IA local para reuniões: o Hedy 3.2 agora pode rodar totalmente no seu dispositivo

O Hedy 3.2 permite executar todo o pipeline de IA no seu próprio dispositivo. Resumos, notas, respostas de chat e sugestões ao vivo acontecem localmente — nada chega a um servidor.

Publicado por Julian Pscheid · 23 de abril de 2026
Ler este artigo online: https://www.hedy.ai/pt/post/local-ai-meetings-hedy-3-2/

Uma advogada e seu cliente em uma consulta privada em um escritório jurídico, com um MacBook entre eles envolto por um holograma suave em ciano e roxo que sugere processamento de IA acontecendo no dispositivo

Com o Hedy 3.2, todo o nosso pipeline de IA agora pode rodar na sua própria máquina. Quando você o ativa, transcrições de reuniões, resumos, notas detalhadas, respostas de chat e sugestões ao vivo acontecem no dispositivo que capturou o áudio. Nada sobre suas conversas vai para um servidor. Veja o que muda quando a sua IA de reuniões funciona assim.

O reconhecimento de fala roda no seu dispositivo desde o dia em que o Hedy foi lançado. As gravações de áudio também sempre ficaram no dispositivo. Suas conversas nunca foram usadas para treinar modelos de IA. Fomos intencionais sobre privacidade desde o começo.

Mas sempre houve uma parte que não conseguíamos trazer para o dispositivo: o trabalho de IA em si. A parte que lê sua transcrição, escreve seu resumo, produz suas notas detalhadas, responde às suas perguntas sobre uma reunião e oferece sugestões enquanto você está nela. Esse trabalho precisava acontecer em servidores, porque os modelos capazes de fazê-lo bem eram grandes demais para rodar em um laptop ou em um celular.

Isso vem mudando rápido. Os dispositivos continuam ficando mais potentes. Os próprios modelos de IA continuam ficando menores e mais inteligentes ao mesmo tempo. Há alguns meses, essas duas tendências se cruzaram para nós. Modelos que cabem em um laptop ou em um iPhone recente agora são fortes o suficiente para lidar com a análise do Hedy em um nível de qualidade útil para reuniões reais.

