Lokale KI für Meetings: Hedy 3.2 kann jetzt vollständig auf Ihrem Gerät laufen

Mit Hedy 3.2 können Sie die gesamte KI-Pipeline auf Ihrem eigenen Gerät ausführen. Zusammenfassungen, Notizen, Chat-Antworten und Live-Vorschläge entstehen lokal – nichts erreicht einen Server.

Veröffentlicht von Julian Pscheid · 23. April 2026
Diesen Artikel online lesen: https://www.hedy.ai/de/post/local-ai-meetings-hedy-3-2/

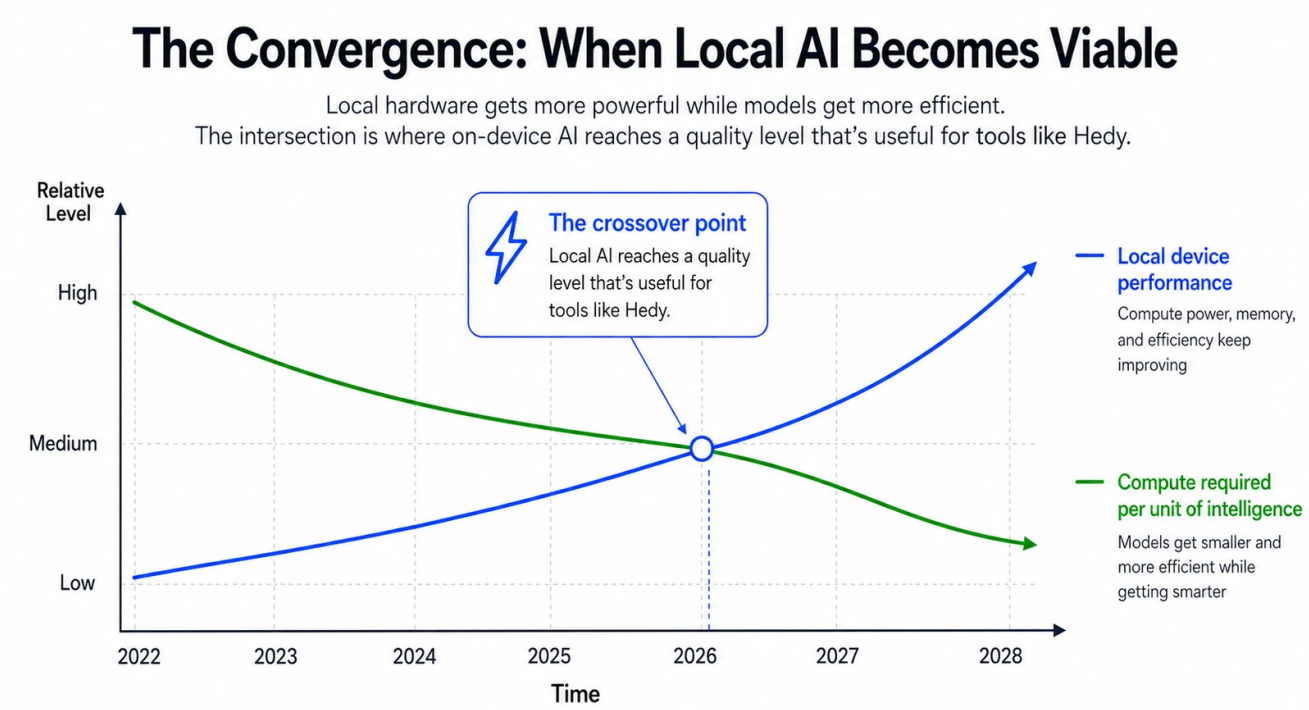
Eine Anwältin und ihr Mandant in einer vertraulichen Beratung in der Kanzlei, mit einem MacBook zwischen ihnen, umgeben von einem sanften cyan- und violettfarbenen Hologramm, das KI-Verarbeitung auf dem Gerät andeutet

Mit Hedy 3.2 kann unsere gesamte KI-Pipeline jetzt auf Ihrem eigenen Gerät laufen. Wenn Sie die Funktion aktivieren, entstehen Meeting-Transkripte, Zusammenfassungen, detaillierte Notizen, Chat-Antworten und Live-Vorschläge auf demselben Gerät, das das Audio erfasst hat. Nichts aus Ihren Gesprächen geht an einen Server. Hier ist, was sich ändert, wenn Ihre Meeting-KI so arbeitet.

Die Spracherkennung läuft schon seit dem ersten Tag von Hedy auf Ihrem Gerät. Audioaufnahmen sind ebenfalls immer auf dem Gerät geblieben. Ihre Gespräche wurden nie zum Training von KI-Modellen verwendet. Datenschutz war von Anfang an eine bewusste Entscheidung.

Aber es gab immer einen Teil, den wir nicht auf das Gerät holen konnten: die KI-Arbeit selbst. Der Teil, der Ihr Transkript liest, Ihre Zusammenfassung schreibt, detaillierte Notizen erstellt, Ihre Fragen zu einem Meeting beantwortet und Ihnen während des Gesprächs Vorschläge macht. Diese Arbeit musste auf Servern stattfinden, weil die Modelle, die dafür gut genug waren, zu groß für einen Laptop oder ein Telefon waren.

