Die besten lokalen KI-Tools für Produktivität 2026

Vergleichen Sie sieben lokale KI-Tools für Produktivität im Jahr 2026: Meeting-KI auf dem Gerät, Diktieren, Desktop-Chat, Notizen und Dokumenten-Q&A mit klaren Cloud-Grenzen.

Veröffentlicht von Julian Pscheid · 1. Juli 2026 · Aktualisiert 20. August 2026
Diesen Artikel online lesen: https://www.hedy.ai/de/post/best-local-ai-tools-for-productivity/

Mann bei einem virtuellen Meeting an seinem Schreibtisch zu Hause, mit dem Videoanruf auf einem externen Monitor, Produktivitäts-Tools auf dem Laptop-Bildschirm und einem mit dem Display nach unten liegenden iPhone, das neben der Tastatur sanft leuchtet

Kurzantwort Lokale KI-Tools führen KI-Modelle auf Ihrem eigenen Gerät statt in der Cloud eines Anbieters aus. Welches das beste ist, hängt von der Aufgabe ab. Für Meetings führt Hedy die Spracherkennung auf jeder Plattform auf dem Gerät aus. Auf Geräten, die leistungsfähige lokale Modelle ausführen können, lässt sich zudem die gesamte KI-Pipeline auf das Gerät verlagern. VoiceInk transkribiert Diktate vollständig auf Ihrem Mac. Jan führt offene Modelle für Desktop-Chats offline aus. Obsidian speichert Notizen vollständig in lokalen Dateien. Jedes der folgenden Tools übernimmt eine Aufgabe. Zusammen bieten sie für jeden Teil eines Arbeitstags einen lokalen, privaten Weg.

Was gilt als lokale KI?

Bei lokaler KI läuft das Modell auf Ihrer Hardware: Ihrem Laptop, Smartphone oder Server. Ihre Daten werden dort verarbeitet, wo sie bereits gespeichert sind, statt in die Cloud eines Anbieters hochgeladen zu werden.

In der Praxis handelt es sich um ein Spektrum. Wer etwas anderes behauptet, führt Sie in die Irre. Es gibt drei ehrliche Positionen:

• Vollständig lokal. Das Modell läuft auf Ihrem Gerät, und das Tool funktioniert bei ausgeschaltetem Netzwerk. Jan, Ollama, PocketPal AI und der Standardmodus von VoiceInk gehören in diese Kategorie.

• Standardmäßig lokal, Cloud optional. Der private Weg ist voreingestellt; Sie können auf Wunsch einen Cloud-Anbieter verbinden. Die optionale Cloud-Transkription von VoiceInk funktioniert so.

• Hybrid mit lokalen Garantien. Ein Teil der Pipeline ist von Grund auf lokal, ein anderer nutzt die Cloud, und das Tool erklärt genau, welcher Teil wo läuft. Hedy gehört in diese Gruppe: Die Spracherkennung läuft standardmäßig auf jeder Plattform auf dem Gerät, und Geräte mit leistungsfähigen lokalen Modellen können die gesamte Pipeline auf das Gerät verlagern.

