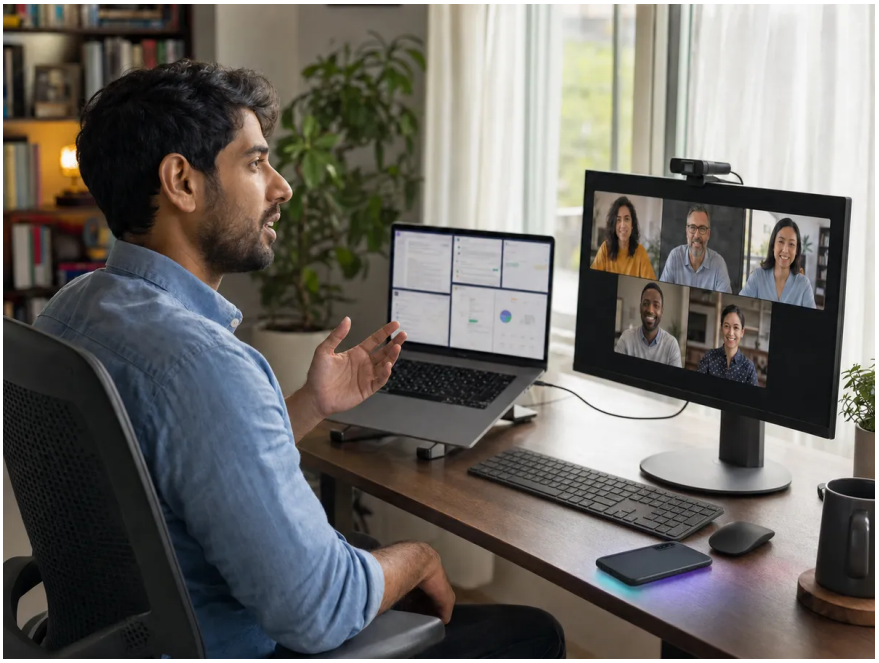
Las mejores herramientas de IA local para la productividad en 2026

Compare siete herramientas de IA local para la productividad en 2026: IA para reuniones en el dispositivo, dictado, chat de escritorio, notas y preguntas y respuestas sobre documentos, con límites claros respecto a la nube.

Publicado por Julian Pscheid · 1 de julio de 2026
Leer este artículo en línea: https://www.hedy.ai/es/post/best-local-ai-tools-for-productivity/



Hombre hablando durante una reunión virtual en el escritorio de su casa, con la videollamada en un monitor externo, herramientas de productividad en la pantalla de su portátil y un iPhone boca abajo que brilla suavemente junto al teclado

Respuesta rápida Las herramientas de IA local ejecutan modelos de IA en su propio dispositivo en lugar de hacerlo en la nube de un proveedor. La mejor depende del trabajo. Para reuniones, Hedy ejecuta el reconocimiento de voz en el dispositivo en todas las plataformas, y en máquinas capaces de ejecutar modelos locales potentes también puede trasladar todo el flujo de IA al dispositivo. Para dictado, VoiceInk transcribe todo en su Mac. Para chat de escritorio, Jan ejecuta modelos abiertos sin conexión. Para notas, Obsidian guarda todo en archivos locales. Cada una de las herramientas siguientes se ocupa de un trabajo y, juntas, ofrecen una vía local y privada para cada parte de una jornada laboral.

¿Qué se considera IA local?

La IA local significa que el modelo se ejecuta en su hardware: su portátil, su teléfono o su servidor. Sus datos se procesan donde ya están, en lugar de subirse a la nube de un proveedor.

En la práctica, es un espectro, y fingir lo contrario es la forma en que algunas herramientas inducen a error. Hay tres posturas honestas:

• Completamente local. El modelo se ejecuta en su dispositivo y la herramienta funciona con la red desconectada. Jan, Ollama, PocketPal AI y el modo predeterminado de VoiceInk están en este grupo.

• Local por defecto, nube opcional. La vía privada es la predeterminada; puede conectar un proveedor en la nube si así lo decide. La transcripción opcional en la nube de VoiceInk funciona de esta manera.

