Melhores ferramentas de IA local para produtividade em 2026

Compare sete ferramentas de IA local para produtividade em 2026: IA para reuniões no dispositivo, ditado, chat no desktop, notas e perguntas e respostas sobre documentos, com limites claros para a nuvem.

Publicado por Julian Pscheid · 1 de julho de 2026
Ler este artigo online: https://www.hedy.ai/pt/post/best-local-ai-tools-for-productivity/

Homem falando durante uma reunião virtual em sua mesa de trabalho em casa, com a videochamada em um monitor externo, ferramentas de produtividade na tela do laptop e um iPhone virado para baixo brilhando suavemente ao lado do teclado

Resposta rápida Ferramentas de IA local rodam modelos de IA no seu próprio dispositivo em vez de usar a nuvem de um fornecedor. A melhor depende do trabalho. Para reuniões, Hedy roda reconhecimento de fala no dispositivo em todas as plataformas e, em máquinas capazes de rodar modelos locais poderosos, também pode mover todo o pipeline de IA para o dispositivo. Para ditado, VoiceInk transcreve inteiramente no seu Mac. Para chat no desktop, Jan roda modelos abertos offline. Para notas, Obsidian mantém tudo em arquivos locais. Cada ferramenta abaixo assume um trabalho, e juntas elas oferecem um caminho local e privado para cada parte de um dia de trabalho.

O que conta como IA local?

IA local significa que o modelo roda no seu hardware: seu laptop, seu celular, seu servidor. Seus dados são processados onde já estão, em vez de serem enviados para a nuvem de um fornecedor.

Na prática, é um espectro, e fingir o contrário é como as ferramentas induzem você ao erro. Há três posições transparentes:

• Totalmente local. O modelo roda no seu dispositivo e a ferramenta funciona sem conexão. Jan, Ollama, PocketPal AI e o modo padrão do VoiceInk estão aqui.

• Local por padrão, nuvem opcional. O caminho privado é o padrão; você pode conectar um provedor de nuvem se quiser. A transcrição opcional na nuvem do VoiceInk funciona assim.

• Híbrida com garantias locais. Parte do pipeline é local por definição, parte usa a nuvem, e a ferramenta deixa claro qual é qual. Hedy está neste grupo: o reconhecimento de fala roda no dispositivo por padrão em todas as plataformas, e máquinas capazes de rodar modelos locais poderosos podem mover todo o pipeline para o dispositivo.