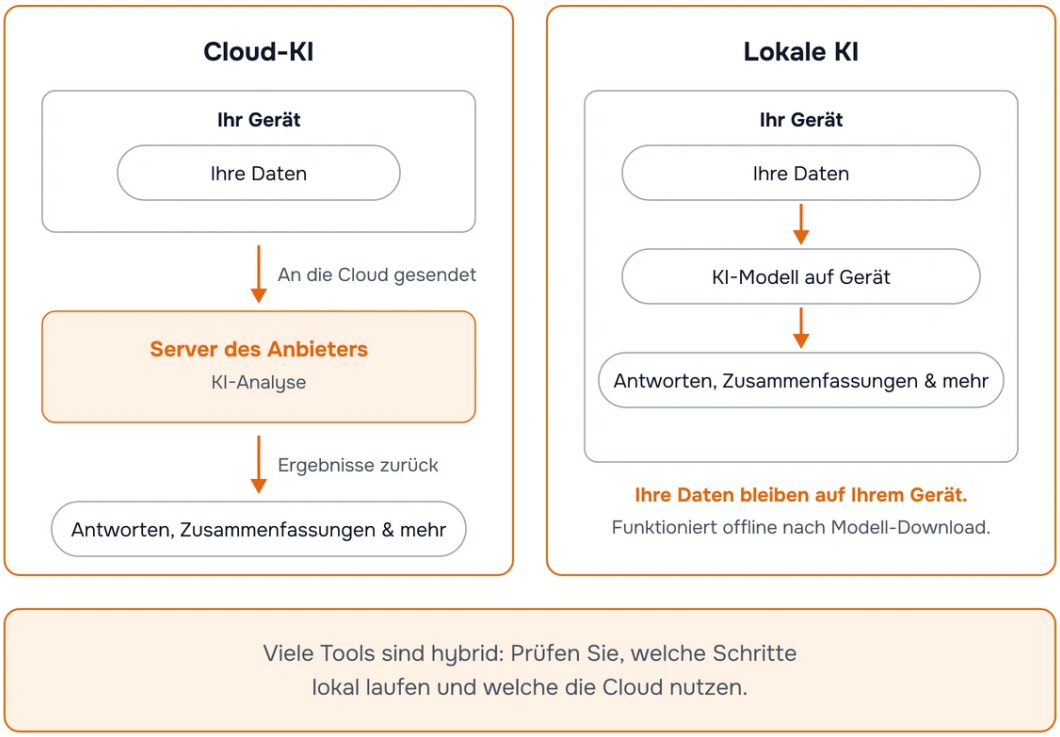

Diagramm zum Vergleich von Cloud-KI, bei der Ihre Daten zur Verarbeitung an die Server des Anbieters gesendet werden, mit lokaler KI, bei der Ihre Daten, das KI-Modell und die Ergebnisse vollständig auf Ihrem Gerät bleiben

Die Nutzung der Cloud schließt ein Tool nicht von dieser Liste aus. Unklare Angaben dagegen schon. Bei jedem Eintrag erfahren Sie, welche Schritte auf Ihrem Gerät ausgeführt werden.

Warum ist lokale KI gerade so gefragt?

Zwei Entwicklungen trafen aufeinander. Geräte werden immer leistungsfähiger, während Modelle zugleich kleiner und intelligenter werden. Irgendwann im vergangenen Jahr waren Modelle, die auf einen Laptop oder ein aktuelles Smartphone passen, leistungsfähig genug für echte Aufgaben: ein Gespräch transkribieren, ein Dokument zusammenfassen oder eine Antwort entwerfen. Beim Aufbau von Hedys lokaler Pipeline haben wir diesen Wendepunkt aus erster Hand erlebt und die Details in unserer technischen Analyse lokaler KI (/de/post/local-ai-engineering-deep-dive-hedy-3-2/) beschrieben.

Im selben Zeitraum wuchs die Sorge darüber, wohin KI-Daten gelangen. Pew Research stellte im Juni 2026 fest (https://www.pewresearch.org/internet/2026/06/17/americans-and-ai-2026-chatbots-smart-devices-and-views-on-impact/) , dass 71% der Erwachsenen in den USA erwarten, eine stärkere KI-

Nutzung werde ihre persönlichen Daten unsicherer machen; nur 3% erwarten das Gegenteil. Aus Sorge verzichten die Menschen jedoch nicht auf KI. Sie geben weiterhin sensibles Material ein: Cyberhaven ermittelte (https://info.cyberhaven.com/hubfs/Content%20PDF/Cyberhaven%20Q2%202024%20AI%20Adoption%20and%20Risk%20Report%20052024.pdf) , dass 27.4% der Daten, die Beschäftigte im Q2 2024 in KI-Tools eingaben, sensibel waren, gegenüber 10.7% ein Jahr zuvor. Das bekannteste warnende Beispiel bleibt der Samsung-Vorfall von 2023 (https://www.engadget.com/three-samsung-employees-reportedly-leaked-sensitive-data-to-chatgpt-190221114.html) , bei dem ein Mitarbeiter eine Aufzeichnung eines internen Meetings zu ChatGPT hochlud, um ein Protokoll zu erstellen.

Unsere Überzeugung, auf der dieser Artikel basiert: Lokale KI weist die Richtung und ist keine Nische. Die Cloud wird bei Spitzenmodellen noch lange führen und für die meisten Menschen bei schwierigen Schlussfolgerungen die bessere Standardeinstellung bleiben. Doch das lokale Leistungsniveau steigt jedes Jahr, und die Hardware steckt bereits in Ihrer Tasche und steht auf Ihrem Schreibtisch (Gartner prognostiziert (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) , dass KI-PCs 2026 einen Marktanteil von 55% erreichen). Sobald sich eine Aufgabe gut auf dem Gerät erledigen lässt, gibt es kein gutes Argument mehr dafür, die Daten an einen anderen Ort zu senden. Die Frage „Wo läuft das?“ wird bei jedem KI-Tool zur Standardfrage, so wie „Ist das verschlüsselt?“ bei Messaging-Diensten. Die folgenden Tools decken Aufgaben ab, bei denen die Antwort schon heute „genau hier“ lauten kann.