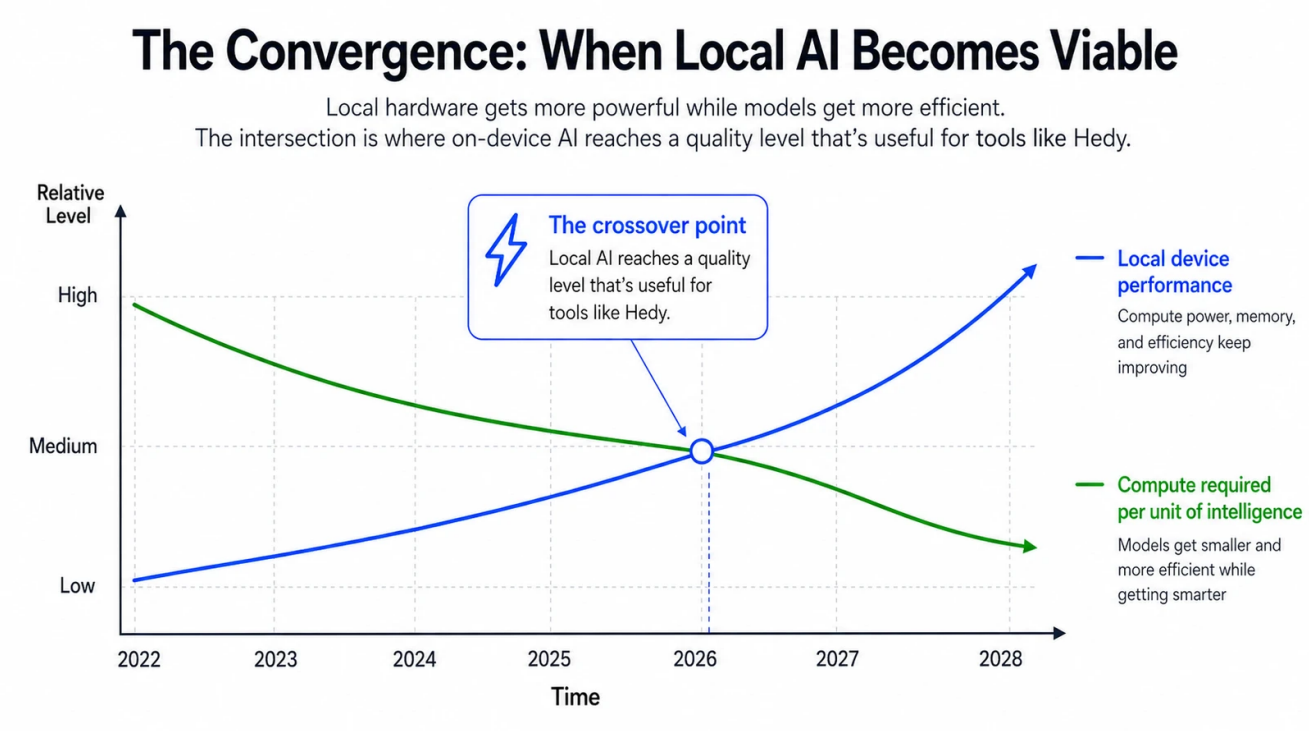

Gráfico conceitual mostrando o desempenho do dispositivo local subindo enquanto a computação necessária por unidade de inteligência cai, com um ponto de cruzamento marcado em torno de 2026 em que a IA no dispositivo se torna viável para ferramentas como o Hedy

Ilustrativo — as curvas mostram a tendência geral, não medições específicas.

Então, com o Hedy 3.2, você pode ativar o Local AI Processing e executar todo o nosso pipeline de IA no seu próprio dispositivo. Resumos, notas, respostas de chat, sugestões. Tudo isso acontece na sua máquina.

Por que a IA no dispositivo importa

Durante a maior parte dos últimos anos, a IA foi algo que um pequeno grupo de grandes empresas operava em seu nome. Você envia seus dados para os servidores delas, os modelos delas processam esses dados e os resultados voltam. Esse modelo tem vantagens óbvias, e impulsionou a maior parte do que conseguimos construir. Ele também tem um custo estrutural: a IA mais útil exige os dados mais pessoais, e esses dados ficam em algum lugar diferente de onde foram gerados.

A IA local muda essa lógica. Sua conversa fica onde aconteceu. O trabalho acontece no mesmo dispositivo que capturou o áudio. Nada sobre a reunião chega a um servidor, a menos que você escolha habilitar o Cloud Sync; e, mesmo nesse caso, o processamento de IA continua local.

Se você mantiver o Cloud Sync desativado, sua conversa existe apenas no dispositivo que a gravou. De ponta a ponta.

Quem se beneficia de uma IA de reuniões privada e no dispositivo

Alguns usuários do Hedy vão ativar isso e nunca mais pensar no assunto. Outros esperam por algo assim há anos. Estes são os grupos que, na nossa opinião, mais se beneficiam.

• Coaches e consultores cujas conversas com clientes carregam expectativas rígidas de confidencialidade. Eles já usavam o Hedy para preparação e chamadas internas. Agora podem usá-lo durante trabalhos com clientes sem que nada saia do laptop.

• Advogados que usam o Hedy em chamadas internas, mas nunca em conversas com clientes. O sigilo advogado-cliente tem uma forma específica, e "prometemos tratar seus dados com cuidado" não se encaixa nessa forma. Dados que não se movem, sim.

• Pacientes entrando em consultas médicas que querem uma recapitulação clara do que o médico disse, mas não querem que sua conversa de saúde fique em um servidor de terceiros. Com IA local, a recapitulação pode acontecer no mesmo celular que gravou a conversa.

• Jornalistas trabalhando em pautas sensíveis que evitam totalmente ferramentas em nuvem. Eles podem gravar uma entrevista, obter uma transcrição e conversar com a reunião, tudo sem que nada chegue a um servidor.

• Qualquer pessoa fora dos EUA que não queira que suas conversas fiquem em servidores americanos. Adicionamos residência de dados na UE no início deste ano. A IA local leva isso um passo adiante: os dados não ficam nos servidores de nenhuma empresa.

• Trabalhadores remotos em Wi-Fi ruim de avião ou em áreas rurais sem cobertura. O Hedy agora funciona totalmente offline. Abra o laptop em um voo, tenha uma conversa real e receba o resumo antes de pousar.

