

# Tiles: 本地优先的AI技术架构

Ankesh Bharti · 2026-09-28 · 中文译文

## 演讲与问答全文

演讲与问答可读版逐字稿 · 中文译文。2026 年 9 月 29 日二次校订：对照 Soniox 转录、Zoom 字幕和演示材料，整理重复起句、断句及术语，保留发言顺序、例子、限定条件与讲者观点。中文逐段对照英文校订。时间从原录像 00:03:05.2 起计；省略等候及设备调试。部分疑点音频另经 Soniox 复核。仍无法确认的表述已标注；未逐句人工回听。

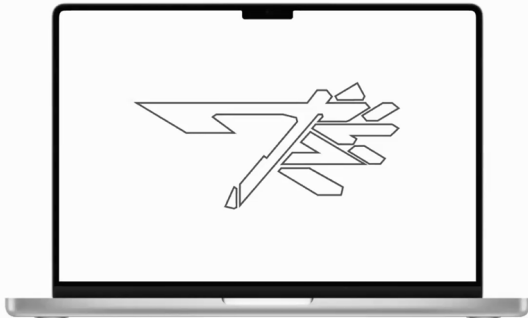
## 开场与背景



1 of 30

### Tiles: Own your AI with open models and decentralized protocols

Ankesh Bharti @feynon  
Founder & CEO, Tiles Privacy



 September 2026, Bengaluru

 © 2026 Tiles Privacy

Tiles 与用户自主权 · 00:00:24

00:00:00 · Ankesh Bharti

谢谢大家邀请我来分享。我是 Ankesh，Tiles Privacy 的创始人兼 CEO。今天我想谈谈，真正拥有自己的 AI 意味着什么：不仅是模型，还包括围绕它的数据、身份和基础设施。

00:00:28 · Ankesh Bharti

我认为，现在有两件事让这成为可能：开放模型最近的发展，以及去中心化协议的采用。



- CEO & Founder, Tiles Privacy
- Co-founder User & Agents community
- Technologist based in Pune
- Ambition is to build a decentralized stack that expands user agency in our digital spaces

分享者与 User & Agents • 00:01:04

00:00:38 • Ankesh Bharti

再介绍一下我自己。除了创办 Tiles Privacy，我也是 User & Agents 的联合创始人和社区维护者。我们在这个社区里讨论用户—智能体系统，希望让人们获得更多自主权、控制权和选择。Tiles 就诞生于这些讨论。我喜欢把自己称为技术工作者，目前住在印度浦那。

00:01:06 • Ankesh Bharti

我的目标是构建一套去中心化的技术栈，扩展用户在数字空间里的自主权。

We need to reclaim  
ownership of our digital lives  
in the age of AI through  
open models, and  
decentralized protocols

重新拥有数字生活 • 00:01:24

00:01:12 • [Ankesh Bharti](#)

这次分享的核心想法很简单：AI 正在成为我们思考、工作、研究和交流的一部分。随着这一变化，谁控制 AI 层，谁就开始控制我们数字生活中相当大的一部分。我认为，我们现在已经有技术来建立另一种关系：让智能在本地运行，在确实需要网络的地方使用开放协议。

# We all use AI agents now



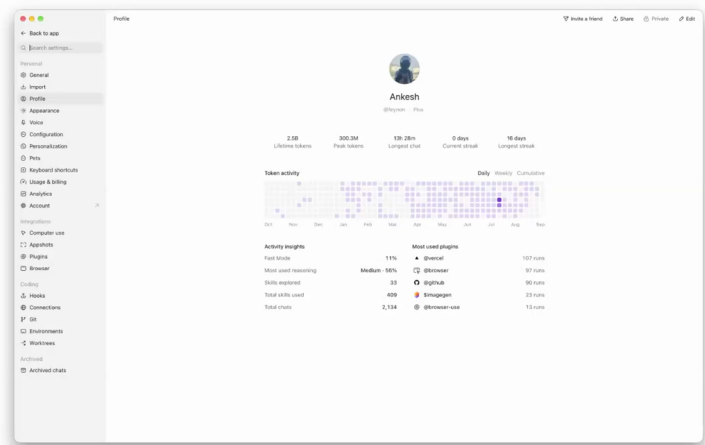
© 2026 Tiles Privacy

AI 智能体进入日常工作 · 00:01:44

00:01:40 · Ankesh Bharti

在座大多数人可能已经在使用其中几种 AI 智能体。它们正在变得远不止是聊天机器人。它们阅读代码、浏览网页、操作工具、处理文件、完成日常任务，而且越来越多地代表我们行动。因此，我们自然会把越来越多关于自己的上下文交给它们。

# I like ChatGPT



© 2026 Tiles Privacy

熟悉的 AI 使用体验 · 00:02:24

00:02:05 • Ankesh Bharti

我其实很喜欢 ChatGPT。让我持续使用它的，不只是模型，还有围绕模型的产品体验、一致性、记忆，以及能够在不同使用界面上接续同一份上下文。它逐渐不像是在用聊天机器人，而像是在用一个真正了解你的软件。但这种体验越好，双方的关系就越不对等。



依赖与控制 • 00:02:44

00:02:31 • Ankesh Bharti

系统越来越了解我，但这些上下文最终都留在一个账户里，留在我无法控制的基础设施中。如果我离开这个平台，日积月累的数字记忆中，越来越大的一部分就会被留在那里。我喜欢这种关系，但不喜欢这种依赖。我得出的结论不是停止使用 AI，而是需要更健康的边界。我希望保留同样的便利，而不必依赖盲目的信任。

# I needed boundaries

7 of 30

© 2025 Tiles Privacy

设定边界 • 00:03:04

00:02:58 • Ankesh Bharti

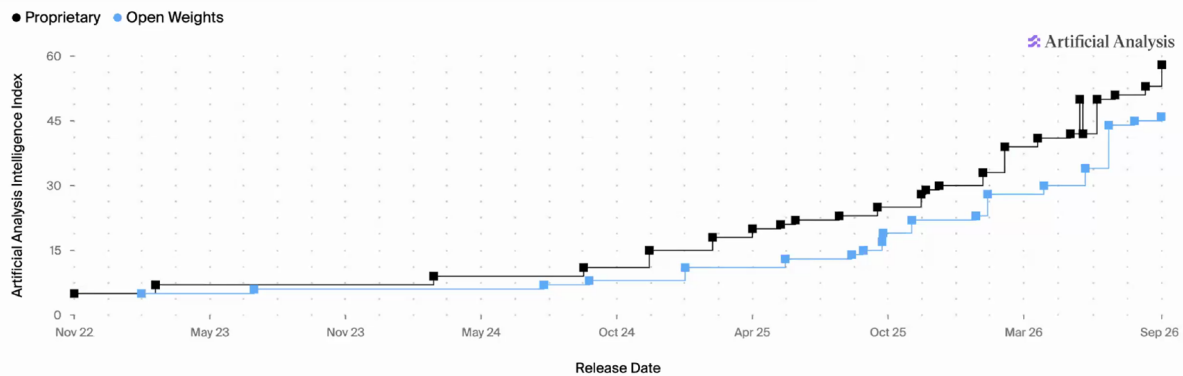
我希望身份和上下文能够跟着我，在所用的不同应用之间迁移；对系统的信任，也应当能通过计算来验证。我仍想保留这种关系，但希望由自己掌握其中的条件。

