fournisseurs de transcription cloud sont disponibles en option, mais doivent être activés volontairement et sont clairement signalés.

Pourquoi VoiceInk plutôt que les outils plus connus Superwhisper ou MacWhisper ? La spécialisation. Tous deux se sont étendus à l'enregistrement de réunions (Superwhisper (https://superwhisper.com/meeting-transcription) , MacWhisper (https://macwhisper.helpscoutdocs.com/article/30-record-meetings)), ce qui les place dans une autre catégorie (et une autre discussion sur la confidentialité) que la simple dictée. L'application Mac de VoiceInk s'en tient à une seule tâche : elle tape ce que vous dites. Son application iOS distincte mentionne la transcription de réunions et de cours parmi ses cas d'usage, mais ne propose ni capture d'appels, ni identification des intervenants, ni résumés ; le produit reste centré sur la dictée. Pour les conversations en direct avec d'autres personnes, c'est le rôle de Hedy, et les deux fonctionnent côte à côte.

Prix : achat unique de $25 (un Mac) à $49 (trois Macs), mises à jour à vie incluses. Une application iOS distincte est actuellement gratuite.

Idéal pour : les auteurs, les développeurs et toute personne qui répond à cinquante e-mails par jour et ne veut ni abonnement ni dépendance au cloud pour ce privilège.

3. Jan : chat IA sur ordinateur

Si ChatGPT est votre outil de réflexion par défaut, mais que vous préférez ne pas envoyer chaque idée incomplète à un serveur, Jan (https://jan.ai/) est l'alternative locale la plus simple à adopter. Téléchargez l'application, choisissez un modèle ouvert (Llama, Qwen, Gemma et des dizaines d'autres au format GGUF), puis discutez. Aucun compte à créer et, une fois le modèle téléchargé, le tout fonctionne même lorsque le Wi-Fi est désactivé.

Jan est open source (https://github.com/janhq/jan) , avec 43,000+ étoiles GitHub et 5.9 millions de téléchargements annoncés. Deux choix de conception en font plus qu'un simple gadget. Premièrement, il expose une API compatible avec OpenAI sur localhost, ce qui permet aux autres applications de cette liste, ainsi qu'à vos propres scripts, d'utiliser votre modèle local comme backend interchangeable. Deuxièmement, il est transparent sur l'utilisation hybride : vous pouvez ajouter les clés API de fournisseurs cloud lorsque vous voulez la qualité d'un modèle de pointe, et Jan indique clairement que ces chats quittent alors votre machine selon les conditions du fournisseur concerné.

Prix : gratuit.

Idéal pour : la rédaction et le brainstorming du quotidien, lorsque « assez bon en local » vaut mieux qu'« excellent, mais téléversé ».

4. Ollama : exécution locale de modèles

Ollama (https://ollama.com/) est la plomberie du mouvement de l'IA locale. Il s'agit d'un outil en ligne de commande et d'un service en arrière-plan qui télécharge et exécute des modèles ouverts avec une seule commande (ollama run llama3), puis les met à disposition via une API REST locale utilisée par un vaste écosystème d'applications. Si vous utilisez plusieurs outils d'IA locale, il est fort probable qu'ils puissent tous partager un même backend Ollama au lieu d'intégrer chacun leur propre moteur d'exécution des modèles.

Il est open source (https://github.com/ollama/ollama) , compte 176,000+ étoiles GitHub, et l'entreprise a déclaré servir 8.9 millions de développeurs en juillet 2026. Tout ce qui est local est gratuit et illimité ; des offres cloud payantes (à partir de $20/month) permettent d'exécuter de plus grands modèles sur le matériel d'Ollama, séparément de la voie locale.

Prix : gratuit pour une utilisation locale illimitée ; offres cloud en option à partir de $20/month.

Idéal pour : les utilisateurs avancés et toute personne qui souhaite un backend de modèles commun, programmable et partagé derrière ses autres outils.