• Híbrida con garantías locales. Parte del flujo es local por diseño, otra parte usa la nube y la herramienta indica con claridad cuál es cuál. Hedy pertenece a este grupo: el reconocimiento de voz se ejecuta en el dispositivo por defecto en todas las plataformas, y las máquinas capaces de ejecutar modelos locales potentes pueden trasladar todo el flujo al dispositivo.

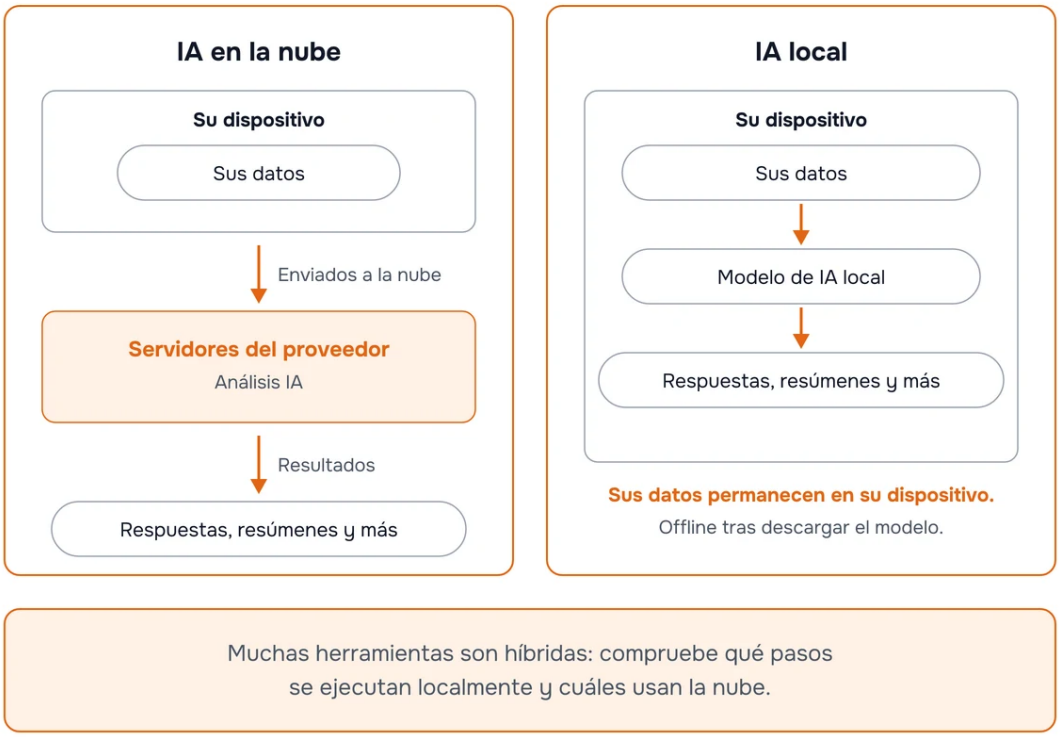


Diagrama que compara la IA en la nube, donde sus datos se envían a los servidores del proveedor para su procesamiento, con la IA local, donde sus datos, el modelo de IA y los resultados permanecen en su dispositivo

Usar la nube no descalifica a una herramienta de esta lista. Ser imprecisa al respecto, sí. Cada opción siguiente indica qué pasos se ejecutan en su máquina.

¿Por qué la IA local está ganando protagonismo?

Dos curvas se cruzaron. Los dispositivos son cada vez más potentes, mientras los modelos se vuelven más pequeños e inteligentes. En algún momento del último año, los modelos que caben en un portátil o teléfono reciente adquirieron la capacidad suficiente para hacer trabajo real: transcribir una conversación, resumir un documento o redactar una respuesta. Vimos ese cruce de primera mano mientras desarrollábamos el flujo local de Hedy y explicamos los detalles en nuestro análisis técnico a fondo de la IA local (/es/post/local-ai-engineering-deep-dive-hedy-3-2/) .