00:03:11 • Ankesh Bharti

因此，我的偏好变成了：私密、本地。好消息是，今天的本地 AI 工具在技术上已经相当出色，也天然适合私密的个人使用场景。

## Frontier is open-source today

9 of 30



© 2026 Tiles Privacy

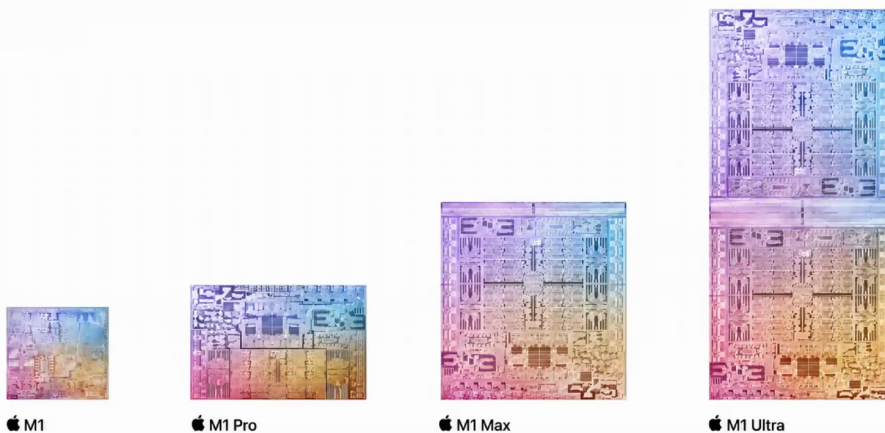
开放权重模型 · 00:03:24

00:03:23 · Ankesh Bharti

前沿 AI 也不再被锁在专有 API 后面。对于很多实际工作负载，开放权重模型已经具备前沿水平的能力，例如 GLM 系列。这里展示的是 Artificial Analysis 的图表。

## And there is compute to run them locally

10 of 30



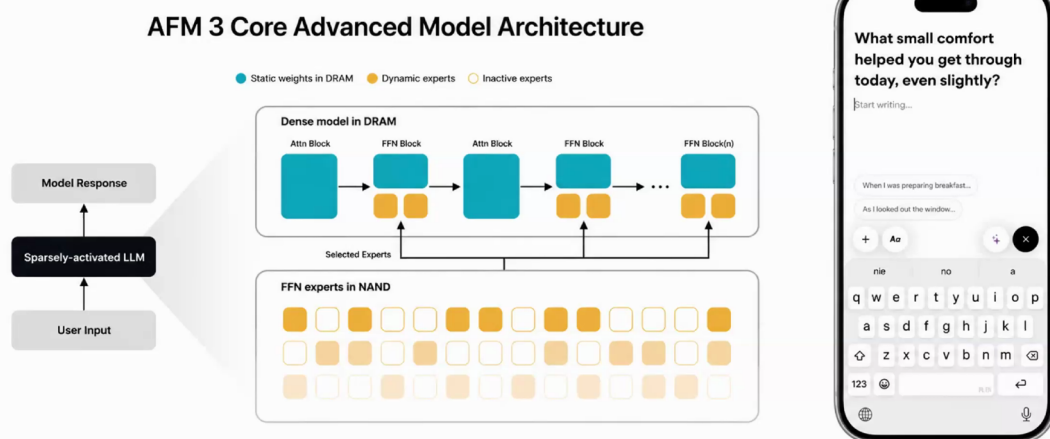
© 2026 Tiles Privacy

本地计算硬件 · 00:03:44

如今消费级设备拥有的算力，已经让我们有可能在本地运行前沿模型。像 UltraFusion 这样的技术，可以把两颗 Apple Silicon Max 芯片结合成一颗 Ultra 芯片，让中等规模的本地模型在设备上运行。我们也在寻找更聪明的方法，在边缘设备上运行更大的模型。苹果即将在 iOS 27 中推出的 Siri 由 Foundation Model 3 驱动，它是 Gemini 的蒸馏版本，可以直接在 iPhone 上运行一个 200 亿参数的混合专家模型。

## 20B models are now pocket-sized

11 of 30

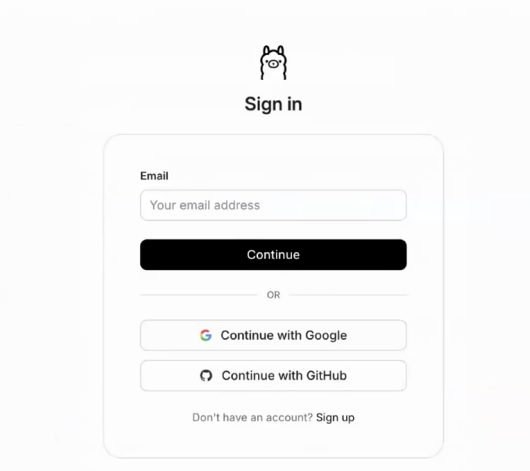


© 2026 Tiles Privacy

讲述中的端侧模型架构 • 00:04:24

它从高速闪存加载专家权重，而不是把整个模型都放在内存里。每次提示词只加载一次，而不是每生成一个 token 就加载一次，这样在很大程度上补偿了闪存较高的延迟。

## But collaboration breaks the private deal



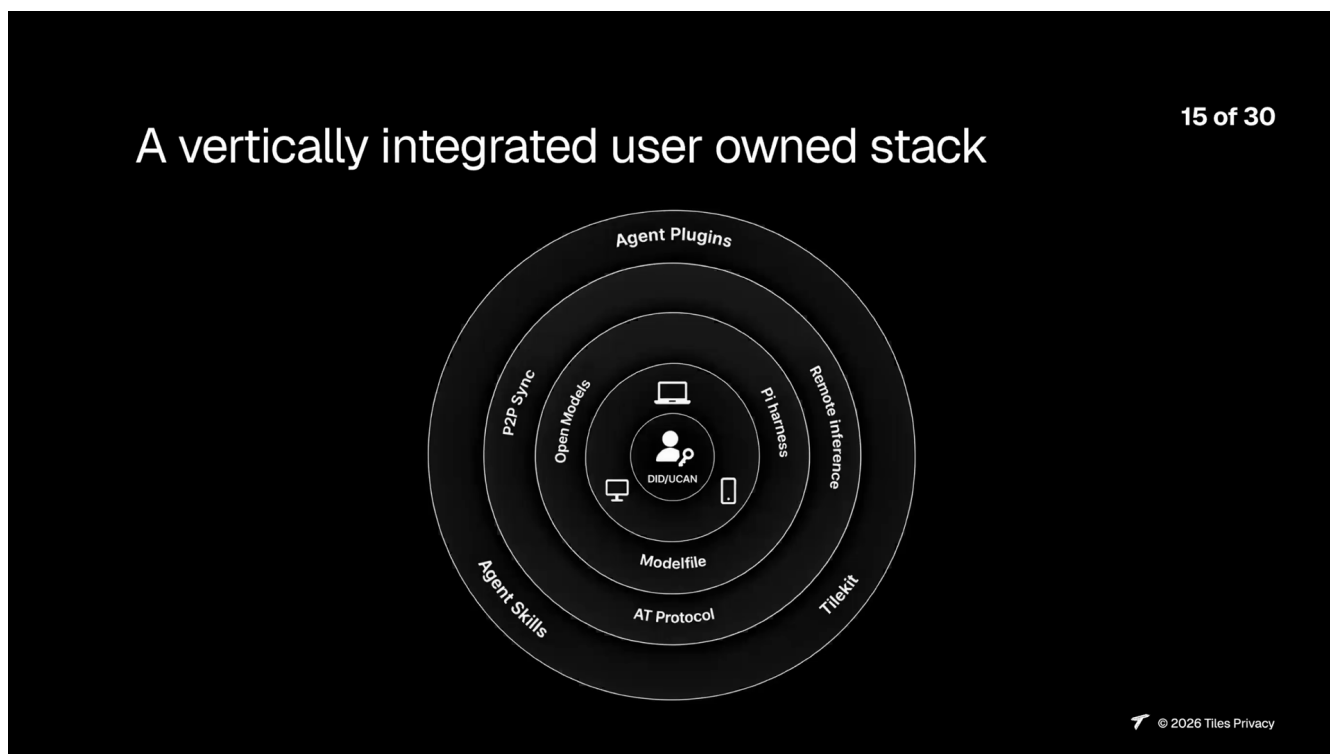
 © 2026 Tiles Privacy

协作与中心化账户 · 00:04:44

00:04:31 · Ankesh Bharti

但只有本地推理，还不能构成一个完整的产品。一旦我们需要身份、多设备、同步或协作，往往又会重新造出原本想摆脱的那套架构。所以，模型可以是本地的，产品关系却再次变得中心化。

## 以用户和密钥为中心



以用户与密钥为中心 • 00:05:04

00:04:51 • Ankesh Bharti

如果我们把用户以及用户的密钥放在中心，再向外构建，会怎样？密钥属于用户，保存在用户自己的设备上，而不是属于服务提供方。它们确定用户的身份，以及用户能够控制什么。然后才是设备、模型和其他一切。关键是，用户的所有权不应该止步于模型层。顺便做一个 AI 使用说明：接下来介绍的项目，主分支包含经过人工审查的代码，核心系统大部分由人工编写。

