渐进式 AI Web 应用 (PAWA)

渐进式 AI Web 应用（PAWA）是一种为了解决 Web 应用在网络不佳、性能限制或安全隐私等情况下，混合使用本地和远程大语言模型的解决方案。

与渐进式 Web 应用（PWA）类似，PAWA 将 Web 应用的 AI 能力扩展到本地设备（例如 PC 端或移动端），确保即使在网络不佳、性能限制或安全隐私等情况下也能提供基本的 AI 功能和服务。它允许 Web 应用渐进式增强 AI 功能，以适应不同场景和资源的限制。

1. PAWA 的定义

PAWA 的核心思想是将 AI Web 应用分为两层：

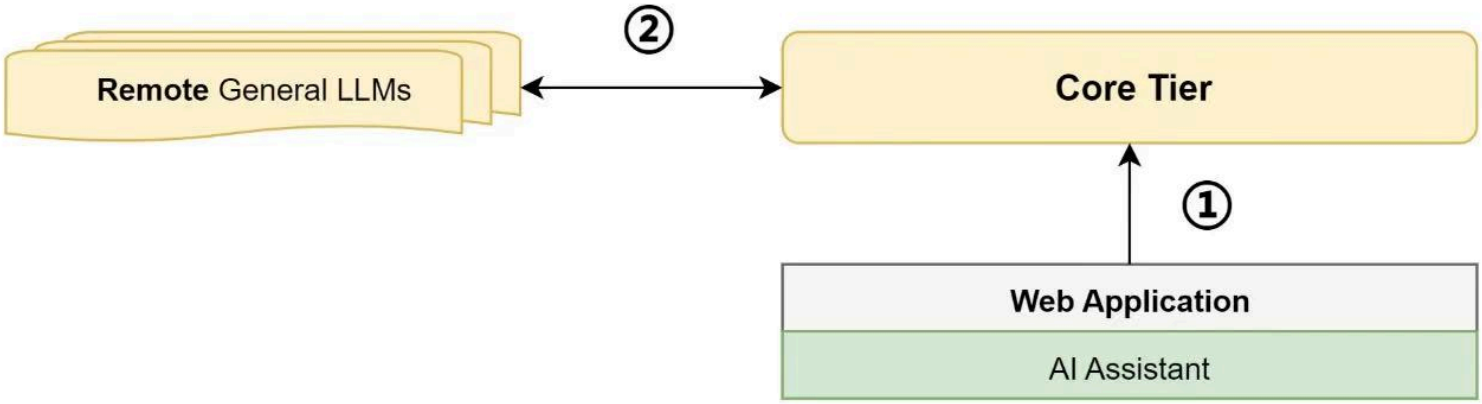
- **核心层：** 包含 AI Web 应用的核心功能和逻辑，通常需要远程模型服务的支持，以及本地层的支撑。
- **本地层：** 因场景和资源的限制无法访问远程模型服务时，可以提供本地模型服务替代方案或降级功能。

通过这种分层架构(该架构在下个章节描述)，PAWA 能够在条件允许的情况下充分利用远程模型的强大功能，并在网络不佳、性能限制或安全隐私等情况下退回到本地层，确保 AI Web 应用的可用性和鲁棒性。