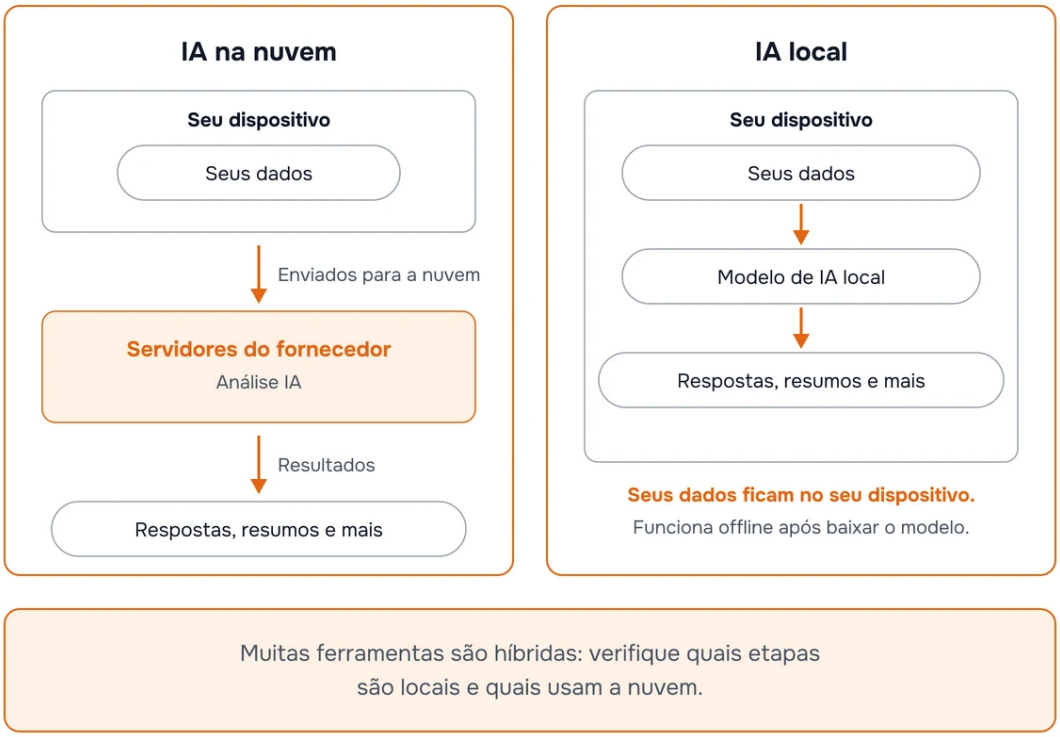

Diagrama comparando a IA na nuvem, em que seus dados são enviados aos servidores do fornecedor para processamento, com a IA local, em que seus dados, o modelo de IA e os resultados ficam no seu dispositivo

Usar a nuvem não desqualifica uma ferramenta desta lista. Ser vaga sobre isso, sim. Cada opção abaixo informa quais etapas rodam na sua máquina.

Por que a IA local está em alta?

Duas curvas se cruzaram. Os dispositivos estão ficando mais poderosos, enquanto os modelos ficam menores e mais inteligentes. Em algum momento do último ano, modelos que cabem em um laptop ou celular recente ficaram fortes o suficiente para fazer trabalho de verdade: transcrever uma conversa, resumir um documento, redigir uma resposta. Vimos essa virada acontecer em primeira mão enquanto construíamos o pipeline local do Hedy e detalhamos o processo em nosso artigo técnico aprofundado sobre IA local (/pt/post/local-ai-engineering-deep-dive-hedy-3-2/) .

A preocupação com o destino dos dados de IA aumentou no mesmo período. Pew Research descobriu em June 2026 (https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/) que 71% dos adultos dos EUA esperam que o aumento do uso de IA

torne suas informações pessoais menos seguras; apenas 3% esperam o contrário. As pessoas não deixam de usar IA por causa dessa preocupação, elas continuam fornecendo material confidencial: a Cyberhaven mediu (https://info.cyberhaven.com/hubfs/Content%20PDF/Cyberhaven%20Q2%202024%20AI%20Adoption%20and%20Risk%20Report%20052024.pdf) que 27.4% dos dados inseridos por funcionários em ferramentas de IA no Q2 2024 eram confidenciais, acima de 10.7% um ano antes. O exemplo clássico que serve de alerta ainda é o incidente da Samsung em 2023 (https://www.engadget.com/three-samsung-employees-reportedly-leaked-sensitive-data-to-chatgpt-190221114.html) , em que um funcionário enviou uma reunião interna gravada ao ChatGPT para gerar uma ata.

Esta é nossa aposta, e este artigo parte dela: IA local é o caminho, não um nicho. A nuvem continuará na fronteira por muito tempo e, para a maioria das pessoas, seguirá como a melhor opção padrão para raciocínio difícil. Mas o nível mínimo local sobe a cada ano, o hardware já está no seu bolso e na sua mesa (a Gartner projeta (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 PCs com IA representarão 55% do mercado em 2026) e, quando um trabalho pode ser bem-feito no dispositivo, não há um bom argumento para enviar esses dados a outro lugar. "Onde isso roda?" está virando uma pergunta padrão para toda ferramenta de IA, assim como "isso é criptografado?" se tornou uma pergunta para serviços de mensagens. As ferramentas abaixo são para os trabalhos em que a resposta já pode ser "aqui mesmo".