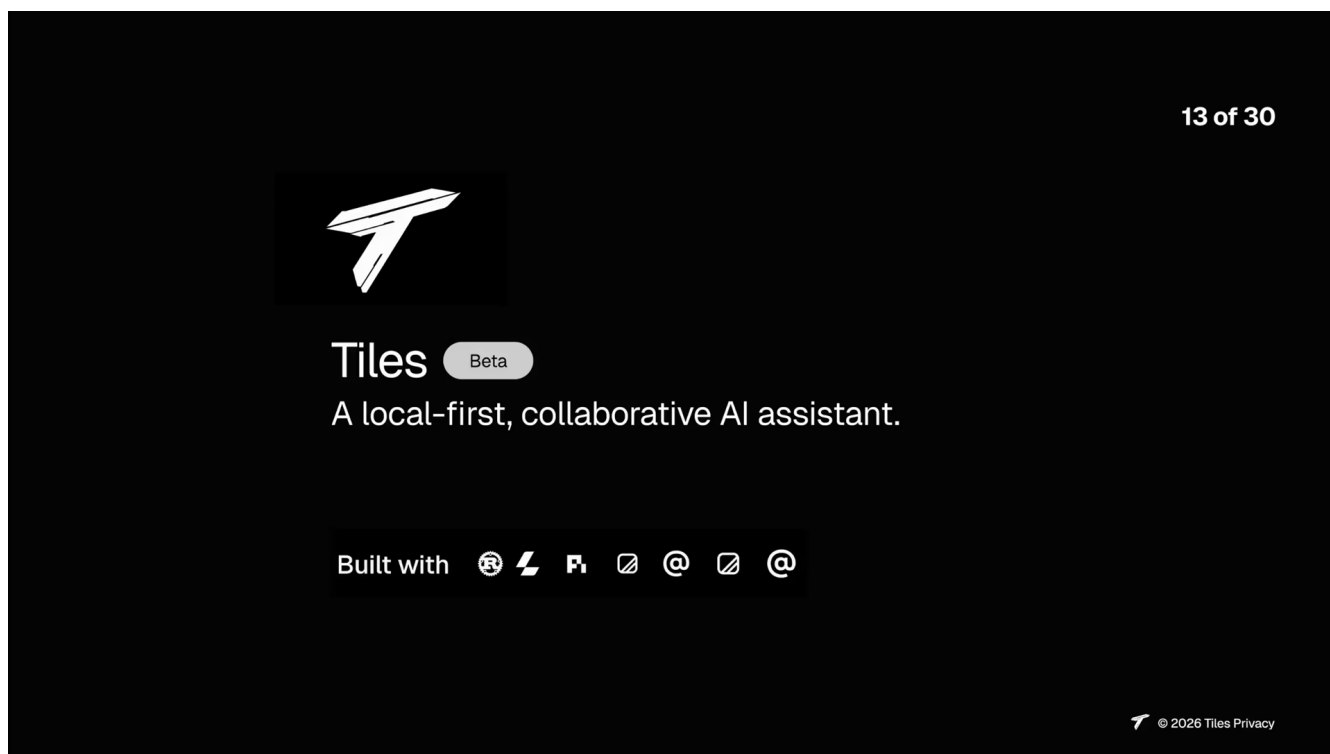
## AI disclosure

AI 辅助开发说明 · 00:06:04

00:05:40 · Ankesh Bharti

刚才说到，我们主分支的代码经过人工审查，核心系统主要是手写的。canary 分支则更积极地试验 AI 生成代码，主要用于客户端功能，以及 API 层之上的 UI/UX 工作。这帮助我们快速制作原型，与用户和团队测试想法，了解哪些东西真正有效。被证明有用的功能，会通过小规模、经过仔细审查的 PR 逐步进入主分支，而不是把实验性代码整批合并过去。

## Tiles 产品与现场演示

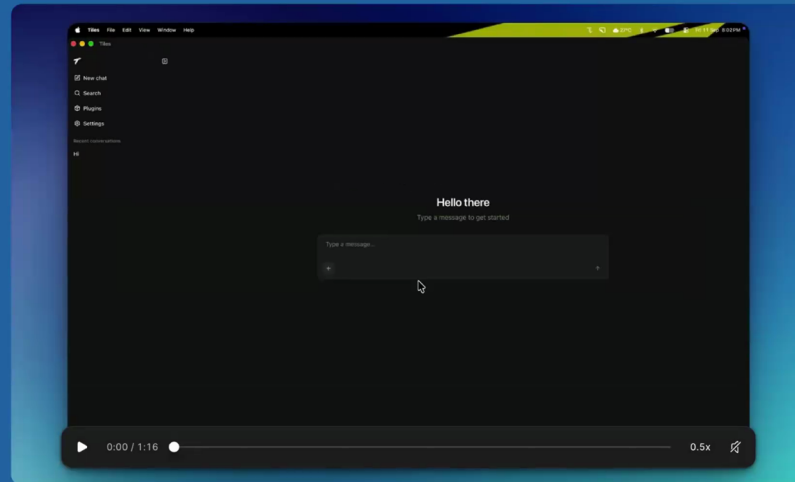


Tiles: 本地优先协作助手 • 00:06:44

00:06:15 • Ankesh Bharti

为了解决这个问题，我们构建了 Tiles。它是一个本地优先的协作式 AI 助手。这里的“本地优先”有很具体的含义：即使我们这个团队停止开发，产品也应该继续工作。你的数据、身份和核心体验，不依赖 Tiles 作为公司或服务继续存在。我们的目标是提供熟悉的 AI 体验，同时改变底层架构，让用户始终处于控制的中心。我们希望这个软件能比开发它的团队存在得更久。

# App and CLI



© 2026 Tiles Privacy

应用与命令行演示 · 00:07:04

00:06:50 • Ankesh Bharti

在继续介绍之前，我想先简短演示一下这个应用的样子。

00:07:36 • Ankesh Bharti

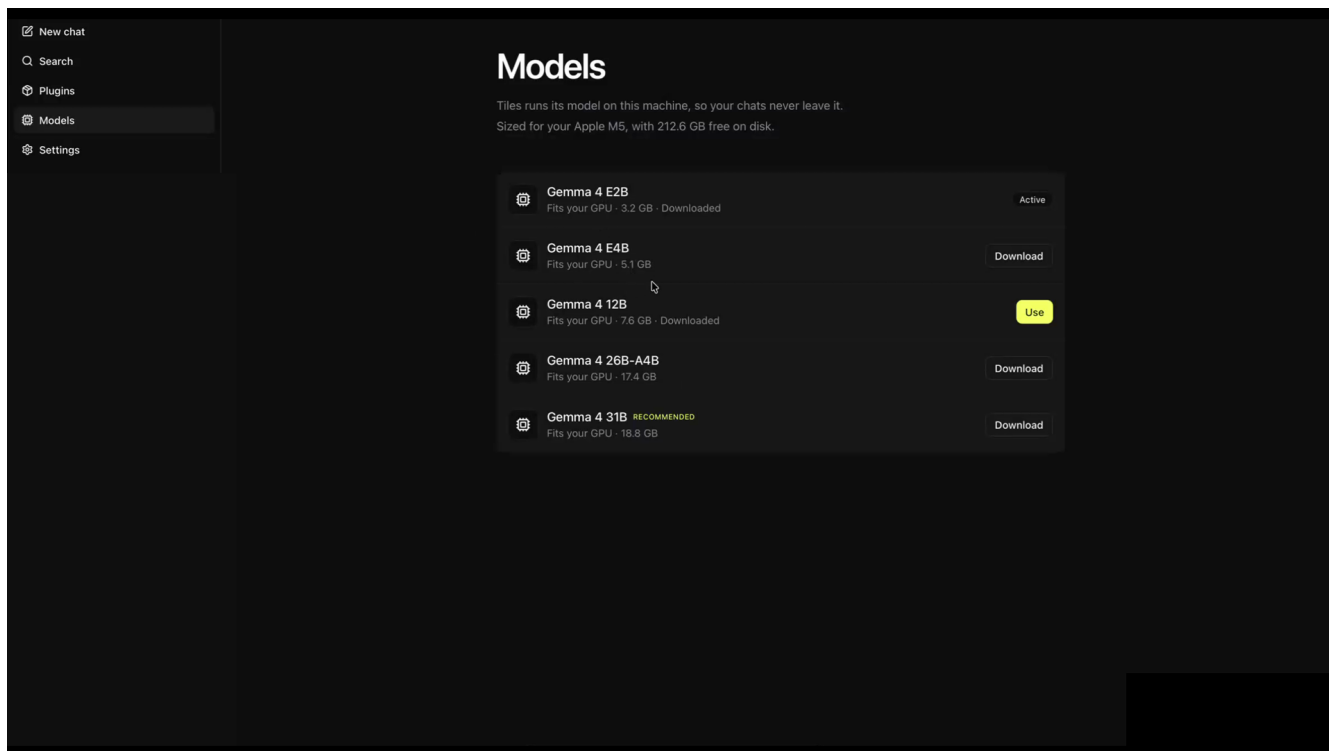
这是事先录制的演示。如果大家感兴趣，我可以在分享结束时现场演示。这是我在使用这个应用。我想查找 Mozilla Festival 2026 的日期，并把它保存到我的日历里。