2. 一个 PAWA 的使用案例

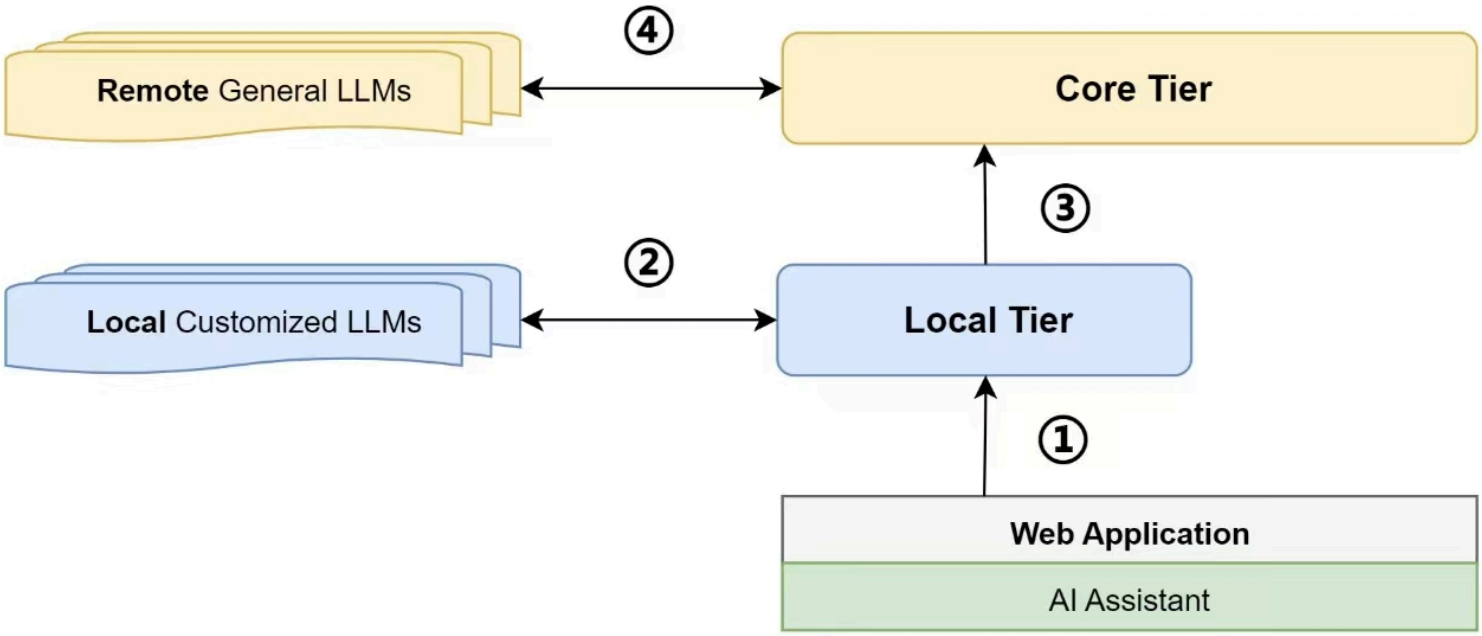
在一个预定酒店的应用里提供 AI 辅助预定功能，假设用户的需求是：“我要到纽约旅游，请帮我预定一家酒店”，以下是可能的场景：

- **直接访问远程模型服务：** 当本地设备缺乏对本地模型服务的支持，Web 应用只能通过核心层访问远程模型服务。服务将返回三种不同价位的酒店清单，分别是每晚约$800 - $1000 的高档酒店，每晚约$300 - $500 的中档酒店，以及每晚约$100 - $200 的经济型酒店。用户要在这份清单中选择适合自己的酒店。