Como escolhemos estas ferramentas

Cinco regras definiram esta lista:

• Uma ferramenta para cada trabalho. Uma stack de especialistas é melhor do que um monte de apps sobrepostos, cada um com uma cópia dos seus dados. Escolhemos uma opção local-first para cada trabalho de produtividade: reuniões, ditado, chat, execução de modelos, notas, documentos e dispositivos móveis.

• O funcionamento local precisa ser real. Cada ferramenta roda inteiramente no dispositivo ou documenta com precisão quais etapas rodam nele. Marketing vago de "privacidade em primeiro lugar" sem detalhes de arquitetura não foi suficiente.

• Ativa em 2026. Verificamos os preços, recursos e a atividade de projeto de cada ferramenta nas páginas e nos repositórios dos fornecedores em July 2026.

• Híbridas são permitidas, mas precisam deixar isso claro. Ferramentas que usam a nuvem em algumas etapas entram quando o limite é explícito e está sob o controle do usuário.

• Transparência total: Hedy é nosso produto. Esta é uma seleção publicada pelo proprietário, não um benchmark independente, e Hedy ocupa intencionalmente a posição de reuniões. Alternativas diretas para reuniões (Granola, Fireflies, modo de reuniões do Superwhisper) não estão ranqueadas aqui porque não somos árbitros neutros nessa categoria; escrevemos comparações separadas e fundamentadas, como nossos guias de alternativa ao Granola (/pt/post/granola-alternative/) e alternativa ao Fireflies (/pt/post/fireflies-alternative-hedy/) , para leitores que estão avaliando essas ferramentas.

Qual ferramenta de IA local é melhor para cada trabalho?

Comparação rápida

#	Ferramenta	Melhor para	Funcionamento local	Plataformas	Preço
1	Hedy (https://www.hedy.ai/)	Reuniões e conversas ao vivo	Reconhecimento de fala no dispositivo por padrão; pipeline totalmente local opcional em hardware adequado	iOS, Android, Mac, Windows, web	Grátis (5 hrs/mo); $12.99/mo ou $99.99/yr
2	VoiceInk (https://tryvoiceink.com/)	Ditado	Totalmente no dispositivo por padrão; provedores de nuvem opcionais	macOS (app iOS disponível)	$25 a $49 em compra única
3	Jan (https://jan.ai/)	Chat de IA no desktop	Totalmente local por padrão; APIs de nuvem opcionais	Windows, Mac, Linux	Grátis
4	Ollama (https://ollama.com/)	Rodar modelos localmente	Totalmente local; nível de nuvem opcional	Mac, Windows, Linux	Grátis; Pro $20/mo
5	Obsidian (https://obsidian.md/)	Notas e base de conhecimento	Arquivos Markdown locais; plugins podem usar modelos locais	Mac, Windows, Linux, iOS, Android	Grátis; Sync a partir de $4/mo
6	PrivateGPT (https://github.com/zylon-ai/private-gpt)	Perguntas e respostas sobre documentos privados	Auto-hospedado; você escolhe o servidor de inferência	Auto-hospedado (Mac, Linux, Windows)	Grátis
7	PocketPal AI (https://pocketpal.dev/)	Chat de IA no seu celular	Totalmente no dispositivo após o download do modelo	iOS, Android	Grátis

Fontes: preços do Hedy (/pt/pricing/) , preços do VoiceInk (https://tryvoiceink.com/pricing) , documentação do Jan (https://jan.ai/docs) , preços do Ollama (https://ollama.com/pricing) , preços do Obsidian (https://obsidian.md/pricing) , repositório do PrivateGPT (https://github.com/zylon-ai/private-gpt) , PocketPal AI (https://pocketpal.dev/) . Os preços foram verificados nas páginas dos fornecedores em July 2026 e mudam com frequência; confira novamente antes de comprar.