00:07:58 • Ankesh Bharti

这里会用到插件。我觉得直接现场演示更好，所以我快速重新共享一下屏幕。

00:08:37 • Ankesh Bharti

好，我把提示词复制到这里。我们就随意地做一个现场演示。它看起来和 ChatGPT 或其他聊天应用差不多。我发送这条提示词：“帮我找到这个节庆会议的日期”，然后把它存进日历。这里使用的是设备上的模型。我们可以把介绍分成两部分：第一部分是 AI，第二部分是点对点、去中心化的网络部分。



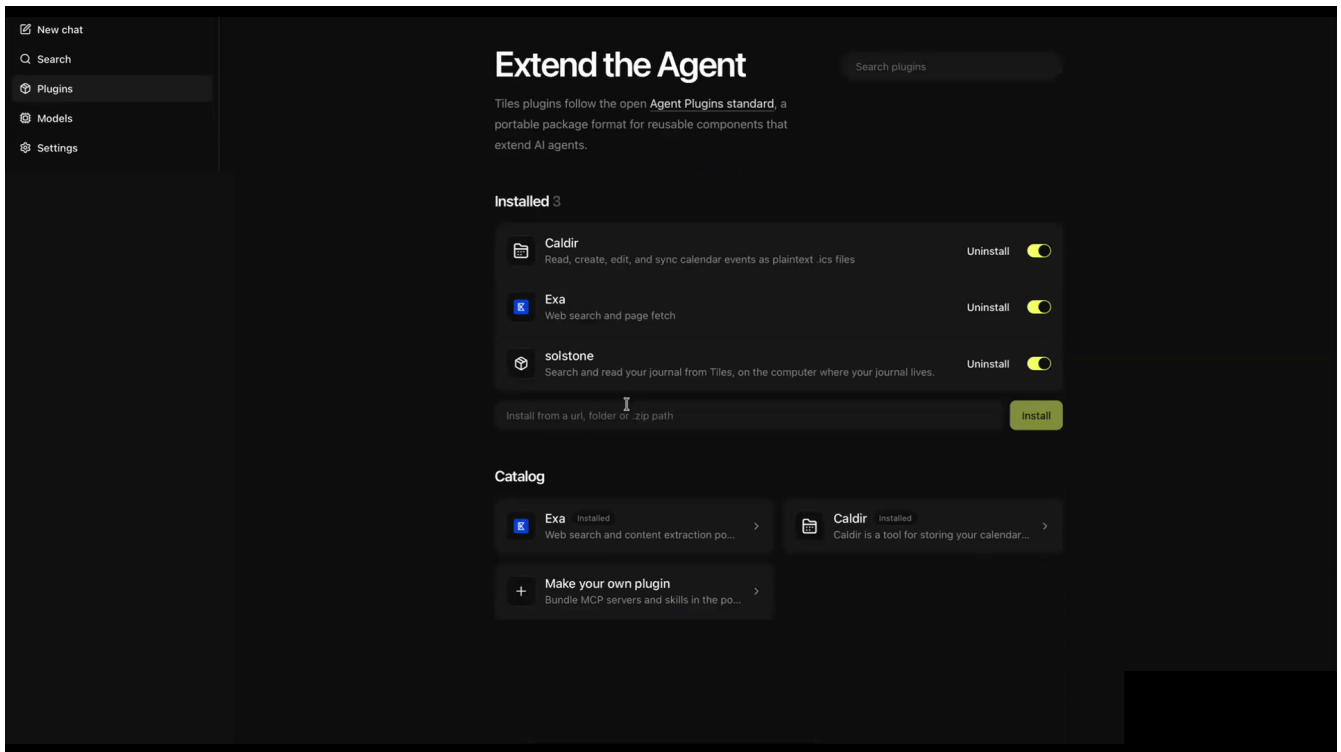
本地模型选择 · 00:09:24

00:09:12 · Ankesh Bharti

先说 AI：与 Ollama 相比有什么不同？我在 Tiles 中作了一些明确的技术选择：支持特定模型，并在 llama.cpp 推理层、Pi 智能体运行框架（harness）和提示词模板中完成配置。目前默认支持 Gemma 4 系列，让用户无需自行配置即可使用。

00:09:39 · Ankesh Bharti

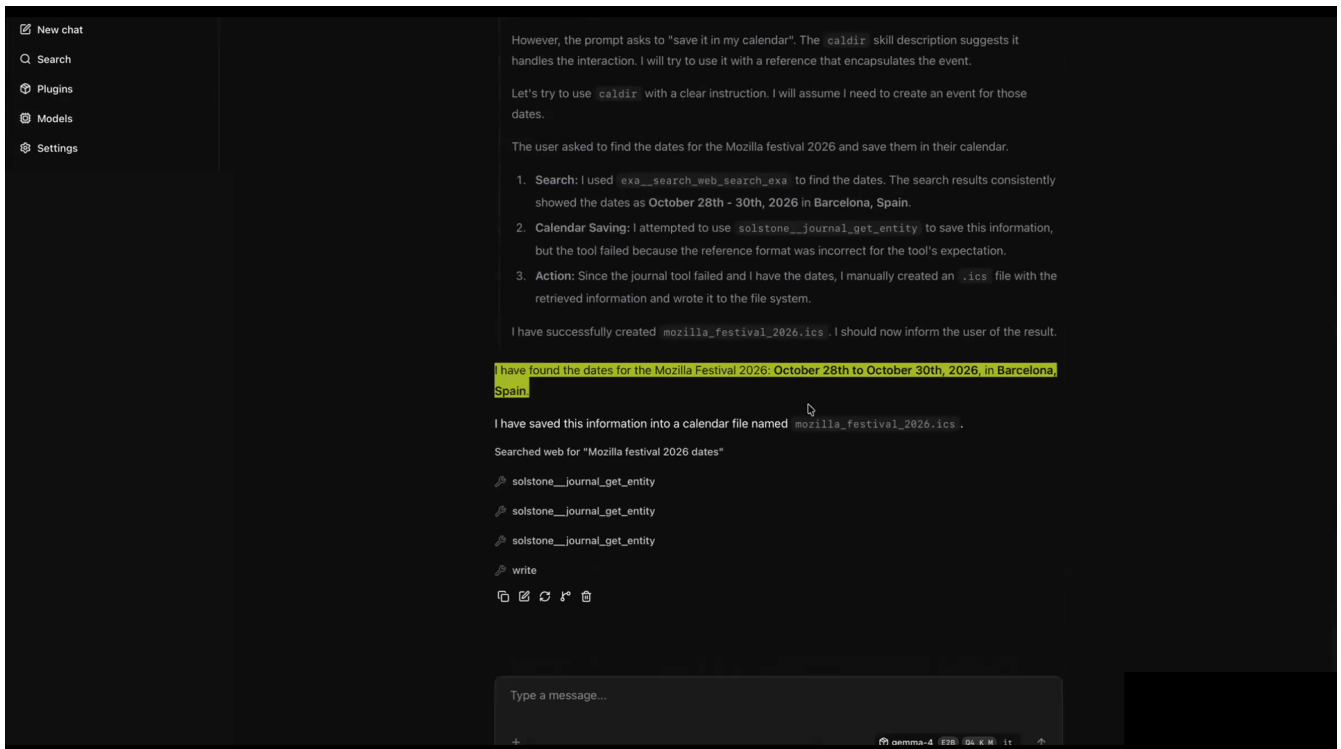
你也可以通过修改模型文件里的 from 字段，更换想使用的模型。目前这里支持任何 Hugging Face 端点，之后还会支持本地路径等方式。这就是我们在模型部分预先做好配置的实现。另外，我们还有插件，遵循 Agent Plugins 标准。