So haben wir diese Tools ausgewählt

Fünf Regeln bestimmten diese Liste:

• Ein Tool pro Aufgabe. Ein Stack aus Spezialisten ist besser als eine Ansammlung sich überschneidender Apps, die jeweils eine Kopie Ihrer Daten besitzen. Wir haben für jede Produktivitätsaufgabe eine Option mit lokalem Schwerpunkt gewählt: Meetings, Diktate, Chats, Modellausführung, Notizen, Dokumente und Mobilgeräte.

• Lokal muss echt sein. Jedes Tool läuft entweder vollständig auf dem Gerät oder dokumentiert genau, welche Schritte dort stattfinden. Vages Marketing mit Begriffen wie „Privacy First“ ohne Angaben zur Architektur reichte nicht aus.

• 2026 aktiv. Wir haben Preise, Funktionen und Projektaktivität jedes Tools im Juli 2026 anhand der Anbieterseiten und Repositorys geprüft.

• Hybride sind erlaubt, werden aber offengelegt. Tools, die für einige Schritte die Cloud nutzen, sind enthalten, wenn die Grenze klar beschrieben ist und von Ihnen gesteuert werden kann.

• Vollständige Offenlegung: Hedy ist unser Produkt. Diese Auswahl stammt vom Anbieter und ist kein unabhängiger Benchmark. Hedy belegt bewusst den Platz für Meetings. Direkte Alternativen unter den Meeting-Tools (Granola, Fireflies und der Meeting-Modus von Superwhisper) werden hier nicht bewertet, weil wir in dieser Kategorie keine neutralen Schiedsrichter sind. Für Leserinnen und Leser, die diese Tools abwägen, haben wir separate, quellenbasierte Vergleiche wie unsere Leitfäden zur Granola-Alternative (/de/post/granola-alternative/) und Fireflies-Alternative (/de/post/fireflies-alternative-hedy/) veröffentlicht.

Welches lokale KI-Tool eignet sich am besten für welche Aufgabe?

Vergleich auf einen Blick

#	Tool	Am besten für	Wie lokal	Plattformen	Preis
1	Hedy (https://www.hedy.ai/)	Meetings und Live-Gespräche	Standardmäßig Spracherkennung auf dem Gerät; optional vollständig lokale Pipeline auf geeigneter Hardware	iOS, Android, Mac, Windows, Web	Kostenlos (5 Std./Monat); $12.99/mo oder $99.99/yr
2	VoiceInk (https://tryvoiceink.com/)	Diktate	Standardmäßig vollständig auf dem Gerät; optionale Cloud-Anbieter	macOS (iOS-App verfügbar)	$25 bis $49 einmalig
3	Jan (https://jan.ai/)	KI-Chats am Desktop	Standardmäßig vollständig lokal; optionale Cloud-APIs	Windows, Mac, Linux	Kostenlos
4	Ollama (https://ollama.com/)	Lokale Modellausführung	Vollständig lokal; optionaler Cloud-Tarif	Mac, Windows, Linux	Kostenlos; Pro $20/mo
5	Obsidian (https://obsidian.md/)	Notizen und Wissensdatenbank	Lokale Markdown-Dateien; Plugins können lokale Modelle nutzen	Mac, Windows, Linux, iOS, Android	Kostenlos; Sync ab $4/mo
6	PrivateGPT (https://github.com/zylon-ai/private-gpt)	Private Dokumenten-Q&A	Selbst gehostet; Sie wählen den Inferenzserver	Selbst gehostet (Mac, Linux, Windows)	Kostenlos
7	PocketPal AI (https://pocketpal.dev/)	KI-Chat auf Ihrem Smartphone	Nach dem Modell-Download vollständig auf dem Gerät	iOS, Android	Kostenlos

Quellen: Hedy-Preise (/de/pricing/) , VoiceInk pricing (https://tryvoiceink.com/pricing) , Jan docs (https://jan.ai/docs) , Ollama pricing (https://ollama.com/pricing) , Obsidian pricing (https://obsidian.md/pricing) , PrivateGPT repository (https://github.com/zylon-ai/private-gpt) , PocketPal AI (https://pocketpal.dev/) . Die Preise wurden im Juli 2026 anhand der Anbieterseiten geprüft und ändern sich häufig; prüfen Sie sie vor dem Kauf erneut.

1. Hedy: Meetings und Live-Gespräche

In Meetings werden die sensibelsten Arbeitsdaten laut ausgesprochen: Kundennamen, Konditionen, Gesundheitsinformationen und Strategien. Gleichzeitig gehören Meetings zu den Workflows, bei denen verbreitete KI-Tools oft den geringsten Datenschutz bieten, denn viele Meeting-Assistenten treten Ihrem Anruf als Bot bei und senden die Audiodaten an einen Server.