1. Hedy: reuniões e conversas ao vivo

Reuniões são onde os dados de trabalho mais confidenciais são ditos em voz alta: nomes de clientes, termos de negócios, detalhes de saúde, estratégia. Elas também são um fluxo de trabalho em que ferramentas populares de IA tendem a oferecer menos privacidade, porque muitos assistentes de reunião funcionam como um bot que entra na sua chamada e envia o áudio para um servidor.

Hedy (https://www.hedy.ai/) adota a abordagem oposta. O reconhecimento de fala roda no dispositivo por padrão em todas as plataformas (Whisper em todas elas, além do mecanismo Nemotron com identificação de interlocutores no dispositivo), então, com os mecanismos padrão, seu áudio nunca sai do dispositivo e nenhum bot entra nas suas chamadas. Além da transcrição, Hedy funciona como um coach de reuniões em tempo real: apresenta sugestões, respostas e perguntas que vale a pena fazer enquanto a conversa ainda está acontecendo e, depois, produz resumos, destaques e tarefas.

Os padrões do Hedy seguem uma divisão deliberada. O reconhecimento de fala roda localmente em todos os dispositivos porque consegue fazer isso, em qualquer lugar, com alta qualidade. A análise de IA roda na nuvem por padrão porque queremos que Hedy funcione no maior número possível de dispositivos com os melhores resultados que podemos entregar, e o que é enviado nunca é armazenado nem usado para treinar modelos. Mas o potencial local é muito maior do que o padrão: qualquer máquina capaz de rodar modelos locais poderosos pode mover todo o pipeline da reunião para o dispositivo com o Processamento de IA Local (/pt/help/local-ai-processing/) . Transcrições, resumos, notas detalhadas, respostas de chat e sugestões ao vivo acontecem na máquina que capturou o áudio, offline se você quiser, em Macs com Apple Silicon, máquinas Windows com GPUs adequadas e iPhones e iPads recentes. Desative o Cloud Sync e a conversa existirá apenas no dispositivo que a capturou, de ponta a ponta. Hedy 3.2 tornou isso possível, como explicamos em IA local para reuniões (/pt/post/local-ai-meetings-hedy-3-2/) . Você também pode desativar a análise de IA na nuvem (/pt/help/cloud-ai-analysis-privacy-control/) mantendo a transcrição completa, escolher residência de dados na UE (/pt/post/eu-data-residency/) e revisar todas as configurações em nosso guia de configurações de privacidade (/pt/post/ai-meeting-privacy-settings-explained/) .

O modo padrão não é totalmente local; o modo totalmente local está disponível quando seu hardware e seus requisitos de privacidade pedem isso e, de qualquer forma, você tem um limite documentado que controla. Hedy também concluiu um exame independente SOC 2 Type I de seus controles de segurança e uma avaliação independente HIPAA como Business Associate em April 2026, além de ter avaliação de 4.8 de 5 em 500+ avaliações feitas por 30,000+ usuários.

Preço: grátis para 5 horas de sessões por mês, sem exigir cartão de crédito. O Pro custa $12.99/mês ou $99.99/ano; uma licença vitalícia de $299 cobre tudo permanentemente.

Melhor para: qualquer pessoa cujas conversas sejam a informação mais confidencial que, de outra forma, enviaria para uma ferramenta na nuvem: consultores, representantes de vendas, jornalistas, pacientes e gestores.

2. VoiceInk: ditado

Ditado é o ganho silencioso de produtividade da IA local. A maioria das pessoas fala muito mais rápido do que digita, e um bom modelo no dispositivo agora transcreve com precisão suficiente para não haver motivo para enviar sua voz.