Agent Plugins 扩展 • 00:10:24

00:10:05 • Ankesh Bharti

ChatGPT 和 Cursor 使用的也是同一个标准，所以能在它们那里运行的插件，应该也能在这里运行。例如 Exa，这是一个网页搜索插件，由 MCP 和一个 skill 组成。你也可以在这里指定 ZIP 文件路径或 URL，使用其他插件。另一个是日历自动化插件：它是一个 CLI，附带的 skill 会说明如何使用这个 CLI。回到刚才的提示词，可以看到它进行了推理，调用了日历工具，也先调用 Exa 搜索了日期。



查找活动日期并写入日历 • 00:10:59

00:10:56 • Ankesh Bharti

这样就成功了，它把活动存进了我的日历。你还可以分享这些聊天。分享聊天时，内容会保存到 AT Protocol。现在好像没有正常工作……好，这次可以了。也就是说，分享聊天并不是由我们把它存进自己的数据库。

00:11:26 • Ankesh Bharti

内容保存在协议上，我会在后面的幻灯片里解释。你可以看到，它是从这个来源获取的。在 AT Protocol 上，各个应用共享同一个数据库，而这个数据库属于用户。用户可以自行托管，也可以由其他服务商代为托管，但用户拥有自己在所有应用中的数据。这就是这段聊天的公开记录。回去看，这些是我之前分享过的其他聊天。

00:12:00 • Ankesh Bharti

如果想私密地分享聊天，也可以做到。目前协议上的记录是公开的，所以我们的做法是把密钥放在 URL 中，将聊天记录以加密数据块的形式存储。在 PDS 上看到的是密文：记录虽然公开，没有密钥仍无法读取内容。

00:12:25 • Ankesh Bharti

密钥就在 URL 里，因此必须妥善保管这个链接。去中心化网络相关的功能还没有全部进入图形界面，目前大部分可以通过 CLI 使用，我们正在逐步把它们加入界面。

00:12:52 • Ankesh Bharti

你可以跨设备共享所有聊天，而不让聊天进入云端。我们采用的是客户端到客户端的数据传输、点对点同步。具体来说，我们有一个中继，稍后会在幻灯片中进一步解释。你可以在自己的设备之间同步聊天。这里还有一个基于去中心化标识符的本地账户，我也会在后面介绍。

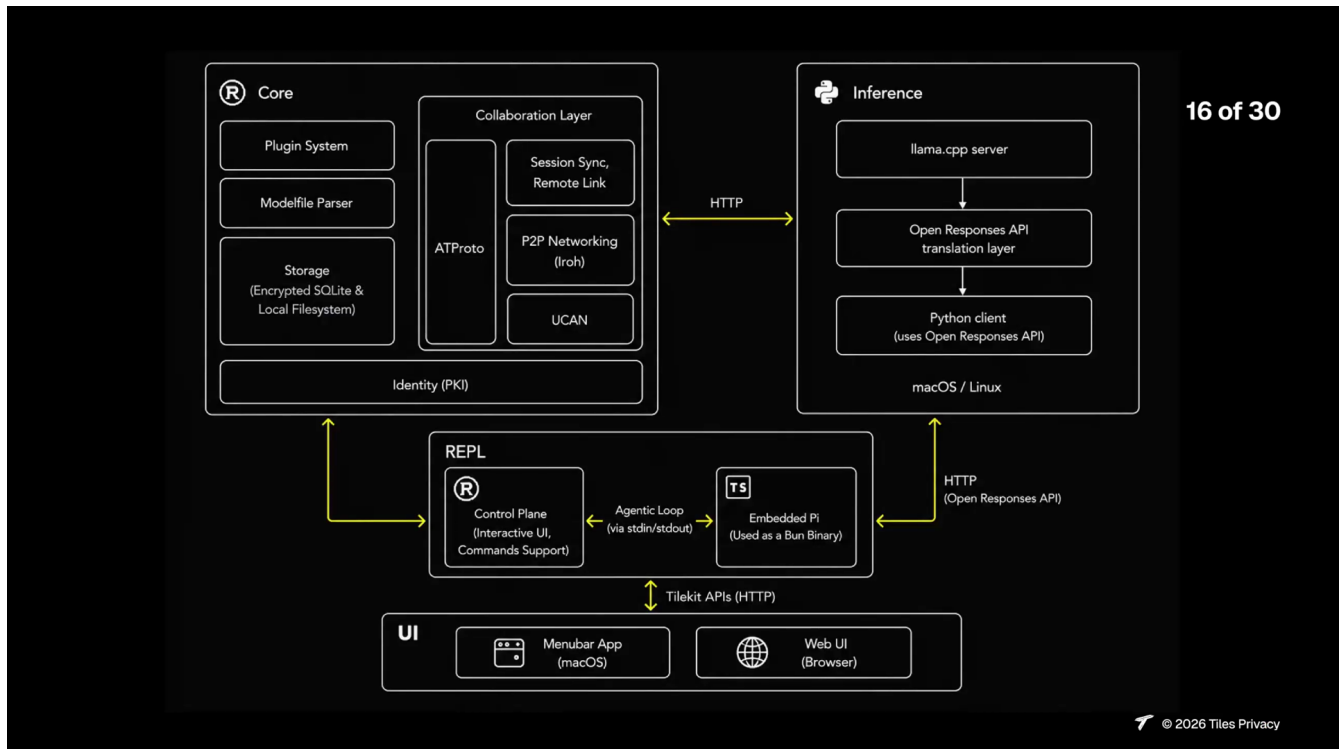
00:13:20 • Ankesh Bharti

这就是应用的大致情况。分享这个令牌后，可以从远程设备调用这台机器上的本地模型。如果想把 Mac 用作持续在线的推理设备，也可以让它在指定时间内保持唤醒。这里是我登录的 AT Protocol 账户，以上基本就是目前应用里的内容。

00:13:43 • Ankesh Bharti

现在回到幻灯片，解释这些东西如何工作的理论部分。我先展示一下应用，是想让大家更容易想象这个产品是什么。现在重新共享幻灯片。

## 技术架构



Tiles 技术架构 · 00:14:24

00:14:18 · Ankesh Bharti

这是技术架构，分成四个不同的模块。首先是 core，它用 Rust 编写，包含插件系统、模型文件解析器、加密的 SQLite、本地文件系统。所有点对点网络功能都基于 iroh，零信任的身份认证与授权则使用 DID 和 UCAN。这里还包括钥匙串、公钥基础设施，以及同步和远程推理等功能。第二个模块是推理部分，它通过 HTTP 与 core 通信。其中包含 llama.cpp server，以及 Open Responses API 的转换层。这基本上是 OpenAI 目前使用的标准，但行业里的采用还不多，因为大部分应用仍在使用 completion API。

00:15:13 · Ankesh Bharti

由于上一代 Tiles 默认支持 GPT-OSS，我们必须实现 Open Responses API，让 GPT-OSS 模型达到最佳效果，所以现在仍然使用它。客户端一侧有 CLI 和 UI，Web UI 全部基于 Tilekit API。Tilekit 实际上是一组 REST API 端点，暴露 Tiles 的全部功能。

00:15:40 · Ankesh Bharti

开发者可以把这些能力嵌入自己的应用，稍后还会谈到。我们也把 Pi harness 作为 Bun 二进制文件嵌入。接下来讲用户持有的密钥与 Tiles 账户：DID，即去中心化标识符，提供身份原语；UCAN，即用户控制的授权网络，提供授权原语。两者共同构成 Tiles 账户，在本地生成，并通过设备钥匙串保护，而不是由中心化的 Tiles 服务器创建和控制。