Hedy (https://www.hedy.ai/) verfolgt den umgekehrten Ansatz. Die Spracherkennung läuft standardmäßig auf jeder Plattform auf dem Gerät (Whisper überall sowie die Nemotron-Engine mit Sprecherzuordnung auf dem Gerät). Mit den Standard-Engines verlassen Ihre Audiodaten daher nie Ihr Gerät, und kein Bot tritt Ihren Anrufen bei. Aufbauend auf dem Transkript arbeitet Hedy als Echtzeit-Meeting-Coach: Während das Gespräch noch läuft, erhalten Sie Vorschläge, Antworten und sinnvolle Fragen. Anschließend erstellt Hedy Zusammenfassungen, Highlights und Aufgaben.

Hedys Standardeinstellungen sind bewusst aufgeteilt. Die Spracherkennung läuft auf jedem Gerät lokal, weil das überall in hoher Qualität möglich ist. Die KI-Analyse läuft standardmäßig in der Cloud, damit Hedy auf möglichst vielen Geräten die bestmöglichen Ergebnisse liefert. Übertragene Inhalte werden weder gespeichert noch zum Trainieren von Modellen verwendet. Lokal ist jedoch deutlich mehr möglich als in der Standardeinstellung: Jedes Gerät, auf dem leistungsfähige lokale Modelle laufen, kann mit Local AI Processing (/de/help/local-ai-processing/) die gesamte Meeting-Pipeline auf das Gerät verlagern. Transkripte, Zusammenfassungen, detaillierte Notizen, Chat-Antworten und Echtzeit-Vorschläge entstehen dann auf dem Gerät, das die Audiodaten erfasst hat, auf Wunsch offline. Das funktioniert auf Apple Silicon Macs, Windows-Rechnern mit leistungsfähigen GPUs sowie aktuellen iPhones und iPads. Deaktivieren Sie Cloud Sync, gelangt in keinem Schritt etwas aus dem Gespräch auf unsere Server, es sei denn, Sie teilen diese Sitzung ausdrücklich. Hedy 3.2 hat dies ermöglicht; mehr dazu erfahren Sie in Lokale KI für Meetings (/de/post/local-ai-meetings-hedy-3-2/) . Sie können außerdem die Cloud-KI-Analyse deaktivieren (/de/help/cloud-ai-analysis-privacy-control/) und dabei die vollständige Transkription beibehalten, EU-Datenspeicherung (/de/post/eu-data-residency/) wählen und jede Einstellung in unserem Leitfaden zu Datenschutzeinstellungen (/de/post/ai-meeting-privacy-settings-explained/) prüfen.

Der Standardmodus ist nicht vollständig lokal. Der vollständig lokale Modus steht bereit, wenn Ihre Hardware und Datenschutzanforderungen ihn verlangen. In beiden Fällen erhalten Sie eine dokumentierte Grenze, die Sie steuern. Hedy hat außerdem im April 2026 eine unabhängige SOC 2 Type I-Prüfung seiner Sicherheitskontrollen und eine unabhängige HIPAA-Bewertung als Business Associate abgeschlossen. Hedy wird bei 500+ Rezensionen von 30,000+ Nutzerinnen und Nutzern mit 4.8 von 5 Sternen bewertet.

Preis: Kostenlos für 5 Sitzungsstunden pro Monat, keine Kreditkarte erforderlich. Pro kostet $12.99/month oder $99.99/year; eine Lifetime-Lizenz für $299 deckt dauerhaft alle Funktionen ab.

Am besten für: alle, deren Gespräche die sensibelsten Inhalte sind, die sie sonst einem Cloud-Tool anvertrauen würden: Beraterinnen und Berater, Vertriebsmitarbeitende, Journalistinnen und Journalisten, Patientinnen und Patienten sowie Führungskräfte.

2. VoiceInk: Diktate

Diktieren ist der unauffällige Produktivitätsgewinn lokaler KI. Die meisten Menschen sprechen deutlich schneller, als sie tippen. Ein gutes Modell auf dem Gerät transkribiert inzwischen so genau, dass es keinen Grund mehr gibt, dafür Ihre Stimme hochzuladen.

VoiceInk (https://tryvoiceink.com/) ist der Diktier-Spezialist für Mac. Die App transkribiert Ihre Sprache mit lokalen Whisper- und Parakeet-Modellen auf dem Gerät, funktioniert offline und speichert Ihren Transkriptionsverlauf mit konfigurierbarer automatischer Löschung auf Ihrem Rechner. Ein persönliches Wörterbuch verarbeitet Namen und Fachbegriffe, und Modi für einzelne Apps passen die Formatierung für E-Mails, Chats oder Code an. VoiceInk ist Open Source (https://github.com/Beingpax/VoiceInk) . Die Datenschutzangaben lassen sich also überprüfen, statt dass Sie ihnen einfach vertrauen müssen. Das Projekt hat 5,500+ GitHub-Sterne, und der Entwickler berichtet von über 200,000 Downloads. Optionale Cloud-Optimierung und Anbieter für Cloud-Transkription sind verfügbar, müssen jedoch bewusst aktiviert werden und sind entsprechend gekennzeichnet.