VoiceInk (https://tryvoiceink.com/) é o especialista em ditado para Mac. Ele transcreve sua fala no dispositivo usando modelos locais Whisper e Parakeet, funciona offline e mantém o histórico de transcrições na sua máquina com exclusão automática configurável. Um dicionário pessoal lida com nomes e jargão, e modos por app ajustam a formatação para e-mail, chat ou código. É open source (https://github.com/Beingpax/VoiceInk) , então as alegações de privacidade podem ser verificadas em vez de aceitas sem prova; o projeto tem 5,500+ estrelas no GitHub, e o desenvolvedor informa mais de 200,000 downloads. Melhorias opcionais na nuvem e provedores de transcrição na nuvem estão disponíveis, mas são opcionais e identificados como tal.

Por que VoiceInk em vez dos mais conhecidos Superwhisper ou MacWhisper? Especialização. Ambos se expandiram para a gravação de reuniões (Superwhisper (https://superwhisper.com/meeting-transcription) , MacWhisper (https://macwhisper.helpscoutdocs.com/article/30-record-meetings)), o que os coloca em uma categoria diferente (e em outra conversa sobre privacidade) do ditado puro. O app do VoiceInk para Mac se dedica a um trabalho: digitar o que você diz. Seu app iOS separado lista transcrição de reuniões e aulas entre os casos de uso, mas não há captura de chamadas, identificação de interlocutores nem geração de resumos; o produto é ditado. Para conversas ao vivo com outras pessoas, esse é o trabalho do Hedy, e os dois funcionam lado a lado.

Preço: compra única de $25 (um Mac) a $49 (três Macs), com atualizações vitalícias incluídas. Um app iOS separado é gratuito no momento.

Melhor para: escritores, desenvolvedores e qualquer pessoa que responda a cinquenta e-mails por dia e não queira uma assinatura ou dependência da nuvem para ter esse benefício.

3. Jan: chat de IA no desktop

Se ChatGPT é sua ferramenta padrão para pensar, mas você prefere não enviar cada ideia ainda incompleta para um servidor, Jan (https://jan.ai/) é a alternativa local com menos atrito. Baixe o app, escolha um modelo aberto (Llama, Qwen, Gemma e dezenas de outros no formato GGUF) e converse. Não é preciso criar uma conta e, depois que o modelo é baixado, ele funciona com o Wi-Fi desligado.

Jan é open source (https://github.com/janhq/jan) , com 43,000+ estrelas no GitHub e 5.9 milhões de downloads informados. Duas escolhas de design fazem dele mais do que um brinquedo. Primeiro, ele disponibiliza uma API compatível com OpenAI no localhost, para que outros apps desta lista (e seus

próprios scripts) possam usar seu modelo local como um backend substituto. Segundo, é transparente sobre o uso híbrido: você pode adicionar chaves de API para provedores de nuvem quando quiser a qualidade de modelos de fronteira, e Jan deixa claro que essas conversas saem da sua máquina de acordo com os termos desse provedor.

Preço: grátis.

Melhor para: rascunhos e brainstorming do dia a dia em que "bom o suficiente localmente" é melhor do que "excelente, mas enviado".

4. Ollama: execução local de modelos

Ollama (https://ollama.com/) é a infraestrutura do movimento de IA local. É uma ferramenta de linha de comando e um serviço em segundo plano que baixa e roda modelos abertos com um comando (ollama run llama3) e depois os disponibiliza por meio de uma API REST local usada por um grande ecossistema de apps. Se você usa várias ferramentas de IA local, é bem possível que todas possam compartilhar um backend do Ollama, em vez de cada uma incluir seu próprio runtime de modelo.

É open source (https://github.com/ollama/ollama) , tem 176,000+ estrelas no GitHub, e a empresa informou atender 8.9 milhões de desenvolvedores em July 2026. Tudo que é local é gratuito e ilimitado; planos pagos na nuvem (a partir de $20/mês) estão disponíveis para rodar modelos maiores no hardware do Ollama, mantidos separados do caminho local.

Preço: grátis para uso local ilimitado; planos opcionais na nuvem a partir de $20/mês.