La preocupación por el destino de los datos de IA creció durante el mismo periodo. Pew Research descubrió en June 2026 (https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/) que el 71% de los adultos estadounidenses espera que el

aumento del uso de IA reduzca la seguridad de su información personal; solo el 3% espera lo contrario. Las personas no dejan de usar IA por esa preocupación, sino que siguen introduciendo material sensible: Cyberhaven midió (https://info.cyberhaven.com/hubfs/Content%20PDF/Cyberhaven%20Q2%202024%20AI%20Adoption%20and%20Risk%20Report%20052024.pdf) que el 27.4% de los datos que los empleados introdujeron en herramientas de IA en Q2 2024 eran sensibles, frente al 10.7% un año antes. El ejemplo más conocido sigue siendo el incidente de Samsung de 2023 (https://www.engadget.com/three-samsung-employees-reportedly-leaked-sensitive-data-to-chatgpt-190221114.html) , en el que un empleado subió una reunión interna grabada a ChatGPT para generar un acta.

Esta es nuestra apuesta, y este artículo parte de ella: la IA local es la dirección, no un nicho. La nube conservará la vanguardia durante mucho tiempo y, para la mayoría de las personas, seguirá siendo la mejor opción predeterminada para el razonamiento difícil. Pero el nivel mínimo de lo local mejora cada año, el hardware ya está en su bolsillo y sobre su escritorio (Gartner proyecta (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 los PC con IA representarán el 55% del mercado en 2026) y, cuando un trabajo puede hacerse bien en el dispositivo, no hay un buen motivo para enviar esos datos a otro lugar. "¿Dónde se ejecuta esto?" se está convirtiendo en una pregunta habitual para toda herramienta de IA, del mismo modo que "¿está cifrado?" se convirtió en una pregunta habitual para la mensajería. Las herramientas siguientes sirven para los trabajos en los que la respuesta ya puede ser "aquí mismo".

Cómo elegimos estas herramientas

Cinco reglas dieron forma a esta lista:

• Una herramienta por trabajo. Un conjunto de especialistas supera a un montón de apps que se solapan y guardan cada una una copia de sus datos. Elegimos una opción que prioriza lo local para cada trabajo de productividad: reuniones, dictado, chat, ejecución de modelos, notas, documentos y móvil.

• Lo local debe ser real. Cada herramienta se ejecuta completamente en el dispositivo o documenta con precisión qué pasos lo hacen. El marketing impreciso de "privacidad primero" sin detalles de arquitectura no cumplió los requisitos.

• Activa en 2026. Verificamos los precios, las funciones y la actividad del proyecto de cada herramienta mediante las páginas y los repositorios de sus proveedores en July 2026.

• Se permiten híbridas, pero deben declararlo. Se incluyen herramientas que usan la nube para algunos pasos cuando el límite es explícito y está bajo el control del usuario.

• Divulgación completa: Hedy es nuestro producto. Esta es una selección publicada por el propietario, no una evaluación independiente, y Hedy ocupa deliberadamente el puesto de reuniones. Las alternativas directas entre herramientas de reuniones (Granola, Fireflies y el modo para reuniones de Superwhisper) no se clasifican aquí porque no somos árbitros neutrales en esa categoría; hemos escrito comparativas separadas y con fuentes, como nuestras guías de alternativa a Granola (/es/post/granola-alternative/) y alternativa a Fireflies (/es/post/fireflies-alternative-hedy/) , para quienes estén valorando esas herramientas.

¿Qué herramienta de IA local es mejor para cada trabajo?

Comparación rápida

| Herramienta | Mejor para | Grado de funcionamiento local | Plataformas | Precio

1 | Hedy (https://www.hedy.ai/) | Reuniones y conversaciones en vivo | Reconocimiento de voz en el dispositivo por defecto; flujo completamente local opcional en hardware capaz | iOS, Android, Mac, Windows, web | Free (5 hrs/mo); $12.99/mo or $99.99/yr
2 | VoiceInk (https://tryvoiceink.com/) | Dictado | Completamente en el dispositivo por defecto; proveedores en la nube opcionales | macOS (iOS app available) | $25 to $49 one-time
3 | Jan (https://jan.ai/) | Chat de IA en el escritorio | Completamente local por defecto; API en la nube opcionales | Windows, Mac, Linux | Free
4 | Ollama (https://ollama.com/) | Ejecutar modelos localmente | Completamente local; nivel en la nube opcional | Mac, Windows, Linux | Free; Pro $20/mo
5 | Obsidian (https://obsidian.md/) | Notas y base de conocimiento | Archivos Markdown locales; los complementos pueden usar modelos locales | Mac, Windows, Linux, iOS, Android | Free; Sync from $4/mo
6 | PrivateGPT (https://github.com/zylon-ai/private-gpt) | Preguntas y respuestas privadas sobre documentos | Autoalojado; usted elige el servidor de inferencia | Self-hosted (Mac, Linux, Windows) | Free
7 | PocketPal AI (https://pocketpal.dev/) | Chat de IA en su teléfono | Completamente en el dispositivo tras descargar el modelo | iOS, Android | Free