Warum VoiceInk statt der bekannteren Apps Superwhisper oder MacWhisper? Wegen der Spezialisierung. Beide haben ihr Angebot auf Meeting-Aufzeichnungen erweitert (Superwhisper (https://superwhisper.com/meeting-transcription) , MacWhisper (https://macwhisper.helpscoutdocs.com/article/30-record-meetings)). Damit gehören sie in eine andere Kategorie und erfordern eine andere Datenschutzdiskussion als reine Diktier-Tools. Die Mac-App von VoiceInk bleibt bei einer Aufgabe: Sie tippt, was Sie sagen. Die separate iOS-App nennt zwar die Transkription von Meetings und Vorlesungen als Anwendungsfälle, bietet aber keine Anruferfassung, Sprecherzuordnung oder Zusammenfassung. Das Produkt ist zum Diktieren gedacht. Bei Live-Gesprächen mit anderen Menschen übernimmt Hedy. Beide Apps können parallel laufen.

Preis: Einmalkauf ab $25 (ein Mac) bis $49 (drei Macs), einschließlich lebenslanger Updates. Eine separate iOS-App ist derzeit kostenlos.

Am besten für: Autorinnen und Autoren, Entwicklerinnen und Entwickler sowie alle, die täglich fünfzig E-Mails beantworten und dafür weder ein Abo noch eine Abhängigkeit von der Cloud möchten.

3. Jan: KI-Chat am Desktop

Wenn ChatGPT Ihr bevorzugtes Tool zum Nachdenken ist, Sie aber nicht jede unfertige Idee an einen Server senden möchten, bietet Jan (https://jan.ai/) den unkompliziertesten lokalen Ersatz. Laden Sie die App herunter, wählen Sie ein offenes Modell (Llama, Qwen, Gemma und Dutzende weitere im GGUF-Format) und beginnen Sie zu chatten. Sie müssen kein Konto erstellen, und nach dem Modell-Download funktioniert die App bei ausgeschaltetem WLAN.

Jan ist Open Source (https://github.com/janhq/jan) , hat 43,000+ GitHub-Sterne und meldet 5.9 Millionen Downloads. Zwei Designentscheidungen machen die App zu mehr als einem Spielzeug. Erstens stellt sie auf localhost eine OpenAI-kompatible API bereit. Andere Apps auf dieser Liste und Ihre eigenen Skripte können Ihr lokales Modell dadurch direkt als Backend verwenden. Zweitens kommuniziert Jan den hybriden Einsatz offen: Wenn Sie die Qualität eines Spitzenmodells benötigen, können Sie API-Schlüssel von Cloud-Anbietern hinzufügen. Jan weist klar darauf hin, dass diese Chats Ihren Rechner verlassen und den Bedingungen des jeweiligen Anbieters unterliegen.

Preis: Kostenlos.

Am besten für: alltägliche Entwürfe und Brainstorming, bei denen „lokal gut genug“ besser ist als „hervorragend, aber hochgeladen“.

4. Ollama: Modelle lokal ausführen

Ollama (https://ollama.com/) ist die technische Grundlage der lokalen KI-Bewegung. Das Kommandozeilen-Tool mit Hintergrunddienst lädt offene Modelle mit einem einzigen Befehl herunter und führt sie aus (ollama run llama3). Anschließend stellt es sie über eine lokale REST API bereit, auf der ein großes App-Ökosystem aufbaut. Wenn Sie mehrere lokale KI-Tools verwenden, können diese mit hoher Wahrscheinlichkeit dasselbe Ollama-Backend nutzen, statt jeweils eine eigene Modell-Laufzeitumgebung mitzubringen.

Ollama ist Open Source (https://github.com/ollama/ollama) und hat 176,000+ GitHub-Sterne. Das Unternehmen berichtete im Juli 2026 von 8.9 Millionen Entwicklerinnen und Entwicklern. Die lokale Nutzung ist kostenlos und unbegrenzt. Kostenpflichtige Cloud-Tarife (ab $20/month) führen größere Modelle auf Ollamas Hardware aus und bleiben vom lokalen Weg getrennt.

Preis: Kostenlos für unbegrenzte lokale Nutzung; optionale Cloud-Tarife ab $20/month.

Am besten für: Power-User und alle, die ein gemeinsames, skriptfähiges Modell-Backend für ihre anderen Tools möchten.