Melhor para: usuários avançados e qualquer pessoa que queira um backend de modelos compartilhado e programável por trás de suas outras ferramentas.

5. Obsidian: notas e conhecimento pessoal

Suas notas são um mapa de tudo em que você está trabalhando, o que faz de uma ferramenta de notas exclusiva da nuvem uma das apostas mais concentradas em privacidade que as pessoas fazem sem pensar. Obsidian (https://obsidian.md/) evita essa aposta por completo: as notas são arquivos Markdown simples em uma pasta no seu dispositivo. O app funciona offline, os arquivos sobrevivem à empresa e nada é sincronizado a menos que você configure uma sincronização (o serviço Sync opcional, a partir de $4/mês com cobrança anual, tem criptografia de ponta a ponta).

A camada de IA vem de plugins da comunidade e pode continuar local. Smart Connections (https://github.com/brianpetro/obsidian-smart-connections) cria embeddings do seu vault no dispositivo para buscar notas relacionadas, e Copilot for Obsidian (https://github.com/logancyang/obsidian-copilot) adiciona chat sobre suas notas com suporte a backends locais como Ollama, que é onde uma stack local compartilhada começa a valer a pena. Com um modelo local por trás, "conversar com minhas notas" não envolve nenhum terceiro. Os dois plugins também oferecem suporte a provedores de nuvem, então leia suas configurações com o mesmo cuidado que teria com qualquer ferramenta híbrida.

Preço: grátis para uso pessoal e comercial; Sync opcional a partir de $4/mês com cobrança anual.

Melhor para: qualquer pessoa que esteja construindo uma base de conhecimento pessoal duradoura e queira busca com IA sobre ela sem enviar seus pensamentos.

6. PrivateGPT: perguntas e respostas sobre documentos privados

"Conversar com seus documentos" foi o primeiro caso de uso decisivo da IA privada, e PrivateGPT (https://github.com/zylon-ai/private-gpt) é sua principal referência open source: 57,000+ estrelas no GitHub e um relançamento completo em June 2026, quando a Zylon lançou a v1.0 (https://www.zylon.ai/

resources/blog/introducing-privategpt-1-0-the-open-source-application-backend-for-private-ai) e incorporou dois anos de desenvolvimento empresarial privado de volta ao repositório público.

PrivateGPT 1.0 é um backend de aplicação: cuida da ingestão de documentos, recuperação com citações, chat e orquestração de ferramentas. Como a Zylon explica no anúncio da 1.0, "PrivateGPT 1.0 não roda modelos por conta própria": você o direciona para o servidor de inferência que escolher. É isso que torna possível uma configuração totalmente auto-hospedada aqui: combine-o com Ollama ou llama.cpp na mesma máquina e o processamento dos documentos continuará dentro da infraestrutura que você controla. Também é a opção desta lista mais adequada para uma pequena equipe que queira criar um assistente privado compartilhado no próprio servidor.

Um aviso justo sobre a adequação: esta é a ferramenta mais técnica daqui. Não há instalador de um clique, e as alternativas mais refinadas para consumidores nesta categoria migraram em grande parte para recursos de transcrição de reuniões, por isso não entraram nesta lista.

Preço: grátis; você fornece o hardware.

Melhor para: usuários técnicos e pequenas equipes que querem perguntas e respostas sobre documentos funcionando inteiramente na infraestrutura que controlam.

7. PocketPal AI: IA local no seu celular

Foi no celular que a IA local se tornou uma realidade para pessoas sem conhecimento técnico, e PocketPal AI (https://pocketpal.dev/) é a maneira mais simples de experimentá-la. É um app gratuito e open source (https://github.com/a-ghorbani/pocketpal-ai) para iOS e Android que baixa pequenos modelos abertos (Llama, Qwen, Gemma, Phi) e os roda inteiramente no dispositivo. Sem conta, sem telemetria, e continua respondendo no modo avião.