Fuentes: precios de Hedy (/es/pricing/) , precios de VoiceInk (https://tryvoiceink.com/pricing) , documentación de Jan (https://jan.ai/docs) , precios de Ollama (https://ollama.com/pricing) , precios de Obsidian (https://obsidian.md/pricing) , repositorio de PrivateGPT (https://github.com/zylon-ai/private-gpt) , PocketPal AI (https://pocketpal.dev/) . Los precios se verificaron en las páginas de los proveedores en July 2026 y cambian con frecuencia; vuelva a comprobarlos antes de comprar.

1. Hedy: reuniones y conversaciones en vivo

Las reuniones son donde se expresan en voz alta los datos de trabajo más sensibles: nombres de clientes, condiciones de acuerdos, detalles de salud y estrategia. También son un flujo de trabajo en el que las herramientas populares de IA tienden a ofrecer menos privacidad, porque muchos asistentes de reuniones funcionan como un bot que se une a la llamada y envía el audio a un servidor.

Hedy (https://www.hedy.ai/) adopta el enfoque opuesto. El reconocimiento de voz se ejecuta en el dispositivo por defecto en todas las plataformas (Whisper en todas partes, además del motor Nemotron con etiquetas de hablantes en el dispositivo), así que con los motores predeterminados su audio nunca sale de su dispositivo y ningún bot se une a sus llamadas. Sobre la transcripción, Hedy funciona como coach de reuniones en tiempo real: ofrece sugerencias, respuestas y preguntas que conviene hacer mientras la conversación sigue en curso, y después genera resúmenes, puntos destacados y tareas pendientes.

La configuración por defecto de Hedy es una división deliberada. El reconocimiento de voz se ejecuta localmente en todos los dispositivos porque puede hacerlo en todas partes y con alta calidad. El análisis de IA se ejecuta en la nube por defecto porque queremos que Hedy funcione en tantos dispositivos como sea posible con los mejores resultados que podemos ofrecer, y lo que se envía nunca se almacena ni se usa para entrenar modelos. Sin embargo, el límite superior de lo local es mucho mayor que la configuración predeterminada: cualquier máquina capaz de ejecutar modelos locales potentes puede trasladar todo el flujo de la reunión al dispositivo con el Procesamiento de IA Local (/es/help/local-ai-processing/) . Las transcripciones, los resúmenes, las notas detalladas, las respuestas de chat y las sugerencias en vivo se generan en la máquina que capturó el audio, sin conexión si así lo desea, en Macs con Apple Silicon, máquinas Windows con GPU capaces y iPhones y iPads recientes. Desactive Cloud Sync y la conversación solo existirá en el dispositivo que la capturó, de principio a fin. Hedy 3.2 lo hizo posible, y lo explicamos en IA local para reuniones (/es/post/local-ai-meetings-hedy-3-2/) . También puede desactivar el análisis de IA en la nube (/es/help/cloud-ai-analysis-privacy-control/) mientras mantiene la transcripción completa, elegir la residencia de datos en la UE (/es/post/eu-data-residency/) y revisar cada ajuste en nuestra guía de configuración de privacidad (/es/post/ai-meeting-privacy-settings-

explained/) .

El modo predeterminado no es completamente local; el modo completamente local está disponible cuando su hardware y sus requisitos de privacidad lo exigen, y en ambos casos obtiene un límite documentado que usted controla. Hedy también completó un examen independiente SOC 2 Type I de sus controles de seguridad y una evaluación HIPAA independiente como Asociado Comercial en April 2026, y tiene una valoración de 4.8 out of 5 en 500+ reseñas de 30,000+ usuarios.