5. Obsidian: Notizen und persönliches Wissen

Ihre Notizen bilden alles ab, woran Sie arbeiten. Ein ausschließlich cloudbasiertes Notiz-Tool gehört daher zu den größeren gebündelten Datenschutzrisiken, die Menschen eingehen, ohne darüber nachzudenken. Obsidian (https://obsidian.md/) vermeidet dieses Risiko vollständig: Notizen sind einfache Markdown-Dateien in einem Ordner auf Ihrem Gerät. Die App funktioniert offline, die Dateien überdauern das Unternehmen, und ohne Ihre Einrichtung wird nichts synchronisiert. Der optionale Sync-Dienst ab $4/month bei jährlicher Abrechnung ist Ende-zu-Ende-verschlüsselt.

Die KI-Ebene stammt aus Community-Plugins und kann lokal bleiben. Smart Connections (https://github.com/brianpetro/obsidian-smart-connections) erstellt Embeddings Ihres Vaults auf dem Gerät, um verwandte Notizen zu finden. Copilot for Obsidian (https://github.com/logancyang/obsidian-copilot) ergänzt einen Chat über Ihre Notizen und unterstützt lokale Backends wie Ollama. Hier zahlt sich ein gemeinsamer lokaler Stack aus. Mit einem lokalen Modell im Hintergrund ist am Chat mit Ihren Notizen kein Drittanbieter beteiligt. Beide Plugins unterstützen auch Cloud-Anbieter. Prüfen Sie ihre Einstellungen daher ebenso sorgfältig wie bei jedem anderen hybriden Tool.

Preis: Kostenlos für private und geschäftliche Nutzung; optionales Sync ab $4/month bei jährlicher Abrechnung.

Am besten für: alle, die eine langfristige persönliche Wissensdatenbank aufbauen und darin mit KI suchen möchten, ohne ihre Gedanken hochzuladen.

6. PrivateGPT: Private Dokumenten-Q&A

„Mit Ihren Dokumenten chatten“ war die erste entscheidende Anwendung privater KI. PrivateGPT (https://github.com/zylon-ai/private-gpt) ist ihr bekanntester Open-Source-Vertreter: 57,000+ GitHub-Sterne und ein vollständiger Neustart im Juni 2026, als Zylon v1.0 (https://www.zylon.ai/resources/blog/introducing-privategpt-1-0-the-open-source-application-backend-for-private-ai) veröffentlichte und zwei Jahre privater Entwicklung für Unternehmen in das öffentliche Repository überführte.

PrivateGPT 1.0 ist ein Anwendungs-Backend: Es übernimmt Dokumentenaufnahme, Retrieval mit Quellenangaben, Chats und Tool-Orchestrierung. Wie Zylon in der Ankündigung von 1.0 schreibt, führt „PrivateGPT 1.0 selbst keine Modelle aus“: Sie verbinden es mit dem Inferenzserver Ihrer Wahl. Genau das ermöglicht hier eine vollständig selbst gehostete Einrichtung. Koppeln Sie PrivateGPT auf demselben Rechner mit Ollama oder llama.cpp, bleibt die Dokumentenverarbeitung innerhalb der von Ihnen kontrollierten Infrastruktur. Unter den Tools auf dieser Liste eignet es sich außerdem am besten für ein kleines Team, das einen gemeinsamen privaten Assistenten auf dem eigenen Server einrichten möchte.

Ein ehrlicher Hinweis zur Eignung: Dies ist das technisch anspruchsvollste Tool auf der Liste. Es gibt kein Ein-Klick-Installationsprogramm, und die ausgereiften Alternativen für Verbraucherinnen und Verbraucher in dieser Kategorie haben sich größtenteils in Richtung Meeting-Transkription entwickelt. Deshalb wurden sie nicht aufgenommen.

Preis: Kostenlos; Sie stellen die Hardware bereit.

Am besten für: technisch versierte Nutzerinnen und Nutzer sowie kleine Teams, die Dokumenten-Q&A vollständig auf ihrer eigenen Infrastruktur ausführen möchten.

7. PocketPal AI: lokale KI auf Ihrem Smartphone

Auf Smartphones wurde lokale KI für technisch weniger versierte Menschen greifbar. PocketPal AI (https://pocketpal.dev/) ist der unkomplizierteste Einstieg. Die kostenlose Open-Source-App (https://github.com/a-ghorbani/pocketpal-ai) für iOS und Android lädt kleine offene Modelle (Llama, Qwen, Gemma, Phi) herunter und führt sie vollständig auf dem Gerät aus. Kein Konto, keine Telemetrie, und im Flugmodus antwortet die App weiterhin.