Das hat sich schnell geändert. Geräte werden immer leistungsfähiger. KI-Modelle werden gleichzeitig kleiner und intelligenter. Vor einigen Monaten haben sich diese beiden Entwicklungen für uns gekreuzt. Modelle, die auf einen Laptop oder ein neueres iPhone passen, sind inzwischen stark genug, um Hedys Analyse mit einer Qualität zu übernehmen, die für echte Meetings nützlich ist.



Konzeptionelle Grafik, die zeigt, wie die lokale Geräteleistung steigt und der Rechenaufwand pro Intelligenzeinheit sinkt, mit einem markierten Schnittpunkt um 2026, ab dem KI auf dem Gerät für Tools wie Hedy einsatzfähig wird

Veranschaulichend – die Kurven zeigen die grobe Richtung, keine konkreten Messwerte.

Mit Hedy 3.2 können Sie deshalb Local AI Processing aktivieren und unsere gesamte KI-Pipeline auf Ihrem eigenen Gerät ausführen. Zusammenfassungen, Notizen, Chat-Antworten, Vorschläge. Alles davon passiert auf Ihrem Gerät.

Warum KI auf dem Gerät wichtig ist

Über weite Teile der letzten Jahre war KI etwas, das einige wenige große Unternehmen in Ihrem Auftrag betrieben haben. Sie senden Ihre Daten an deren Server, deren Modelle verarbeiten sie, und die Ergebnisse kommen zurück. Dieses Modell hat offensichtliche Vorteile und hat den größten Teil dessen möglich gemacht, was wir bisher bauen konnten. Es hat aber auch strukturelle Kosten: Die nützlichste KI braucht die persönlichsten Daten, und diese Daten liegen an einem anderen Ort als dem, an dem sie entstanden sind.

Lokale KI dreht das um. Ihr Gespräch bleibt dort, wo es stattgefunden hat. Die Arbeit passiert auf demselben Gerät, das das Audio erfasst hat. Nichts über das Meeting erreicht einen Server, es sei denn, Sie entscheiden sich dafür, Cloud Sync zu aktivieren. Und selbst dann bleibt die KI-Verarbeitung lokal.

Wenn Sie Cloud Sync deaktiviert lassen, existiert Ihr Gespräch nur auf dem Gerät, das es aufgezeichnet hat. Durchgehend.

Wer von privater Meeting-KI auf dem Gerät profitiert

Manche Hedy-Nutzer werden diese Funktion aktivieren und danach nicht mehr darüber nachdenken. Andere warten seit Jahren auf genau so etwas. Aus unserer Sicht profitieren diese Gruppen am meisten.

• Coaches und Berater , deren Kundengespräche strenge Vertraulichkeit erwarten lassen. Sie haben Hedy für Vorbereitung und interne Gespräche genutzt. Jetzt können sie Hedy auch in der Kundenarbeit einsetzen, ohne dass etwas den Laptop verlässt.

• Anwälte , die Hedy für interne Gespräche nutzen, aber nie für Mandantengespräche. Das Mandatsgeheimnis hat eine klare Form, und „wir versprechen, Ihre Daten sorgfältig zu behandeln" passt nicht in diese Form. Daten, die sich nicht bewegen, passen.

• Patientinnen und Patienten , die in Arzttermine gehen und eine klare Zusammenfassung dessen möchten, was die Ärztin oder der Arzt gesagt hat, aber nicht möchten, dass ihr Gesundheitsgespräch auf einem Drittanbieter-Server liegt. Mit lokaler KI kann die Zusammenfassung auf demselben Telefon entstehen, das das Gespräch aufgezeichnet hat.

• Journalistinnen und Journalisten , die an sensiblen Recherchen arbeiten und Cloud-Tools vollständig meiden. Sie können ein Interview aufzeichnen, ein Transkript erhalten und mit dem Meeting chatten, ohne dass etwas einen Server erreicht.

• Alle außerhalb der USA , die nicht möchten, dass ihre Gespräche auf US-Servern liegen. Wir haben EU-Datenresidenz Anfang dieses Jahres eingeführt. Lokale KI geht einen Schritt weiter: Die Daten liegen auf den Servern gar keines Unternehmens.

• Remote-Arbeitende mit schlechtem Flugzeug-WLAN oder in ländlichen Regionen ohne Netzabdeckung. Hedy funktioniert jetzt vollständig offline. Laptop im Flugzeug aufklappen, ein echtes Gespräch führen, die Zusammenfassung vor der Landung erhalten.

• Datenschutzbewusste , die keinen regulierten Beruf und kein spezifisches Bedrohungsmodell haben. Sie fanden einfach die Idee besser, dass das Tool, das ihren Meetings zuhört, das Audio nirgendwohin sendet. Bisher konnten sie das nicht haben. Jetzt können sie es.