Precio: gratis por 5 horas de sesiones al mes, sin tarjeta de crédito. Pro cuesta $12.99 al mes o $99.99 al año; una licencia lifetime de $299 cubre todo de forma permanente.

Mejor para: cualquier persona cuyas conversaciones sean lo más sensible que, de otro modo, introduciría en una herramienta en la nube: consultores, representantes de ventas, periodistas, pacientes y gerentes.

2. VoiceInk: dictado

El dictado es la discreta ventaja de productividad de la IA local. La mayoría de las personas habla mucho más rápido de lo que escribe, y un buen modelo en el dispositivo ya transcribe con tanta precisión que no hay motivo para subir su voz.

VoiceInk (https://tryvoiceink.com/) es el especialista en dictado para Mac. Transcribe su voz en el dispositivo mediante modelos locales Whisper y Parakeet, funciona sin conexión y guarda el historial de transcripciones en su máquina con eliminación automática configurable. Un diccionario personal gestiona nombres y jerga, mientras los modos por app ajustan el formato para correo electrónico, chat o código. Es de código abierto (https://github.com/Beingpax/VoiceInk) , por lo que sus afirmaciones de privacidad se pueden comprobar en lugar de aceptarse sin más; el proyecto tiene 5,500+ estrellas en GitHub y el desarrollador informa de más de 200,000 descargas. Existen mejoras y proveedores de transcripción opcionales en la nube, pero requieren activación y están etiquetados como tales.

¿Por qué VoiceInk en lugar de los más conocidos Superwhisper o MacWhisper? Por la especialización. Ambos se han expandido a la grabación de reuniones (Superwhisper (https://superwhisper.com/meeting-transcription) , MacWhisper (https://macwhisper.helpscoutdocs.com/article/30-record-meetings)), lo que los sitúa en una categoría distinta (y en una conversación sobre privacidad diferente) del dictado puro. La app para Mac de VoiceInk se limita a un trabajo: escribe lo que usted dice. Su app separada para iOS incluye la transcripción de reuniones y clases entre sus casos de uso, pero no ofrece captura de llamadas, etiquetas de hablantes ni resúmenes; el producto es el dictado. Para conversaciones en vivo con otras personas, ese es el trabajo de Hedy, y ambas se ejecutan en paralelo.

Precio: compra única desde $25 (un Mac) hasta $49 (tres Macs), con actualizaciones lifetime incluidas. Actualmente, una app separada para iOS es gratis.

Mejor para: escritores, desarrolladores y cualquier persona que responda cincuenta correos electrónicos al día y no quiera una suscripción ni depender de la nube por ese privilegio.

3. Jan: chat de IA en el escritorio

Si ChatGPT es su herramienta predeterminada para pensar, pero prefiere no enviar cada idea a medio formar a un servidor, Jan (https://jan.ai/) es el sustituto local con menos fricción. Descargue la app, elija un modelo abierto (Llama, Qwen, Gemma y decenas más en formato GGUF) y empiece a chatear. No necesita crear una cuenta y, una vez descargado el modelo, funciona con el Wi-Fi desactivado.

Jan es de código abierto (https://github.com/janhq/jan) , tiene 43,000+ estrellas en GitHub y reporta 5.9 millones de descargas. Dos decisiones de diseño hacen que sea más que un juguete. Primero, expone una API compatible con OpenAI en localhost, de modo que otras apps de esta lista (y sus propios scripts) pueden usar su modelo local como backend intercambiable. Segundo, es transparente sobre el uso híbrido: puede añadir claves de API de proveedores en la nube cuando quiera la calidad de un modelo de vanguardia, y Jan deja claro que esos chats salen de su máquina bajo las condiciones de ese proveedor.

Precio: gratis.

Mejor para: redacción y lluvia de ideas cotidianas donde "suficientemente bueno en local" supera a "excelente, pero subido".

4. Ollama: ejecución local de modelos

Ollama (https://ollama.com/) es la infraestructura del movimiento de IA local. Es una herramienta de línea de comandos y un servicio en segundo plano que descarga y ejecuta modelos abiertos con un solo comando (ollama run llama3) y después los sirve mediante una API REST local sobre la que se apoya un amplio ecosistema de apps. Si usa varias herramientas de IA local, es muy probable que todas puedan compartir un backend de Ollama en lugar de incluir cada una su propio entorno de ejecución de modelos.