5. Obsidian : notes et connaissances personnelles

Vos notes cartographient tout ce sur quoi vous travaillez. Un outil de notes exclusivement cloud constitue donc l'un des paris les plus concentrés sur la confidentialité que les gens font sans y penser. Obsidian (https://obsidian.md/) évite entièrement ce pari : les notes sont de simples fichiers Markdown dans un dossier de votre appareil. L'application fonctionne hors ligne, les fichiers survivent à l'entreprise et rien ne se synchronise sauf si vous configurez une synchronisation (le service Sync en option, à partir de $4/month facturé annuellement, est chiffré de bout en bout).

La couche IA provient de plugins communautaires et peut rester locale. Smart Connections (https://github.com/brianpetro/obsidian-smart-connections) crée sur l'appareil des embeddings de votre coffre pour rechercher des notes connexes, tandis que Copilot for Obsidian (https://github.com/logancyang/obsidian-copilot) ajoute un chat sur vos notes avec la prise en charge de backends locaux comme

Ollama. C’est ici qu’une pile locale partagée commence à porter ses fruits. Avec un modèle local derrière ces outils, « discuter avec mes notes » ne fait intervenir aucun tiers. Les deux plugins prennent aussi en charge des fournisseurs cloud ; lisez donc leurs paramètres avec la même attention que pour tout outil hybride.

Prix : gratuit pour un usage personnel et commercial ; Sync en option à partir de $4/month facturé annuellement.

Idéal pour : toute personne qui construit une base de connaissances personnelle pérenne et souhaite y effectuer des recherches IA sans téléverser ses réflexions.

6. PrivateGPT : questions-réponses sur des documents privés

« Discuter avec vos documents » a été la première application phare de l’IA privée, et PrivateGPT (https://github.com/zylon-ai/private-gpt) en est la référence open source : 57,000+ étoiles GitHub et une relance complète en juin 2026, lorsque Zylon a publié la v1.0 (https://www.zylon.ai/resources/blog/introducing-privategpt-1-0-the-open-source-application-backend-for-private-ai) et réintégré au dépôt public deux ans de développement privé pour les entreprises.

PrivateGPT 1.0 est un backend d’application : il gère l’ingestion de documents, la recherche avec citations, le chat et l’orchestration des outils. Comme Zylon l’indique dans l’annonce de la version 1.0, « PrivateGPT 1.0 n’exécute pas lui-même les modèles » : vous le reliez au serveur d’inférence de votre choix. C’est ce qui rend possible une configuration entièrement auto-hébergée : associez-le à Ollama ou llama.cpp sur la même machine et le traitement des documents reste au sein de l’infrastructure que vous contrôlez. C’est aussi l’outil de cette liste le mieux adapté à une petite équipe qui souhaite déployer un assistant privé partagé sur son propre serveur.

Une mise en garde honnête concernant son adéquation : c’est l’outil le plus technique de cette liste. Il n’existe pas de programme d’installation en un clic, et les alternatives grand public plus abouties de cette catégorie ont pour la plupart évolué vers des fonctions de transcription de réunions, raison pour laquelle elles ne figurent pas dans cette liste.

Prix : gratuit ; vous fournissez le matériel.

Idéal pour : les utilisateurs techniques et les petites équipes qui souhaitent que les questions-réponses sur leurs documents s’exécutent entièrement sur une infrastructure qu’ils contrôlent.

7. PocketPal AI : IA locale sur votre téléphone

Le téléphone est l’appareil sur lequel l’IA locale est devenue concrète pour les personnes sans compétences techniques, et PocketPal AI (https://pocketpal.dev/) est la façon la plus simple de l’essayer. Il s’agit d’une application gratuite et open source (https://github.com/a-ghorbani/pocketpal-ai) pour iOS et Android, qui télécharge de petits modèles ouverts (Llama, Qwen, Gemma, Phi) et les exécute entièrement sur l’appareil. Aucun compte, aucune télémétrie, et elle continue de répondre en mode avion.

