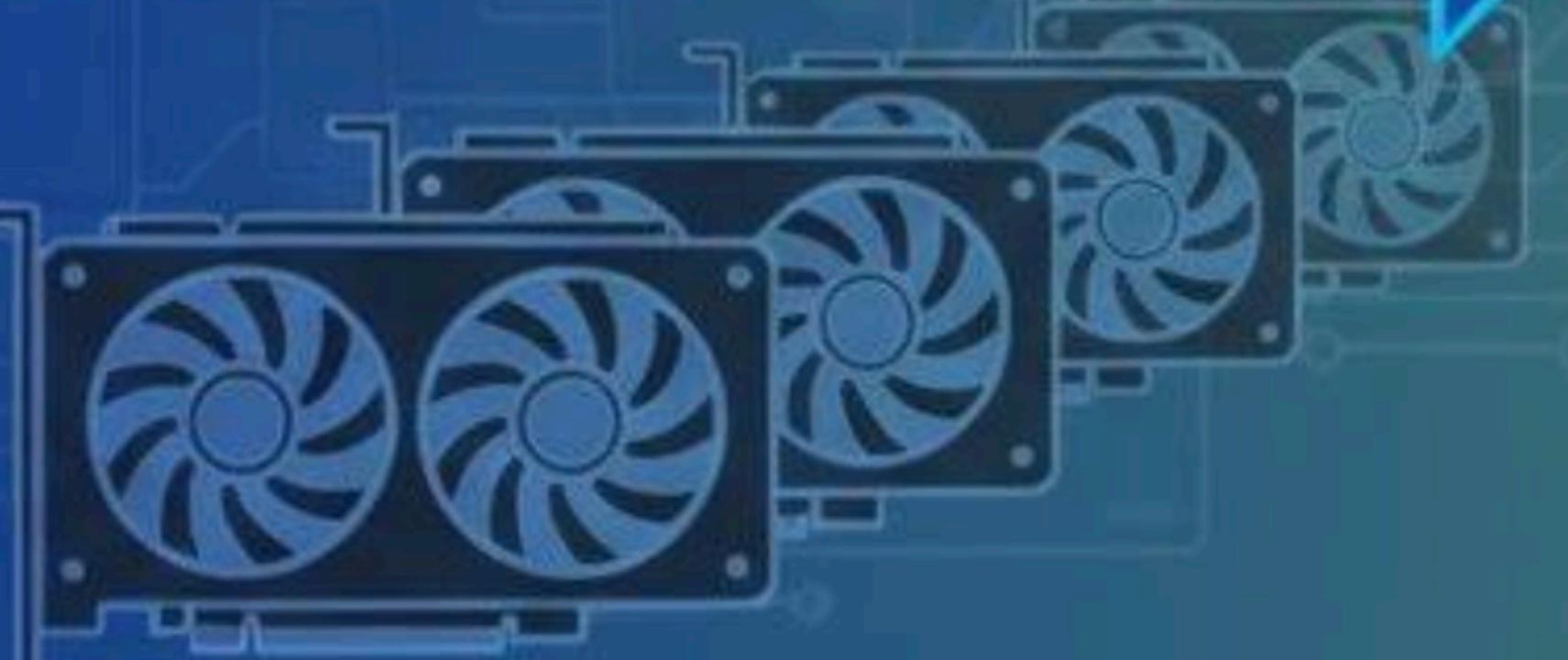




NVIDIA.[®]

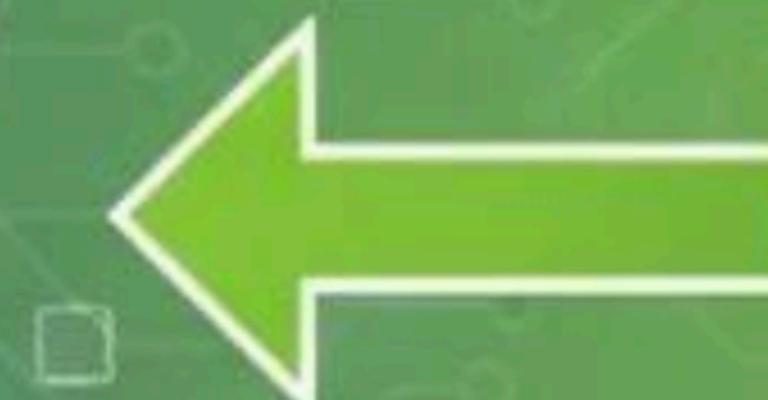
AMD



LM helper

Local LLMS

Antonín Ečer, DiS. 14.12.2025



HUAWEI



Lokální AI Asistent pro IT & Helpdesk

Soukromý • On-Prem • Na základě lmhelper (open-source)

- Rutinní incidenty zabírají hodiny – manuální hledání v dokumentech
- Cloud AI (ChatGPT/Copilot) = data odchází ven → bezpečnostní riziko
- Řešení: lmhelper + lokální LLM → vše zůstává u nás, odpovědi v sekundách



Start během týdne – 2× Mac mini HA setup

Hardware: 2× Apple Mac mini (M4/M2, 16–24 GB RAM) → tiché, nízká spotřeba

Software: lmhelper (git clone), lokální model (Qwen/Mistral 7–8B via MLX/LM Studio)

Kroky: Ingest KB → Spust' UI/API → Integruj do helpdesku

Náklady: ~45–55k CZK celkem (HA, redundancy)

Čas na PoC: 2-3 dny



Proč začít + Upgrade path

Výhody: Bezpečné, rychlé, nízké náklady → vyšší produktivita IT týmu

Když vyroste (10+ uživatelů, větší modely): Přejdi na NVIDIA GPU

Strategie: Start s Apple MLX → Scale na CUDA jen při potřebě

| Kritérium | Apple Mac mini | NVIDIA GPU |
|---------------------|----------------|------------|
| Náklady | Nízké | Vysoké |
| Složitost | Jednoduchá | Vyšší |
| IT RAG/Helpdesk | Ideální | Overkill |
| Scaling (10+ users) | Omezeně | Výborně |

„Lokální AI s lmhelper – data zůstávají u nás, odpovědi přichází v sekundách.

Začneme PoC tento týden?“