Es de código abierto (https://github.com/ollama/ollama) , tiene 176,000+ estrellas en GitHub y la empresa informó que daba servicio a 8.9 millones de desarrolladores en July 2026. Todo el uso local es gratuito e ilimitado; existen planes de pago en la nube (desde $20 al mes) para ejecutar modelos más grandes en el hardware de Ollama, separados de la vía local.

Precio: gratis para uso local ilimitado; planes opcionales en la nube desde $20 al mes.

Mejor para: usuarios avanzados y cualquiera que quiera un único backend compartido y programable para modelos detrás de sus demás herramientas.

5. Obsidian: notas y conocimiento personal

Sus notas son un mapa de todo aquello en lo que trabaja, lo que convierte a una herramienta de notas solo en la nube en una de las apuestas de privacidad más concentradas que las personas hacen sin pensarlo. Obsidian (https://obsidian.md/) evita por completo esa apuesta: las notas son archivos Markdown simples en una carpeta de su dispositivo. La app funciona sin conexión, los archivos sobreviven a la empresa y nada se sincroniza a menos que usted configure la sincronización (el servicio Sync opcional, desde $4 al mes con facturación anual, está cifrado de extremo a extremo).

La capa de IA proviene de complementos de la comunidad y puede mantenerse local. Smart Connections (https://github.com/brianpetro/obsidian-smart-connections) crea incrustaciones de su bóveda en el dispositivo para buscar notas relacionadas, y Copilot for Obsidian (https://github.com/logancyang/obsidian-copilot) añade chat sobre sus notas con soporte para backends locales como Ollama, que es donde un conjunto local compartido empieza a dar frutos. Con un modelo local detrás, "chatear con mis notas" no involucra a ningún tercero. Ambos complementos también admiten proveedores en la nube, así que lea sus ajustes con el mismo cuidado que los de cualquier herramienta híbrida.

Precio: gratis para uso personal y comercial; Sync opcional desde $4 al mes con facturación anual.

Mejor para: cualquier persona que construya una base de conocimiento personal duradera y quiera buscar en ella con IA sin subir sus ideas.

6. PrivateGPT: preguntas y respuestas privadas sobre documentos

"Chatear con sus documentos" fue la primera aplicación decisiva de la IA privada, y PrivateGPT (https://github.com/zylon-ai/private-gpt) es su referente de código abierto: 57,000+ estrellas en GitHub y un relanzamiento completo en June 2026, cuando Zylon publicó la v1.0 (https://www.zylon.ai/resources/blog/introducing-privategpt-1-0-the-open-source-application-backend-for-private-ai) e incorporó dos años de desarrollo empresarial privado al repositorio público.

PrivateGPT 1.0 es un backend de aplicación: gestiona la ingesta de documentos, la recuperación con citas, el chat y la orquestación de herramientas. Como explica Zylon en el anuncio de la versión 1.0, "PrivateGPT 1.0 no ejecuta modelos por sí mismo": usted lo conecta al servidor de inferencia que elija. Eso es lo que permite una configuración completamente autoalojada: combínelo con Ollama o llama.cpp en la misma máquina y el procesamiento de documentos permanecerá dentro de la infraestructura que usted controla. También es la opción de esta lista más adecuada para un equipo pequeño que quiera poner en marcha un asistente privado compartido en su propio servidor.

Una advertencia justa sobre su idoneidad: esta es la herramienta más técnica de la lista. No hay un instalador con un clic, y las alternativas de consumo más pulidas de esta categoría se han desplazado en su mayoría hacia funciones de transcripción de reuniones, por lo que no entraron en esta lista.

Precio: gratis; usted aporta el hardware.

Mejor para: usuarios técnicos y equipos pequeños que quieran preguntas y respuestas sobre documentos ejecutándose por completo en una infraestructura que controlan.