Le projet annonce 1.2 million de téléchargements, et sa communauté a soumis plus de 7,000 résultats de benchmarks sur l’appareil couvrant 100+ modèles de téléphone, ce qui offre aussi une réponse pratique à la question « Est-ce que cela fonctionnera sur mon téléphone ? » (Règle générale : les iPhones récents et les téléphones Android de milieu de gamme ou mieux gèrent confortablement les petits modèles.) Il n’égalera pas un modèle de pointe sur les raisonnements complexes, mais pour les brouillons rapides et les traductions, il démontre concrètement qu’une IA utile tient désormais dans une poche sans aucun serveur. C’est le même pari qu’a fait Apple en plaçant le traitement sur l’appareil au cœur d’Apple Intelligence.

Prix : gratuit.

Idéal pour : essayer l’IA locale sans aucune configuration et disposer d’un assistant de poche hors ligne pour les tâches rapides.

Comment construire une pile d’IA locale avec ces outils ?

Commencez par la tâche qui contient vos données les plus sensibles. Pour de nombreux professionnels, il s’agit des réunions et des conversations, la catégorie couverte par Hedy ; pour d’autres, ce sont les documents ou les notes. Faites passer ce flux de travail à un outil axé sur le local et vous aurez obtenu l’essentiel du gain de confidentialité en une seule étape.

Ajoutez ensuite les spécialistes un par un : un outil de dictée, un outil de chat, un outil de notes. Évitez les chevauchements. Deux outils qui accomplissent la même tâche signifient deux copies des mêmes données sensibles à deux endroits, soit l’exact opposé de l’objectif. La pile issue de cette liste partage bien son infrastructure : Ollama fournit les modèles, Jan, les plugins Obsidian et PrivateGPT peuvent tous s’appuyer dessus, VoiceInk tape ce que vous dites et Hedy couvre tout ce qui est prononcé avec d’autres personnes.

Enfin, vérifiez plutôt que de faire confiance. Désactivez le Wi-Fi et voyez ce qui continue de fonctionner. Lisez la page de confidentialité et recherchez les détails de chaque étape (audio, transcription, analyse, stockage) plutôt que des adjectifs. Tous les outils ci-dessus réussissent ce test de lecture ; c’est pourquoi ils sont ici. Pour les réunions en particulier, notre guide sur l'utilisation privée des notes de réunion IA (/fr/post/how-to-use-ai-meeting-notes-privately/) examine ce contrôle en détail : ce qui quitte votre appareil à chaque étape et pourquoi un test hors ligne vous en apprend moins que vous ne le pensez.

Questions fréquentes

Quel est le meilleur outil d’IA locale pour la productivité ?

Cela dépend de la tâche. Pour les réunions et les conversations en direct, Hedy exécute la reconnaissance vocale sur l’appareil sur toutes les plateformes, et les machines capables d’exécuter de puissants modèles locaux peuvent aussi faire tourner l’ensemble de son pipeline IA sur l’appareil. Pour la dictée, VoiceInk transcrit entièrement sur votre Mac. Pour le chat IA sur ordinateur, Jan exécute des modèles ouverts localement sans compte. Pour les notes, Obsidian conserve tout dans des fichiers locaux. La plupart des gens finissent par combiner deux ou trois outils qui couvrent chacun une tâche.

Que signifie « IA locale » ?

C’est un spectre, pas une étiquette binaire. Les outils entièrement locaux comme Jan, Ollama et PocketPal exécutent le modèle sur votre matériel et fonctionnent hors ligne. Les outils hybrides exécutent une partie du pipeline localement : Hedy, par exemple, exécute la reconnaissance vocale sur l’appareil par défaut et propose un mode optionnel de Traitement IA local qui conserve également l’étape d’analyse sur votre appareil. L’important est de savoir quelles étapes restent sur votre machine, lesquelles n’y restent pas et si vous pouvez le contrôler.

Les outils d’IA locale sont-ils aussi performants que l’IA cloud ?

Pour de nombreuses tâches de productivité, oui. La reconnaissance vocale sur l’appareil est désormais excellente, les embeddings locaux gèrent bien la recherche dans les notes et les petits modèles ouverts couvrent les résumés et la rédaction. Pour le raisonnement de pointe, les modèles cloud restent en tête.