## 1. User owned keys with Tiles Account

17 of 30

**DID**  
Decentralized Identifiers

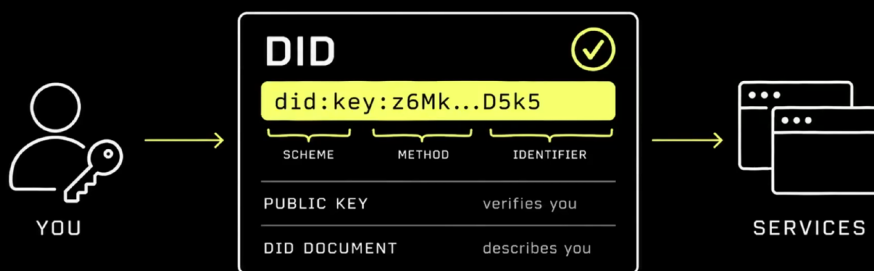
 **UCAN**  
User Controlled Authorization Network

© 2026 Tiles Privacy

用户持有的密钥 • 00:16:04

## DID is who you are

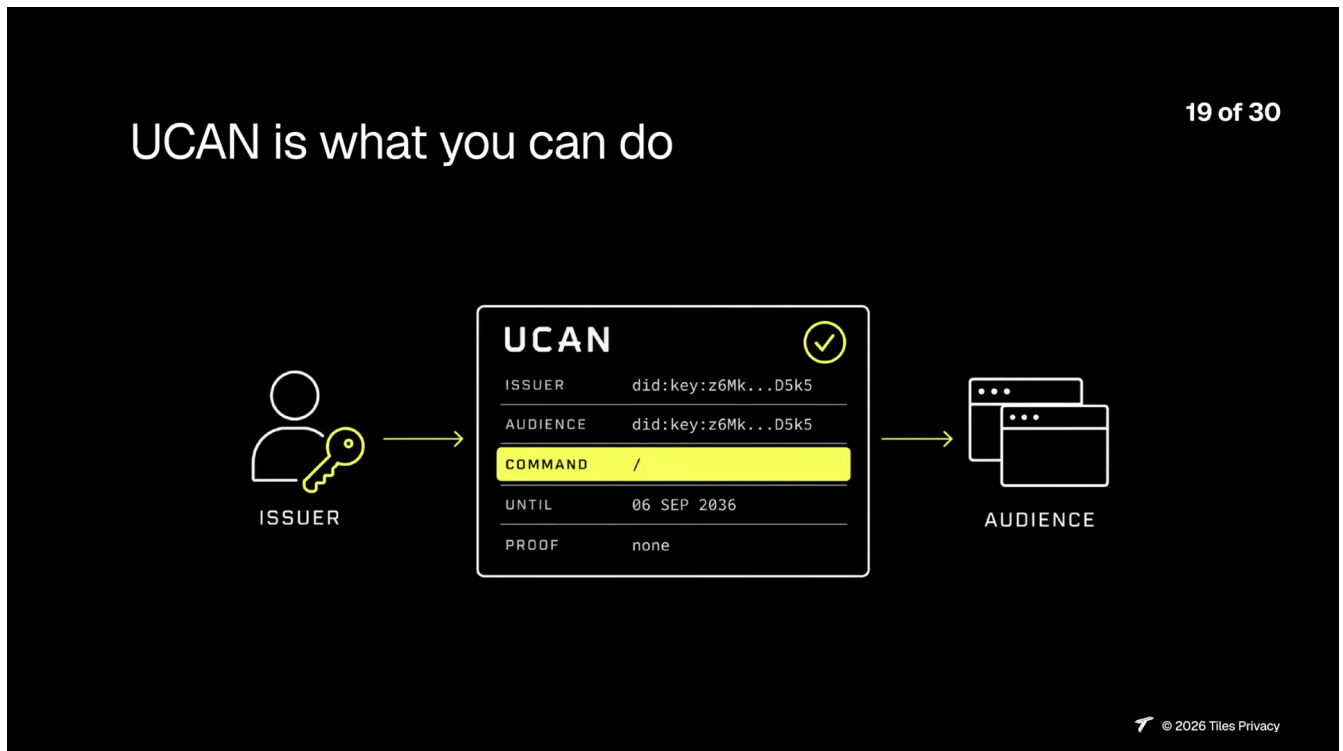
18 of 30



© 2026 Tiles Privacy

DID: 身份 • 00:16:44

也就是说，我们不使用本质上中心化的 SSO 或 OAuth，而是使用可以点对点工作的 DID 和 UCAN。DID 回答的是“你是谁”：这是一个由你的密钥控制、可迁移的标识符。服务可以验证你的身份，但服务本身不必成为你身份的来源。



UCAN: 权限 • 00:17:04

如果 DID 回答“你是谁”，UCAN 回答的就是“你被允许做什么”。在 Tiles 中，你可以提供有范围限制的授权，用于在设备之间、不同用户之间同步聊天。UCAN 使用基于能力的授权，而不是中心化系统中常见的访问控制列表，也就是 ACL。你不需要让中央服务器去检查权限清单，而是出示一个经过签名、授予特定权限、也可以带有有效期的能力凭证。

## 2. Local-first AI with Tiles Network

20 of 30



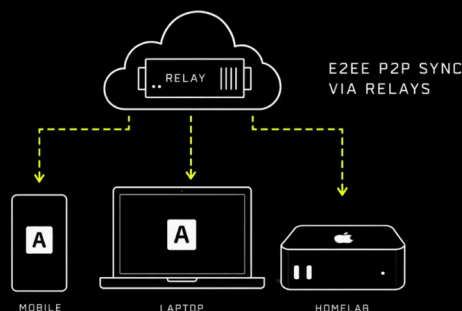
© 2026 Tiles Privacy

Tiles 网络 • 00:17:44

00:17:14 • Ankesh Bharti

因此，我不必给另一台设备一个永久密码，而是只授予它所需的权限。这更适合本地优先和点对点系统。第二部分是通過 Tiles 网络实现本地优先 AI，覆盖本地、在线和社交几个层面。你可以使用本地设备、身边的算力、私密的 P2P 网络和开放的社交层，无需让一个中央 Tiles 服务器位于中间。基本原则很简单：让所有权靠近用户，只在网络能带来价值的地方引入网络。

## Iroh based P2P sync through public relays



© 2026 Tiles Privacy

iroh 点对点网络 • 00:18:24

00:17:47 • Ankesh Bharti

除了日常使用的手机、笔记本，还有我称为“身边的 AI 算力”（ambient AI compute）的设备：家庭实验室、办公室工作站，或者任何由你控制、能在后台按需提供推理的电脑。它们可以通过局域网直接通信，完全没有互联网也能工作。跨网络连接时，Tiles 使用端到端加密的点对点网络，以及基于 iroh 的公共中继。

00:18:15 • Ankesh Bharti

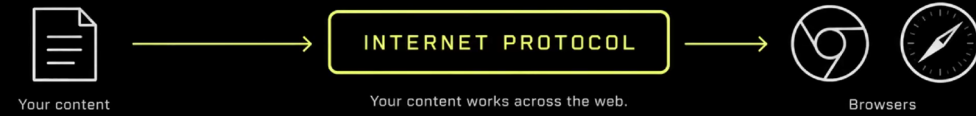
也就是说，中继只是用于发现 IP 地址，实际的数据传输直接通过 UDP 数据包在客户端之间进行。

00:18:25 • Ankesh Bharti

数据不会到达中继。中继只看到 IP，仅此而已，从而让网络上的设备能够被发现。〔编注：以上为讲者的描述。一般而言，iroh 中继在建连阶段或无法直连时可以转发端到端加密的数据，但无法读取明文。这条说明不代表已核实本次 Tiles 版本的中继策略。〕另外，这些中继也可以运行在本地 Wi-Fi 或局域网中。如果你有自己的局域网，就不需要那个托管中继。接下来是社交和协作功能：我们使用 AT Protocol，以及用户自己的个人数据服务器。

