Lokální hostování AI modelů

Lokální provozování jazykových modelů (On-device AI) umožňuje využívat umělou inteligenci bez nutnosti odesílat data na servery třetích stran (OpenAI, Google). To zajišťuje **soukromí**, **nezávislost na internetu** a **nulové provozní náklady**.

Požadavky na hardware

Klíčovým faktorem pro výkon není procesor (CPU), ale **grafická paměť (VRAM)** a **propustnost RAM**.

- **Operační paměť (RAM):** Minimum je 8 GB pro nejmenší modely (3B), doporučeno 16 GB+.
- **Grafická karta (GPU):** Čipy NVIDIA (architektura CUDA) jsou nejlépe podporovány.
- **Apple Silicon (M1/M2/M3):** Velmi efektivní díky sdílené paměti (Unified Memory).
- **Disk:** SSD s dostatkem místa (modely mají 2 GB až 50 GB+).

Doporučené nástroje (Software)

Pro začátečníky i pokročilé existují tři hlavní cesty, jak model zprovoznit:

1. Ollama (Nejjednodušší cesta)

Ollama je terminálový nástroj pro macOS, Linux a Windows, který spravuje stahování i běh modelů.

- **Příkaz:** `ollama run llama3`
- **Výhoda:** Extrémně jednoduché, funguje jako server na pozadí.
- **Webové rozhraní:** Lze propojit s **Open WebUI** pro zážitek podobný ChatGPT.

2. LM Studio (Grafické rozhraní)

Aplikace s plnohodnotným GUI, která umožňuje vyhledávat modely přímo z portálu **Hugging Face**.

- **Vhodné pro:** Uživatelé, kteří nechtějí používat příkazovou řádku.
- **Funkce:** Snadné nastavení parametrů a sledování vytížení hardwaru.

3. LocalAI / vLLM (Pro vývojáře)

Nástroje určené pro nasazení v rámci lokální infrastruktury přes Docker.

- **Výhoda:** Poskytují API kompatibilní s OpenAI (lze v kódu jen přepsat URL adresu).

Formáty modelů a kvantizace

Většina lokálních modelů využívá formát **GGUF**. Protože jsou modely v plné přesnosti příliš velké, používá se tzv. **kvantizace** (snížení bitové přesnosti).

Kvantizace	Vliv na kvalitu	Využití RAM
Q8_0 (8-bit)	Téměř nerozeznatelný od originálu	Vysoké
Q4_K_M (4-bit)	Zlatá střední cesta (doporučeno)	Střední
Q2_K (2-bit)	Výrazná ztráta logiky	Minimální

Postup nasazení (Rychlý start)

```
1. **Stáhnout Ollama:** Z oficiálních stránek
[[https://ollama.com|ollama.com]].
2. **Výběr modelu:** Pro začátek doporučujeme ''phi3'' (malý a rychlý) nebo
''llama3'' (všestranný).
3. **Spuštění:** V terminálu zadejte:
```

```
ollama run phi3
```

```
4. **Integrace:** Propojte lokální instanci s vaším editorem kódu (např.
pomocí pluginu **Continue** ve VS Code).
```

> **Tip:** Pokud máte málo VRAM, hledejte modely s označením „Instruct“, které jsou vyladěny pro plnění úkolů a chatování.

— **Související dokumentace:**

- Small Language Models
- Nastavení GPU akcelerace
- Práce s portálem Hugging Face

— **Správce IT sekce:** @AI_Admin