7. PocketPal AI: IA local en su teléfono

El teléfono es donde la IA local se volvió real para las personas sin conocimientos técnicos, y PocketPal AI (https://pocketpal.dev/) es la forma más sencilla de probarla. Es una app gratuita y de código abierto (https://github.com/a-ghorbani/pocketpal-ai) para iOS y Android que descarga modelos abiertos pequeños (Llama, Qwen, Gemma, Phi) y los ejecuta completamente en el dispositivo. Sin cuenta, sin telemetría y sigue respondiendo en modo avión.

El proyecto informa de 1.2 millones de descargas, y su comunidad ha enviado más de 7,000 resultados de pruebas comparativas en el dispositivo para 100+ modelos de teléfono, lo que también ofrece una respuesta práctica a "¿funcionará esto en mi teléfono?" (Regla general: los iPhones recientes y los teléfonos Android de gama media o superior manejan con comodidad los modelos pequeños). No igualará a un modelo de vanguardia en razonamiento difícil, pero para borradores rápidos y traducciones es una demostración funcional de que la IA útil ya cabe en un bolsillo sin necesidad de un servidor. Es la misma apuesta que hizo Apple al situar el procesamiento en el dispositivo en el centro de Apple Intelligence.

Precio: gratis.

Mejor para: probar la IA local sin ninguna configuración y usarla como asistente de bolsillo sin conexión para tareas rápidas.

¿Cómo puede crear un conjunto de IA local con estas herramientas?

Empiece por el trabajo que contiene sus datos más sensibles. Para muchos profesionales, son las reuniones y conversaciones, el espacio que cubre Hedy; para otros, son los documentos o las notas. Traslade ese flujo de trabajo a una herramienta que priorice lo local y habrá obtenido la mayor parte del beneficio de privacidad en un solo paso.

Después, añada especialistas de uno en uno: una herramienta de dictado, una de chat y una de notas. Evite el solapamiento. Dos herramientas que hacen el mismo trabajo implican dos copias de los mismos datos sensibles en dos lugares, justo lo contrario del objetivo. El conjunto que surge de esta lista comparte bien la infraestructura: Ollama sirve los modelos; Jan, los complementos de Obsidian y PrivateGPT pueden apoyarse en él; VoiceInk escribe lo que usted dice; y Hedy cubre todo lo que se habla con otras personas.

Por último, verifique en lugar de confiar. Desactive el Wi-Fi y compruebe qué sigue funcionando. Lea la página de privacidad y busque datos concretos de cada paso (audio, transcripción, análisis y almacenamiento) en lugar de adjetivos. Todas las herramientas anteriores superan esa lectura; por eso están aquí. En concreto para las reuniones, nuestra guía para usar notas de reuniones con IA de forma privada (/es/post/how-to-use-ai-meeting-notes-privately/) analiza esa comprobación en detalle: qué sale en cada paso y por qué una prueba sin conexión le dice menos de lo que podría pensar.

Preguntas frecuentes

¿Cuál es la mejor herramienta de IA local para la productividad?

Depende del trabajo. Para reuniones y conversaciones en vivo, Hedy ejecuta el reconocimiento de voz en el dispositivo en todas las plataformas, y las máquinas capaces de ejecutar modelos locales potentes también pueden trasladar todo su flujo de IA al dispositivo. Para dictado, VoiceInk transcribe todo en su Mac. Para chat de IA en el escritorio, Jan ejecuta modelos abiertos localmente sin necesidad de una cuenta. Para notas, Obsidian guarda todo en archivos locales. La mayoría de las personas termina combinando dos o tres herramientas que cubren un trabajo cada una.

¿Qué significa "IA local"?

Es un espectro, no una etiqueta de sí o no. Las herramientas completamente locales como Jan, Ollama y PocketPal ejecutan el modelo en su hardware y funcionan sin conexión. Las herramientas híbridas ejecutan parte del flujo localmente: Hedy, por ejemplo, ejecuta el reconocimiento de voz en el dispositivo por defecto y ofrece un modo opcional de Procesamiento de IA Local (/es/help/local-ai-processing/) que también mantiene el paso de análisis en su dispositivo. Lo importante es saber qué pasos permanecen en su máquina, cuáles no y si usted puede controlarlo.

¿Son las herramientas de IA local tan buenas como la IA en la nube?