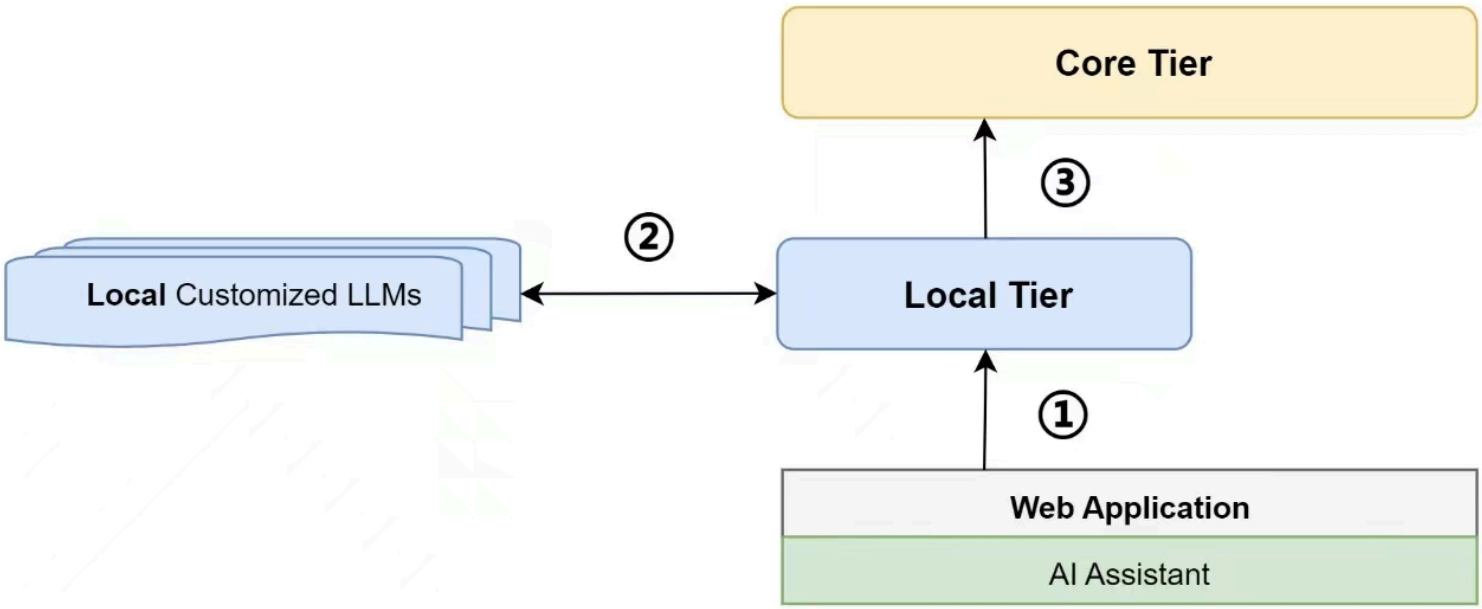


- **访问本地设备提供模型服务：** 当本地设备支持本地提供模型服务，Web 应用首先通过本地层访问本地模型服务，获取到用户的个人信息和偏好，然后根据用户偏好优化用户的需求，再通过核心层向远程模

型服务发送请求：“作为一名在校女大学生，我要到纽约独自旅游，请推荐一家适合我的酒店”。服务将返回安全性高、交通便利和每晚约$80 - $150 的酒店清单。



- **远程模型服务响应慢或不可用：** 假设本地设备支持本地模型服务，同时远程模型服务响应慢或不可用，Web 应用通过本地层访问本地模型服务。假设本地模型服务已同步远程模型的数据，拥有最近更新的酒店清单，根据用户画像匹配用户的需求之后，再交由核心层完成酒店预定的流程。



3. PAWA 的使用场景

PAWA 的使用场景适用于以下情况：

- **离线或网络连接不稳定：** 在本地设备处于离线状态，或者网络信号差的区域，PAWA 可以确保 Web 应用基本的 AI 功能可用。当远程模型服务无法访问时，本地模型服务可以作为快速且可靠的备用解决方案。

- **对隐私敏感的应用：** 某些 AI 应用或局部功能可能直接涉及用户的隐私和敏感数据，用户可能不愿将这些数据传输到远程模型服务。PAWA 可以将数据处理转移到本地模型服务，减轻用户隐私和敏感数据泄露的风险。此外，本地和远程模型之间可以进行多次交互，使得远程模型无法访问任何敏感数据。
- **缓解远程模型服务压力：** 当 AI 应用的用户呈爆发式增长时，远程模型服务可能会遇到性能瓶颈。PAWA 可以将部分计算任务转移到本地模型服务，减轻远程服务器的负载并提高响应速度。

4. PAWA 的局限性

PAWA 的使用存在以下局限性：

- **模型大小和性能：** 本地模型通常比远程模型更小，无法达到与远程模型相同的性能水平。
- **本地硬件限制：** 由于用户本地配备的硬件不同，导致 AI 处理速度不同，可能无法承载复杂巨大的模型。
- **维护成本：** 同时维护远程模型和本地模型会增加 Web 应用的开发和运维成本。
- **兼容性问题：** 不同平台和设备可能对本地模型的支持程度不同，从而引发兼容性问题。

5. PAWA 与 Hybrid AI 的关系

> Hybrid AI 总体目标是通过为 Web 开发者提供在服务器和客户端之间分配数据和计算资源的工具，从而最大化机器学习应用的用户体验 -- **Hybrid AI Exploration**

PAWA 与 Hybrid AI 的总体目标基本一致，面临的问题也基本相同，但 PAWA 侧重于针对不同预算 AI 功能渐进增强的策略：

- **设计理念：** PAWA 的设计理念是将本地模型服务作为一种渐进增强功能，即使本地设备不支持模型，Web 应用里的 AI 功能也能正常使用。
- **技术实现：** 当侦测到远程模型服务响应变慢甚至无法访问时，又或者识别出正在处理用户隐私数据的场景，Web 应用将尝试把任务转移给本地模型服务。
- **混合场景：** 由于本地模型可以使用用户的个性化信息来训练，因此 Web 应用可能混合使用远程模型和本地模型服务，以增强 AI 定制化的能力。

现今 AI PC 尚未普及，支持 NPU 的本地设备也比较少，可以预见在相当长的一段时间里，Web 应用需要应对用户的本地设备逐步支持本地模型的现状，PAWA 使用本地模型渐进性增强 Web 应用 AI 功能的策略，符合当前 AI 发展的背景和趋势。