Das Projekt meldet 1.2 Millionen Downloads. Seine Community hat über 7,000 Benchmark-Ergebnisse auf dem Gerät für 100+ Smartphone-Modelle eingereicht. Damit beantwortet sie zugleich die praktische Frage „Läuft das auf meinem Smartphone?“. Als Faustregel gilt: Aktuelle iPhones und Android-Smartphones der Mittelklasse oder besser bewältigen kleine Modelle problemlos. Bei schwierigen Schlussfolgerungen kann PocketPal AI nicht mit einem Spitzenmodell mithalten. Für schnelle Entwürfe und Übersetzungen zeigt die App jedoch überzeugend, dass nützliche KI inzwischen ohne Server in eine Hosentasche passt. Apple setzt mit der Verarbeitung auf dem Gerät im Zentrum von Apple Intelligence auf dieselbe Entwicklung.

Preis: Kostenlos.

Am besten für: lokale KI ohne Einrichtung auszuprobieren und schnelle Aufgaben mit einem Offline-Assistenten für unterwegs zu erledigen.

Wie bauen Sie aus diesen Tools einen lokalen KI-Stack auf?

Beginnen Sie mit der Aufgabe, die Ihre sensibelsten Daten enthält. Für viele Fachleute sind das Meetings und Gespräche, die Hedy abdeckt; für andere sind es Dokumente oder Notizen. Verlagern Sie diesen einen Workflow auf ein Tool mit lokalem Schwerpunkt, und Sie erzielen den größten Teil des Datenschutzvorteils in einem einzigen Schritt.

Fügen Sie anschließend nach und nach Spezialisten hinzu: ein Diktier-Tool, ein Chat-Tool und ein Notiz-Tool. Vermeiden Sie Überschneidungen. Zwei Tools für dieselbe Aufgabe bedeuten zwei Kopien derselben sensiblen Daten an zwei Orten. Das ist das Gegenteil des eigentlichen Ziels. Der Stack aus dieser Liste teilt Infrastruktur auf sinnvolle Weise: Ollama stellt Modelle bereit; Jan, die Obsidian-Plugins und PrivateGPT können darauf aufbauen; VoiceInk tippt, was Sie sagen; und Hedy deckt alles ab, was Sie mit anderen Menschen besprechen.

Überprüfen Sie am Ende selbst, statt Angaben einfach zu vertrauen. Schalten Sie das WLAN aus und sehen Sie nach, was weiter funktioniert. Lesen Sie die Datenschutzseite und suchen Sie nach konkreten Angaben zu den einzelnen Schritten (Audio, Transkription, Analyse und Speicherung) statt nach Adjektiven. Jedes der oben genannten Tools besteht diesen Test und ist deshalb auf der Liste. Speziell für Meetings geht unser Leitfaden zur privaten Nutzung von KI-Meeting-Notizen (/de/post/how-to-use-ai-meeting-notes-privately/) diese Prüfung im Detail durch: was bei jedem Schritt das Gerät verlässt und warum ein Offline-Test weniger aussagt, als Sie vielleicht denken.

Häufig gestellte Fragen

Welches ist das beste lokale KI-Tool für Produktivität?

Das hängt von der Aufgabe ab. Für Meetings und Live-Gespräche führt Hedy die Spracherkennung auf jeder Plattform auf dem Gerät aus. Geräte, auf denen leistungsfähige lokale Modelle laufen, können zudem Hedys gesamte KI-Pipeline auf das Gerät verlagern. Für Diktate transkribiert VoiceInk vollständig auf Ihrem Mac. Für KI-Chats am Desktop führt Jan offene Modelle lokal und ohne Konto aus. Obsidian speichert Ihre Notizen vollständig in lokalen Dateien. Die meisten Menschen kombinieren am Ende zwei oder drei Tools, die jeweils eine Aufgabe abdecken.

Was bedeutet lokale KI?

Es ist ein Spektrum, kein Ja-Nein-Etikett. Vollständig lokale Tools wie Jan, Ollama und PocketPal führen das Modell auf Ihrer Hardware aus und funktionieren offline. Hybride Tools führen einen Teil der Pipeline lokal aus: Hedy beispielsweise nutzt standardmäßig Spracherkennung auf dem Gerät und bietet optional Local AI Processing an, sodass auch der Analyseschritt auf Ihrem Gerät bleibt. Entscheidend ist, dass Sie wissen, welche Schritte auf Ihrem Gerät bleiben, welche nicht und ob Sie das steuern können.

Sind lokale KI-Tools so gut wie Cloud-KI?

Für viele Produktivitätsaufgaben ja. Die Spracherkennung auf dem Gerät ist inzwischen hervorragend, lokale Embeddings eignen sich gut für die Notizsuche und kleine offene Modelle bewältigen Zusammenfassungen und Entwürfe. Bei komplexen Schlussfolgerungen auf Spitzenniveau führen Cloud-Modelle weiterhin. Deshalb sind mehrere Tools auf dieser Liste ehrliche Hybride: lokal, wo Datenschutz am wichtigsten ist (Ihre Audiodaten und Dateien), und in der Cloud, wo Qualität am wichtigsten ist, mit einem Schalter für die vollständig lokale Verarbeitung, wenn Sie sie benötigen.