## Collaborative features on AT Protocol

### OPEN WEB



### OPEN SOCIAL WEB



© 2026 Tiles Privacy

AT Protocol 协作层 • 00:19:04

00:18:54 • Ankesh Bharti

这里可以类比开放网络。你的网站能在不同浏览器中工作，因为浏览器不拥有底层的 Web。AT Protocol 把类似的想法带到社交层。你的数据存在自己的仓库中，可以把它想成数字生活的 GitHub 仓库。不同应用可以读取和写入同一个仓库，而不是每个应用都把身份和数据困在自己的数据库里。

### 3. Opinionated AI stack



模型与插件技术栈 • 00:19:44

00:19:21 • Ankesh Bharti

因此，你的身份和内容可以在多个应用中使用，更换应用不意味着从头开始数字生活。第三部分是我们预先选定并调优的技术栈，默认使用 Gemma 4 系列。我们处理 llama.cpp、Pi harness 和量化中的模型特定调优，你无需自行配置所有东西。扩展方面，我们支持 Agent Plugins 标准，可以使用来自 Cursor 或 ChatGPT 等客户端的兼容插件。

**Server**

```
GET    /server/start
GET    /server/stop
GET    /server/ping
```

**Agent**

```
GET    /agent/start
GET    /agent/state
POST   /agent/prompt
GET    /agent/end_session
GET    /agent/reload
```

**Session**

```
POST   /session/new
GET    /session/list
GET    /session/{session_id}/chats
POST   /session/chat
```

**Account**

```
POST   /account/create
GET    /account/status
```

**ATproto**

```
POST   /atproto/login
POST   /atproto/logout
GET    /atproto/status
POST   /atproto/share-session/{session_id}
```

Tilekit API • 00:20:04

00:19:54 • Ankesh Bharti

最后，你还可以通过 Tilekit，把 Tiles 嵌入自己的产品。Tilekit 是一组 REST API 端点，面向希望在自己应用中使用底层运行时和能力的开发者。它让 Tiles 可以被嵌入，从而让大家在同一套底层基础设施上构建自己的界面和产品。这与 Codex 通过 Codex App Server 让你嵌入其 harness 的方式类似。这是相似的设计。

00:20:27 • Ankesh Bharti

在结束之前，还有一件事想和大家分享。我们计划下个月宣布与 Mozilla Foundation 的一项重要合作，也正在寻找组织和客户，作为试点用户和设计合作伙伴，一起完善我们构建的产品。



We're announcing a major partnership with the Mozilla Foundation soon and are looking for organizations to help pilot and test what we're building.

Advisors



**Boris Mann**  
[@bmann.ca](#)

COO, Ink & Switch  
Project Lead, AT Community Fund  
Co-founder & CEO, Fission Labs



**Dietrich Ayala**  
[@autonome](#)

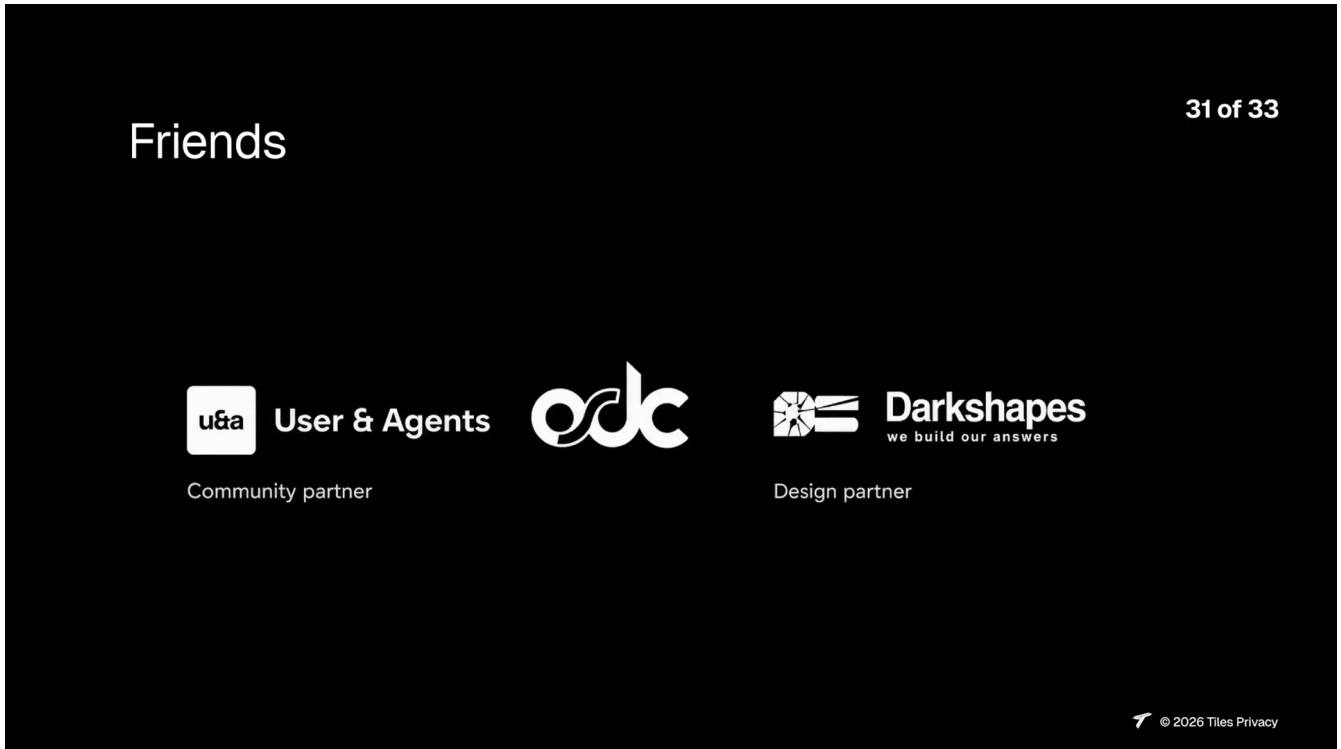
Co-founder, Users & Agents  
Founder, Web Transitions  
Ecosystem Lead and Advisor, Protocol Labs  
Developer Advocate, Mozilla



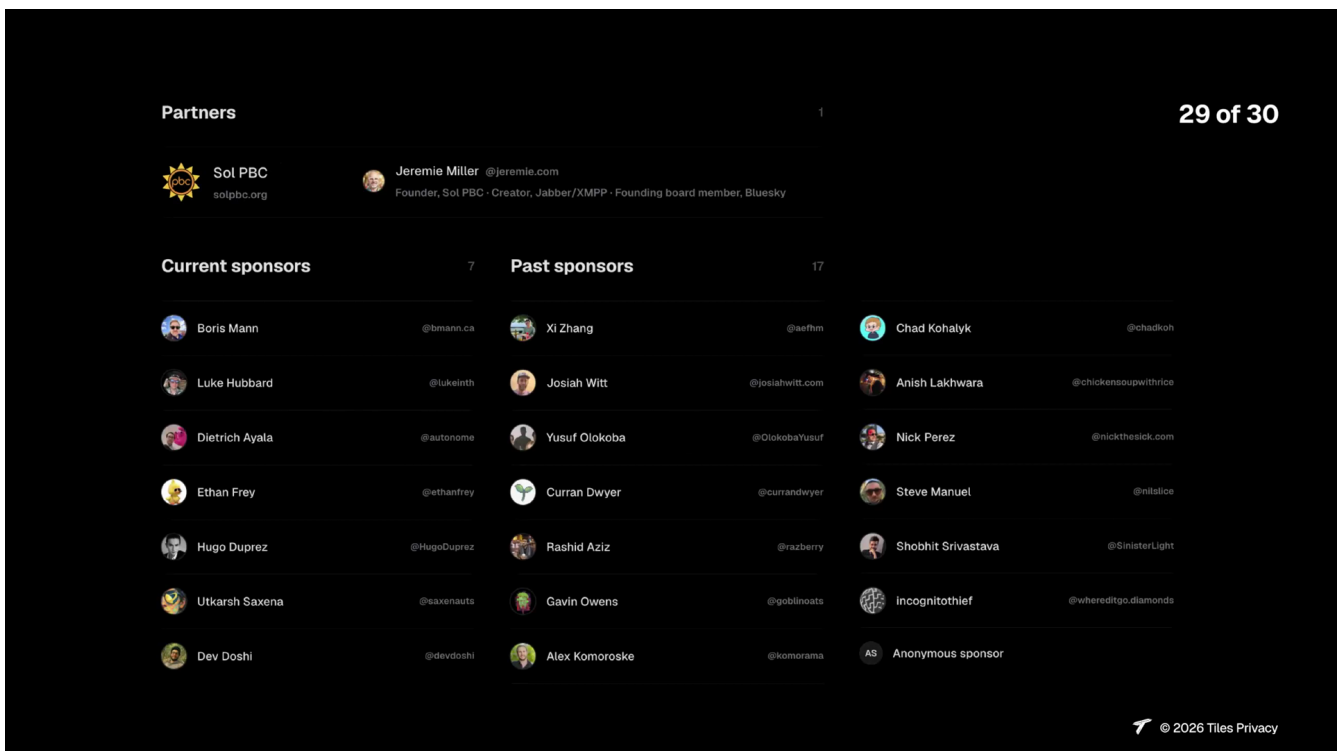
**Gordon Brander**  
[@gordon](#)