Para muchas tareas de productividad, sí. El reconocimiento de voz en el dispositivo ya es excelente, las incrustaciones locales gestionan bien la búsqueda en notas y los modelos abiertos pequeños cubren resúmenes y redacción. Para razonamiento de vanguardia, los modelos en la nube siguen a la cabeza. Por eso varias herramientas de esta lista son híbridas honestas: locales donde más importa la privacidad (su audio, sus archivos), en la nube donde más importa la calidad y con un interruptor para pasar a un

funcionamiento completamente local cuando lo necesite.

¿Hedy es completamente local?

La configuración por defecto de Hedy es una división deliberada. El reconocimiento de voz se ejecuta en el dispositivo en todas las plataformas (Whisper en todas partes, además de Nemotron con etiquetas de hablante en el dispositivo), así que con los motores predeterminados su audio nunca sale de su dispositivo y ningún bot se une a sus llamadas. El análisis de IA se ejecuta en la nube por defecto para que Hedy funcione en tantos dispositivos como sea posible con resultados de alta calidad. Sin embargo, cualquier máquina capaz de ejecutar modelos locales potentes (Macs con Apple Silicon, máquinas Windows con GPU capaces, iPhones y iPads recientes) puede trasladar todo el flujo al dispositivo con el modo opcional de Procesamiento de IA Local (/es/help/local-ai-processing/) , que también funciona sin conexión. Desactive Cloud Sync y la conversación solo existirá en el dispositivo que la capturó.

¿Qué hardware necesito para ejecutar herramientas de IA local?

Menos de lo que podría pensar, aunque depende del modelo y del dispositivo. El dictado y el reconocimiento de voz funcionan bien en teléfonos y portátiles recientes. Para chat local con LLM, la regla general es 8 GB de RAM para modelos pequeños y 16 GB o más para modelos medianos; los Macs con Apple Silicon y los PC con IA más nuevos manejan bien ambos. Gartner proyectó que los PC capaces de ejecutar IA representarían el 31% del mercado de PC en 2025, así que el hardware necesario se está convirtiendo rápidamente en la norma.

¿Son gratuitas las herramientas de IA local?

En su mayoría. Jan, Ollama, PrivateGPT y PocketPal AI son completamente gratuitas. Obsidian es gratis para uso personal y comercial, con Sync opcional desde $4 al mes con facturación anual. VoiceInk es una compra única de entre $25 y $49 para Mac. Hedy tiene un nivel gratuito con 5 horas de sesiones al mes; Pro cuesta $12.99 al mes o $99.99 al año.

¿Qué herramientas de IA local funcionan completamente sin conexión?

Jan, Ollama y PocketPal AI siguen funcionando sin conexión una vez descargados sus modelos. VoiceInk transcribe sin conexión con sus modelos locales y Obsidian funciona sin conexión por defecto. PrivateGPT funciona sin conexión cuando se combina con un servidor de inferencia local. Hedy funciona sin conexión cuando el Procesamiento de IA Local está activado en una máquina capaz de ejecutar modelos locales y se combina con un motor de voz en el dispositivo.

¿Cuáles de estas herramientas funcionan en Windows?

Hedy, Jan, Ollama y Obsidian tienen apps para Windows, y PrivateGPT puede autoalojarse en Windows. VoiceInk es solo para Mac (con una app separada para iOS) y PocketPal AI es solo para móviles. Una nota específica sobre Windows: el Procesamiento de IA Local de Hedy requiere una GPU compatible con Vulkan; el reconocimiento de voz en el dispositivo funciona independientemente de ello.

¿Por qué importa la privacidad en las herramientas de IA para la productividad?

Porque los datos de trabajo son exactamente lo que las personas introducen en estas herramientas. Cyberhaven descubrió que el 27.4% de los datos que los empleados introdujeron en herramientas de IA en Q2 2024 eran sensibles, aproximadamente 2.6 veces la proporción de un año antes. El muy citado

incidente de Samsung de 2023 incluyó a un empleado que subió una reunión interna grabada a ChatGPT para generar un acta. El procesamiento local elimina esa vía de fallo: lo que nunca sale de su dispositivo no puede terminar en los registros de otra empresa.