Ist Hedy vollständig lokal?

Hedys Standardeinstellungen sind bewusst aufgeteilt. Die Spracherkennung läuft auf jeder Plattform auf dem Gerät (Whisper überall sowie Nemotron mit Sprecherlabels auf dem Gerät), sodass Ihr Audio mit den Standard-Engines nie Ihr Gerät verlässt und kein Bot Ihren Anrufen beitritt. Die KI-Analyse läuft standardmäßig in der Cloud, damit Hedy auf möglichst vielen Geräten hochwertige Ergebnisse liefert. Jedes Gerät, auf dem leistungsfähige lokale Modelle laufen (Apple Silicon Macs, Windows-Rechner mit leistungsfähigen GPUs sowie aktuelle iPhones und iPads), kann mit dem optionalen Local AI Processing (/de/help/local-ai-processing/) jedoch die gesamte Pipeline auf das Gerät verlagern. Dieser Modus funktioniert auch offline. Deaktivieren Sie Cloud Sync, gelangt nichts aus dem Gespräch auf unsere Server, es sei denn, Sie teilen diese Sitzung ausdrücklich.

Welche Hardware benötige ich für lokale KI-Tools?

Weniger, als Sie vielleicht denken, wobei es vom Modell und Gerät abhängt. Diktate und Spracherkennung funktionieren auf aktuellen Smartphones und Laptops gut. Für lokale LLM-Chats gilt als Faustregel: 8 GB RAM für kleine Modelle und mindestens 16 GB für mittelgroße Modelle. Apple Silicon Macs und neuere KI-PCs bewältigen beides gut. Gartner prognostizierte, dass KI-fähige PCs 2025 einen Anteil von 31% am PC-Markt erreichen würden. Die passende Hardware wird also schnell zum Standard.

Sind lokale KI-Tools kostenlos?

Größtenteils. Jan, Ollama, PrivateGPT und PocketPal AI sind vollständig kostenlos. Obsidian ist für private und geschäftliche Nutzung kostenlos; das optionale Sync kostet bei jährlicher Abrechnung ab $4/month. VoiceInk ist für Mac als Einmalkauf zwischen $25 und $49 erhältlich. Hedy bietet einen kostenlosen Tarif mit 5 Sitzungsstunden pro Monat; Pro kostet $12.99/month oder $99.99/year.

Welche lokalen KI-Tools funktionieren vollständig offline?

Jan, Ollama und PocketPal AI funktionieren ohne Verbindung weiter, sobald ihre Modelle heruntergeladen wurden. VoiceInk transkribiert mit seinen lokalen Modellen offline, und Obsidian funktioniert standardmäßig offline. PrivateGPT läuft offline, wenn es mit einem lokalen Inferenzserver verbunden ist. Hedy funktioniert offline, wenn Local AI Processing auf einem Gerät aktiviert ist, das lokale Modelle ausführen kann, und eine Spracherkennungs-Engine auf dem Gerät verwendet wird.

Welche dieser Tools laufen unter Windows?

Hedy, Jan, Ollama und Obsidian bieten Windows-Apps, und PrivateGPT kann unter Windows selbst gehostet werden. VoiceInk ist nur für Mac verfügbar (mit einer separaten iOS-App), PocketPal AI nur für Mobilgeräte. Ein Hinweis speziell zu Windows: Hedys Local AI Processing erfordert eine Vulkan-fähige GPU; die Spracherkennung auf dem Gerät funktioniert unabhängig davon.

Warum ist Datenschutz bei KI-Produktivitäts-Tools wichtig?

Weil Menschen gerade ihre Arbeitsdaten in diese Tools eingeben. Cyberhaven stellte fest, dass 27.4% der Daten, die Beschäftigte im Q2 2024 in KI-Tools eingaben, sensibel waren. Das war etwa das 2.6-Fache des Anteils ein Jahr zuvor. Beim viel zitierten Samsung-Vorfall von 2023 lud ein Mitarbeiter unter anderem die Aufzeichnung eines internen Meetings zu ChatGPT hoch, um ein Protokoll zu erstellen. Lokale Verarbeitung schließt diese Fehlerquelle aus: Was Ihr Gerät nie verlässt, kann nicht in den Protokollen eines anderen Unternehmens landen.

Hedy AI · Live-KI-Coaching für wichtige Gespräche

Hedy kostenlos testen: https://www.hedy.ai/de/downloads/

https://www.hedy.ai/de/post/best-local-ai-tools-for-productivity/