O projeto informa 1.2 milhão de downloads, e sua comunidade enviou mais de 7,000 resultados de benchmark no dispositivo em 100+ modelos de celular, o que também funciona como resposta prática para "isso vai rodar no meu celular?" (Regra prática: iPhones recentes e celulares Android intermediários ou melhores lidam confortavelmente com os modelos pequenos.) Não vai igualar um modelo de fronteira em raciocínio difícil, mas, para rascunhos rápidos e traduções, é uma demonstração funcional de que uma IA útil agora cabe no bolso sem envolver um servidor. É a mesma aposta que a Apple fez ao colocar o processamento no dispositivo no centro do Apple Intelligence.

Preço: grátis.

Melhor para: experimentar IA local sem nenhuma configuração e usá-la como assistente de bolso offline para tarefas rápidas.

Como criar uma stack de IA local com estas ferramentas?

Comece pelo trabalho que envolve seus dados mais confidenciais. Para muitos profissionais, são reuniões e conversas, a função coberta pelo Hedy; para outros, são documentos ou notas. Mova esse único fluxo de trabalho para uma ferramenta local-first e você já terá obtido a maior parte do benefício de privacidade em uma única etapa.

Depois, adicione especialistas um de cada vez: uma ferramenta de ditado, uma de chat e uma de notas. Evite sobreposição. Duas ferramentas fazendo o mesmo trabalho significam duas cópias dos mesmos dados confidenciais em dois lugares, o oposto do objetivo. A stack formada por esta lista compartilha bem a infraestrutura: Ollama disponibiliza modelos, Jan, os plugins do Obsidian e PrivateGPT podem funcionar sobre ele, VoiceInk digita o que você diz, e Hedy cobre tudo que é falado com outras pessoas.

Por fim, verifique em vez de confiar. Desative o Wi-Fi e veja o que continua funcionando. Leia a página de privacidade e procure detalhes de cada etapa (áudio, transcrição, análise, armazenamento), em vez de adjetivos. Todas as ferramentas acima passam nesse teste de leitura; é por isso que estão aqui. Para a categoria de reuniões especificamente, nosso guia sobre como usar notas de reunião com IA de forma privada (/pt/post/how-to-use-ai-meeting-notes-privately/) analisa essa verificação em detalhes: o que sai em cada etapa e por que um teste offline revela menos do que você imagina.

Perguntas frequentes

Qual é a melhor ferramenta de IA local para produtividade?

Depende do trabalho. Para reuniões e conversas ao vivo, Hedy roda reconhecimento de fala no dispositivo em todas as plataformas, e máquinas capazes de rodar modelos locais poderosos também podem mover todo o pipeline de IA para o dispositivo. Para ditado, VoiceInk transcreve inteiramente no seu Mac. Para chat de IA no desktop, Jan roda modelos abertos localmente sem exigir uma conta. Para notas, Obsidian mantém tudo em arquivos locais. A maioria das pessoas acaba combinando duas ou três ferramentas, cada uma responsável por um trabalho.

O que significa "IA local"?

É um espectro, não um rótulo de sim ou não. Ferramentas totalmente locais como Jan, Ollama e PocketPal rodam o modelo no seu hardware e funcionam offline. Ferramentas híbridas rodam parte do pipeline localmente: Hedy, por exemplo, roda reconhecimento de fala no dispositivo por padrão e oferece um modo opcional de Processamento de IA Local que também mantém a etapa de análise no seu dispositivo. O importante é saber quais etapas ficam na sua máquina, quais não ficam e se você pode controlar isso.

As ferramentas de IA local são tão boas quanto a IA na nuvem?