Research Fellow, Future of Life Foundation  
Research Fellow, Cosmos Institute  
Founder and CEO, Subconscious  
Web Platform Designer, Google  
Design Engineer, Mozilla

过去一年里，我们成长为五人团队，拿到 Mozilla Foundation 的资助后，将从下个月开始全职工作。我很感谢 Boris Mann、Dietrich Ayala 和 Gordon Brander，感谢他们根据自己在去中心化系统、浏览器和开放网络方面的经验提供建议和指导。Tiles 确实是开源社区的产物。User & Agents 和我们大学的开源社区 OSDC 都给了我很大的帮助。



社区与设计伙伴 • 00:21:24



项目支持者 • 00:21:44

00:21:21 • Ankesh Bharti

还有 Darkshapes，他们设计了标志，并帮助我们想法转化为 Tiles 的基础视觉形象。特别感谢 Solstone 及其创始人 Jeremie Miller，作为紧密合作伙伴和赞助者支持 Tiles。也感谢其他所有在过去一年里赞助 Tiles、帮助项目持续运行的人。

## 演讲小结



参与和交流 • 00:22:04

00:21:46 • Ankesh Bharti

回顾一下，我喜欢 ChatGPT，但不想在便利和控制权之间妥协。本地 AI 与开放协议让我们保留这种便利，同时使数据、身份和记忆可迁移，真正属于我们自己。希望大家在 Mac 或 Linux 上试用 Tiles，并告诉我你的想法。如果想参与贡献，或聊聊本地优先 AI，欢迎扫描二维码加入我们的新 Discord 服务器。其他事情也可以发邮件到这个地址。

00:22:14 • Ankesh Bharti

我的分享就是这些。如果大家还想了解 Tiles，有任何问题，或者想看看 CLI 的样子、有哪些取舍，以及其他技术问题，现在都可以交流。

## 交流与问答

00:22:46 • Ankesh Bharti (代读问题并回答)

“当两台设备上的智能体各自写入记忆，可能重复或相互矛盾时，你计划如何合并？”目前，聊天同步由用户控制，我们只是共享聊天记录，也就是对话线程。现在可以把聊天同步到另一台设备，但还不能在那里接续原来的会话。

00:23:17 • Ankesh Bharti

我们还没有深入处理这些合并问题。项目很早期，目前是 alpha 阶段，Tiles 的第二次迭代才开发了六七个月。我们可能在接下来几个月里考虑这个问题。

00:23:49 • Ankesh Bharti

一个正在探索的方向，是让每段聊天成为一个沙箱。目前分享聊天，只会分享记录；之后，每段聊天都可以成为基于 `microsandbox.dev` 的微型虚拟机沙箱。例如，研究者安装了 Python 包，再把聊天分享给朋友，朋友无需重新安装，因为共享的是整个计算状态。微型虚拟机也能让这个环境保持轻量。

00:24:41 • Ankesh Bharti

我很期待这个方向。微型虚拟机可以创建快照、暂停和恢复。我想，这能让智能体在彼此的计算环境中协作，就像 Grok bot [名称仍听辨不清] 那样，同时帮助解决合并问题。

00:25:07 • Ankesh Bharti

我看到一个链接：`centaur.run`。面向 Slack 的多人、自托管、安全的智能体。好的。

00:25:16 • Ankesh Bharti

我们也在开发群聊，计划在接下来六个月里做这件事，所以看起来有些相似。

00:25:27 • Ankesh Bharti

我的目标就是，让自己的整个数字生活处在一种由我控制账户、模型、数据和一切的状态。

## 智能体协作与共享环境

00:25:54 • 观众（聊天提问）

你们是否计划建立某种结构，让智能体之间直接协作？例如使用 A2A 协议让它们彼此交流，或者让它们访问一个共享看板？

00:25:58 • Ankesh Bharti

有这个计划。具体协议还没有想好，我们可以不绑定某一种协议。基本想法是让每段聊天本身就是一个微型虚拟机。这样，你就可以邀请一个智能体进入某个对话线程并与它协作，也可以邀请朋友进入这个线程，一起协作。

00:26:26 • Ankesh Bharti

关于共享看板，我最近看到 LM Studio 的一个产品，似乎集成了可供智能体访问的画布。我还没试过，但看起来与这个想法接近。它是在 LM Studio 的新本地 AI 应用 Bionic [应用名仍听辨不清] 里发布的。

00:26:51 • Ankesh Bharti

在最初的十二到十八个月里，这个产品的用户主要会是学生。至少在开始阶段，我们会与印度的教育机构紧密合作，了解适合他们的使用场景。

## 本地模型与硬件限制

00:27:09 • 观众

你好，谢谢你的分享。我对电脑算力这部分有些好奇，也就是本地运行语言模型。我试过几种，我的电脑是配备 24 GB 内存的 MacBook M4，也试过 Gemma。问题是，即便运行一个很小的任务，电脑也会很烫。所以我想知道，怎样才能让它在中等配置的电脑上更可行？

00:27:53 • Ankesh Bharti

现在也有更小的模型，比如 Gemma E2B。关键是本地模型是否擅长工具调用；如果擅长，我认为就能处理大部分场景。至于发热降频，现在还很早。我记得最近的 iPhone 19，或者 iPhone 18，NPU 容量翻了一倍；还有 [芯片或设备型号听辨不清] Pro Max，也有翻倍的 NPU，帮助处理预填充 (prefill)。  
[讲者对型号并不确定；此处保留原述，不作为已核实的产品规格。]

00:28:31 • Ankesh Bharti

我们都是喜欢动手试验的人。我认为，硬件可能还需要一两年才能赶上，但最终会做到。现在很适合进行实验和探索，参与这个过程，影响它未来的发展方向。

00:28:53 • 观众

好的，谢谢。

00:28:59 • Ankesh Bharti

坦白说，我在自己的设备上运行这些本地模型时，也会发热。但我并不太担心，因为我认为更好的硬件最终会解决这些问题。我甚至希望 AI 泡沫破裂，这样我们就能买到更便宜的硬件。以前互联网泡沫破裂的时候，也发生过类似的事。

00:29:29 • 观众

我个人觉得要看场景。一些确定性的任务，可以先用脚本解决一部分，再重新设计工作流程，把部分能力交给本地语言模型，这样就很有用。

00:30:07 • Ankesh Bharti

还有，我们集成 DID 和 UCAN，也就是去中心化的授权与身份认证，我认为这会非常有助于 A2A 之类的协议。因为目前即使用 A2A，智能体能够彼此交流，仍然需要通过某处的中心化服务器来协调。它们没有那样的自由或自主权，用户没有，智能体也没有。

## 结束交流

00:31:23 • Ankesh Bharti

很高兴和大家交流。希望有一天可以在线下见到这个社区，参加在中国举办的线下活动。

00:31:55 • Ankesh Bharti

也想为今天有些疲惫向大家致歉。过去一周非常忙，我在 IndiaFOSS 做了分享，也设了会议展位。这是印度最大的数字公地与开源会议。所以确实很累，但今天已经尽力分享了。欢迎通过社交平台联系我。我会把个人网站链接发到聊天里，在网页脚可以找到所有社交链接。

00:32:28 • Ankesh Bharti

我也会把正在开发的产品链接发出来，大家可以看看。我这边就到这里，谢谢大家邀请我。