C’est pourquoi plusieurs outils de cette liste sont des hybrides transparents : local lorsque la confidentialité compte le plus (votre audio, vos fichiers), cloud lorsque la qualité compte le plus, avec une option pour passer entièrement en local quand vous en avez besoin.

Hedy est-il entièrement local ?

Les réglages par défaut de Hedy reposent sur une séparation délibérée. La reconnaissance vocale s’exécute sur l’appareil sur toutes les plateformes (Whisper partout, ainsi que Nemotron avec identification des intervenants sur l’appareil). Avec les moteurs par défaut, votre audio ne quitte donc jamais votre appareil et aucun bot ne rejoint vos appels. L’analyse IA s’exécute dans le cloud par défaut afin que Hedy fonctionne sur le plus grand nombre d’appareils possible avec des résultats de grande qualité. Mais toute machine capable d’exécuter de puissants modèles locaux (Macs Apple Silicon, machines Windows équipées de GPU adaptés, iPhones et iPads récents) peut faire tourner l’ensemble du pipeline sur l’appareil grâce au mode optionnel de Traitement IA local (/fr/help/local-ai-processing/) , qui fonctionne aussi hors ligne. Désactivez Cloud Sync et la conversation n’existe que sur l’appareil qui l’a capturée.

De quel matériel ai-je besoin pour exécuter des outils d’IA locale ?

Moins que vous ne le pensez, même si cela varie selon le modèle et l’appareil. La dictée et la reconnaissance vocale fonctionnent bien sur les téléphones et ordinateurs portables récents. Pour le chat avec un LLM local, la règle générale est de prévoir 8 Go de RAM pour les petits modèles et 16 Go ou plus pour les modèles intermédiaires ; les Macs Apple Silicon et les nouveaux PC IA gèrent bien les deux. Gartner prévoyait que les PC compatibles avec l’IA représenteraient 31% du marché des PC en 2025. Côté matériel, cela devient donc rapidement la norme.

Les outils d’IA locale sont-ils gratuits ?

Pour la plupart. Jan, Ollama, PrivateGPT et PocketPal AI sont entièrement gratuits. Obsidian est gratuit pour un usage personnel et commercial, avec Sync en option à partir de $4/month facturé annuellement. VoiceInk s’achète une seule fois, entre $25 et $49 pour Mac. Hedy propose une offre gratuite avec 5 heures de sessions par mois ; Pro coûte $12.99/month ou $99.99/year.

Quels outils d’IA locale fonctionnent entièrement hors ligne ?

Jan, Ollama et PocketPal AI continuent de fonctionner sans connexion une fois leurs modèles téléchargés. VoiceInk transcrit hors ligne avec ses modèles locaux et Obsidian fonctionne hors ligne par défaut. PrivateGPT fonctionne hors ligne lorsqu’il est associé à un serveur d’inférence local. Hedy fonctionne hors ligne lorsque le Traitement IA local est activé sur une machine capable d’exécuter des modèles locaux et associé à un moteur de reconnaissance vocale sur l’appareil.

Lesquels de ces outils fonctionnent sous Windows ?

Hedy, Jan, Ollama et Obsidian ont tous des applications Windows, et PrivateGPT peut être auto-hébergé sous Windows. VoiceInk est réservé au Mac (avec une application iOS distincte) et PocketPal AI est réservé aux appareils mobiles. Une remarque propre à Windows : le Traitement IA local de Hedy nécessite un GPU compatible Vulkan ; la reconnaissance vocale sur l’appareil fonctionne dans tous les cas.

Pourquoi la confidentialité est-elle importante pour les outils de productivité IA ?

Parce que les données professionnelles sont précisément ce que les gens fournissent à ces outils. Cyberhaven a constaté que 27.4% des données saisies par les employés dans des outils d’IA au Q2 2024 étaient sensibles, soit environ 2.6 fois la proportion de l’année précédente. Lors de l’incident Samsung de 2023 souvent cité, un employé a notamment téléversé l’enregistrement d’une réunion interne dans ChatGPT pour en générer le compte rendu. Le traitement local élimine ce mode de défaillance : ce qui ne quitte jamais votre appareil ne peut pas finir dans les journaux de quelqu’un d’autre.