Para muitas tarefas de produtividade, sim. O reconhecimento de fala no dispositivo agora é excelente, embeddings locais lidam bem com a busca em notas e pequenos modelos abertos dão conta de resumos e rascunhos. Para raciocínio no nível dos modelos de fronteira, os modelos na nuvem ainda lideram. É por isso que várias ferramentas desta lista são híbridas com transparência: locais onde a privacidade mais importa (seu áudio, seus arquivos), na nuvem onde a qualidade mais importa, com uma opção para funcionar de forma totalmente local quando você precisar.

Hedy é totalmente local?

Os padrões do Hedy seguem uma divisão deliberada. O reconhecimento de fala roda no dispositivo em todas as plataformas (Whisper em todas elas, além do Nemotron com identificação de interlocutores no dispositivo), então, com os mecanismos padrão, seu áudio nunca sai do dispositivo e nenhum bot entra nas suas chamadas. A análise de IA roda na nuvem por padrão para que Hedy funcione no maior número possível de dispositivos com resultados de alta qualidade. Mas qualquer máquina capaz de rodar modelos locais poderosos (Macs com Apple Silicon, máquinas Windows com GPUs adequadas, iPhones e iPads recentes) pode mover todo o pipeline para o dispositivo com o modo opcional de Processamento de IA Local (/pt/help/local-ai-processing/) , que também funciona offline. Desative o Cloud Sync e a conversa existirá apenas no dispositivo que a capturou.

De qual hardware preciso para rodar ferramentas de IA local?

Menos do que você imagina, embora isso varie conforme o modelo e o dispositivo. Ditado e reconhecimento de fala rodam bem em celulares e laptops recentes. Para chat com LLM local, a regra prática é 8 GB de RAM para modelos pequenos e 16 GB ou mais para modelos intermediários; Macs com Apple Silicon e PCs com IA mais recentes lidam bem com ambos. A Gartner projetou que PCs capazes de rodar IA representariam 31% do mercado de PCs em 2025, então o hardware para isso está rapidamente se tornando o padrão.

As ferramentas de IA local são gratuitas?

Em sua maioria. Jan, Ollama, PrivateGPT e PocketPal AI são totalmente gratuitos. Obsidian é gratuito para uso pessoal e comercial, com Sync opcional a partir de $4/mês com cobrança anual. VoiceInk é uma compra única entre $25 e $49 para Mac. Hedy tem um plano gratuito com 5 horas de sessões por mês; o Pro custa $12.99/mês ou $99.99/ano.

Quais ferramentas de IA local funcionam totalmente offline?

Jan, Ollama e PocketPal AI continuam funcionando sem conexão depois que seus modelos são baixados. VoiceInk transcreve offline com seus modelos locais, e Obsidian funciona offline por padrão. PrivateGPT funciona offline quando combinado com um servidor de inferência local. Hedy funciona offline quando o Processamento de IA Local está ativado em uma máquina capaz de rodar modelos locais, combinado com um mecanismo de fala no dispositivo.

Quais dessas ferramentas rodam no Windows?

Hedy, Jan, Ollama e Obsidian têm apps para Windows, e PrivateGPT pode ser auto-hospedado no Windows. VoiceInk é exclusivo para Mac (com um app iOS separado), e PocketPal AI é exclusivo para dispositivos móveis. Uma observação específica sobre Windows: o Processamento de IA Local do Hedy exige uma GPU compatível com Vulkan; o reconhecimento de fala no dispositivo funciona de qualquer forma.

Por que a privacidade importa para ferramentas de produtividade com IA?

Porque dados de trabalho são exatamente o que as pessoas inserem nessas ferramentas. A Cyberhaven descobriu que 27.4% dos dados inseridos por funcionários em ferramentas de IA no Q2 2024 eram confidenciais, cerca de 2.6 vezes a proporção de um ano antes. O muito citado incidente da Samsung em 2023 incluiu um funcionário que enviou uma reunião interna gravada ao ChatGPT para gerar uma ata. O processamento local elimina esse modo de falha: o que nunca sai do seu dispositivo não pode acabar nos registros de outra pessoa.