• Pessoas curiosas sobre privacidade que não têm uma profissão regulamentada nem um modelo de ameaça específico. Elas simplesmente preferiam a ideia de que a ferramenta ouvindo suas reuniões não estivesse enviando o áudio para lugar nenhum. Antes, não podiam ter isso. Agora podem.

O que une essas pessoas é que o Hedy já era útil para elas em conceito, mas o modelo de dados não se encaixava em suas restrições. A IA local remove essa restrição.

A versão honesta

Queremos ser claros sobre o que isso é e o que não é, porque preferimos que você comece com expectativas realistas em vez de se decepcionar depois.

O Local AI Processing é opcional e vem desativado por padrão. Cloud AI ainda é mais rápido, ainda produz resultados melhores e funciona em todas as plataformas compatíveis com o Hedy. Se você não tem um motivo específico para querer processamento no dispositivo, a opção em nuvem é a melhor experiência neste momento.

Um resumo que parece instantâneo na nuvem pode levar de 30 segundos a vários minutos localmente, dependendo do seu hardware e do modelo que você escolher. Modelos menores são bons para resumos curtos, mas podem se atrapalhar com conversas longas ou cheias de nuance. Modelos maiores chegam perto da qualidade da nuvem, mas precisam de hardware real para rodar bem. E nós não fazemos fallback silencioso para a nuvem quando algo falha localmente. Você optou pelo local por um motivo, e uma nova tentativa discreta nos nossos servidores trairia essa intenção. Você verá um erro em vez disso.

A IA local é compatível com Macs Apple Silicon, máquinas Windows com GPUs capazes, iPhones recentes (15 Pro e posteriores) e iPads M-series. Android e Web estão no roadmap, mas ainda não estão prontos. A variação no hardware Android e as restrições de rodar modelos dentro de um navegador tornam difícil entregar uma experiência consistente hoje.

Você escolhe o modelo que combina com seu hardware. Oferecemos três níveis de qualidade, desde modelos compactos que cabem em um celular, passando por opções intermediárias que funcionam bem na maioria dos laptops modernos, até modelos maiores que se aproximam da qualidade da nuvem em hardware capaz. O seletor de modelos mostra o que cabe antes de você fazer o download.

Para onde isso está indo

A IA local no Hedy 3.2 é um ponto de partida, não um produto acabado. Os modelos continuarão melhorando. O hardware de consumo continuará ficando mais capaz. A diferença entre a qualidade local e a da nuvem continuará diminuindo. Vamos expandir o suporte para mais plataformas conforme a tecnologia permitir.

Isso faz parte de uma mudança mais ampla que vale reconhecer. Durante anos, os maiores avanços em IA vieram de empresas rodando modelos cada vez maiores em fazendas de servidores cada vez maiores. Uma mudança mais silenciosa está acontecendo ao mesmo tempo: modelos com pesos abertos ficam mais inteligentes e menores a cada poucos meses. O hardware para rodá-los está no bolso e na mesa da maioria das pessoas. A diferença de capacidade entre a IA de reuniões na nuvem e a IA de reuniões no dispositivo está diminuindo rapidamente.

O panorama maior importa mais do que qualquer lançamento específico. Acreditamos que os próximos anos da IA serão definidos por uma mudança: de um mundo em que um pequeno grupo de empresas opera a IA em seu nome para um mundo em que você pode rodar seu próprio pipeline, no seu próprio dispositivo, com seus próprios dados, de ponta a ponta. Quer você escolha fazer isso ou não, ter a opção é o que dá à tecnologia a forma certa. Isso mantém o poder equilibrado.

O Hedy foi construído para estar na linha de frente dessa mudança. A IA local é o primeiro passo concreto. Haverá mais.

Para ativar, atualize para o Hedy 3.2 e procure em Settings, na seção Speech & AI. O botão está identificado como Local AI Processing. Escolha um modelo que se encaixe no seu hardware e pronto.

Para ver os detalhes de engenharia — quais modelos escolhemos, como eles cabem no Mac, Windows e iPhone, e quanto a inferência local custa em latência — veja nosso mergulho técnico em engenharia de IA local (/pt/post/local-ai-engineering-deep-dive-hedy-3-2/) .

Se você testar, adoraríamos saber o que achou.

Hedy AI · Coaching de IA ao vivo para conversas importantes

Experimente o Hedy grátis: https://www.hedy.ai/pt/downloads/

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