Was diese Menschen verbindet: Hedy war für sie konzeptionell bereits nützlich, aber das Datenmodell passte nicht zu ihren Anforderungen. Lokale KI entfernt diese Einschränkung.

Die ehrliche Version

Wir möchten klar sagen, was das ist und was nicht, weil wir lieber möchten, dass Sie mit realistischen Erwartungen starten, statt später enttäuscht zu sein.

Local AI Processing ist opt-in und standardmäßig deaktiviert. Cloud AI ist weiterhin schneller, liefert weiterhin bessere Ergebnisse und läuft auf jeder Plattform, die Hedy unterstützt. Wenn Sie keinen konkreten Grund haben, Verarbeitung auf dem Gerät zu wollen, ist die Cloud-Option im Moment die bessere Erfahrung.

Eine Zusammenfassung, die in der Cloud sofort wirkt, kann lokal je nach Hardware und gewähltem Modell zwischen 30 Sekunden und mehreren Minuten dauern. Kleinere Modelle sind gut bei kurzen Zusammenfassungen, geraten aber bei langen oder nuancierten Gesprächen an Grenzen. Größere Modelle kommen nahe an Cloud-Qualität heran, brauchen aber echte Hardware, um gut zu laufen. Und wir fallen nicht still auf die Cloud zurück, wenn lokal etwas fehlschlägt. Sie haben sich aus einem Grund für lokal entschieden, und ein stiller Wiederholungsversuch auf unseren Servern würde diese Absicht unterlaufen. Stattdessen sehen Sie eine Fehlermeldung.

Lokale KI wird auf Apple Silicon Macs, Windows-Rechnern mit leistungsfähigen GPUs, neueren iPhones (15 Pro und neuer) sowie M-series iPads unterstützt. Android und Web stehen auf der Roadmap, sind aber noch nicht bereit. Die große Vielfalt an Android-Hardware und die Einschränkungen beim Ausführen von Modellen im Browser machen es heute schwierig, eine konsistente Erfahrung zu liefern.

Sie wählen das Modell, das zu Ihrer Hardware passt. Wir bieten drei Qualitätsstufen an: von kompakten Modellen, die auf ein Telefon passen, über Mittelklasse-Optionen, die auf den meisten modernen Laptops gut funktionieren, bis hin zu größeren Modellen, die auf leistungsfähiger Hardware nahe an Cloud-Qualität herankommen. Der Modell-Picker zeigt Ihnen vor dem Download, was passt.

Wohin das führt

Lokale KI in Hedy 3.2 ist ein Ausgangspunkt, kein fertiges Produkt. Modelle werden besser werden. Consumer-Hardware wird leistungsfähiger werden. Die Lücke zwischen lokaler Qualität und Cloud-Qualität wird weiter schrumpfen. Wir werden die Unterstützung auf weitere Plattformen ausweiten, sobald die Technologie es erlaubt.

Das ist Teil einer größeren Entwicklung, die Anerkennung verdient. Über Jahre kamen die größten Fortschritte bei KI von Unternehmen, die immer größere Modelle auf immer größeren Serverfarmen betrieben. Parallel dazu findet eine leisere Verschiebung statt: Open-Weight-Modelle werden alle paar Monate intelligenter und kleiner. Die Hardware, auf der sie laufen können, steckt in den Taschen und steht auf den Schreibtischen der meisten Menschen. Die Fähigkeitslücke zwischen Meeting-KI in der Cloud und Meeting-KI auf dem Gerät schließt sich schnell.

Das größere Bild ist wichtiger als jede einzelne Veröffentlichung. Wir glauben, dass die nächsten Jahre der KI von einem Wechsel geprägt sein werden: weg von einer Welt, in der einige wenige Unternehmen KI in Ihrem Auftrag betreiben, hin zu einer Welt, in der Sie Ihre eigene Pipeline auf Ihrem eigenen Gerät mit Ihren eigenen Daten vollständig selbst ausführen können. Ob Sie diese Option nutzen oder nicht: Dass es sie gibt, gibt der Technologie ihre richtige Form. Es hält die Macht im Gleichgewicht.

Hedy ist dafür gebaut, bei dieser Entwicklung vorne zu stehen. Lokale KI ist der erste konkrete Schritt. Weitere werden folgen.

Um sie zu aktivieren, aktualisieren Sie auf Hedy 3.2 und öffnen Sie in den Einstellungen den Bereich Speech & AI. Der Schalter heißt Local AI Processing. Wählen Sie ein Modell, das zu Ihrer Hardware passt, und Sie sind startklar.

Die technischen Details – welche Modelle wir ausgewählt haben, wie sie auf Mac, Windows und iPhone passen und was lokale Inferenz an Latenz kostet – finden Sie in unserem technischen Deep Dive zu lokaler KI (/de/post/local-ai-engineering-deep-dive-hedy-3-2/) .

Wenn Sie es ausprobieren, freuen wir uns auf Ihre Meinung.

Hedy AI · Live-KI-Coaching für wichtige Gespräche

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

https://www.hedy.ai/de/post/local-ai-meetings-hedy-3-2/