精选快讯　软件快讯　2026年11月23日 05:01

新手指南：用Unsloth Desktop和Qwen在Windows PC上免费搭建本地AI，10-15分钟可局域网全家共用

要点

- 该指南介绍使用Unsloth Desktop在本地PC上搭建免费的个人AI（Claude和ChatGPT的本地替代品），支持在家中Windows环境运行并通过局域网供全家使用。
- 量化通过压缩权重、降低精度缩小模型体积，代价是牺牲部分准确性以换取能在GPU上运行且速度更快。
- 使用Inkling 2 Mini需先更新Unsloth Desktop，因其需要较新版本。

该指南介绍使用Unsloth Desktop在本地PC上搭建免费的个人AI（Claude和ChatGPT的本地替代品），支持在家中Windows环境运行并通过局域网供全家使用。硬件就绪后，整个搭建过程约需10-15分钟（不含下载时间），无需终端命令、Python环境或订阅。搭建分三步：安装Unsloth Desktop、下载模型、设置上下文大小；还需开启局域网访问。显存需求大致为：8-12GB显存可运行9B-12B的较小模型，32GB显存可舒适运行27B-35B模型并支持长对话。完整版Qwen模型约56GB，通常无法装入消费级显卡（除非使用统一内存的Mac）。

量化通过压缩权重、降低精度缩小模型体积，代价是牺牲部分准确性以换取能在GPU上运行且速度更快。指南作者在RTX 5090（32GB显存）上使用4-bit量化的Qwen3.8-27B模型，选择该档位以保留更大的上下文窗口。4-bit量化下，该模型加128K上下文可装入32GB显卡并有富余。Thinking Machines Lab于2026年11月17日发布Inkling 2 Mini：26.4B参数、Apache 2.0许可、支持视觉和100万token上下文。Unsloth推荐的Inkling 2 Mini量化文件UD-Q4_K_XL为17.6GB，可在32GB显卡上安装并留下约14GB给上下文。

使用Inkling 2 Mini需先更新Unsloth Desktop，因其需要较新版本。Inkling 2 Small为594B模型，Unsloth发布的最小文件为171.7GB，需要256GB Mac Studio或DGX Station等统一内存设备。Mac设备凭借统一内存可运行更大模型，但性能并非最佳；使用时需选择专为Apple Silicon构建的MLX版本。

报道来源：gpupartner

来源网站：https://www.inews.today/