

Lokální hostování AI modelů (Self-hosting)

Lokální hostování znamená, že model běží na vašem vlastním procesoru (CPU) nebo grafické kartě (GPU). Veškerá data zůstávají u vás a model funguje i bez připojení k internetu.

Hardwarové požadavky

Nejdůležitějším parametrem je **VRAM** (video paměť na grafické kartě) nebo **Unified Memory** (u počítačů Apple).

- **Minimální (SLM modely 1B-4B):** 8 GB RAM / 4 GB VRAM.
- **Optimální (Střední modely 7B-14B):** 16 GB - 32 GB RAM / 8 GB - 12 GB VRAM.
- **Profesionální (Velké modely 30B+):** 64 GB+ RAM / 24 GB+ VRAM (např. NVIDIA RTX 3090/4090).

Hlavní nástroje pro rok 2026

1. Ollama (Doporučeno pro začátečníky)

Ollama je nejjednodušší cesta. Funguje jako služba na pozadí, která automaticky spravuje paměť.

- **Platformy:** Windows, macOS, Linux.
- **Příkaz:** `ollama run llama3.1`
- **Výhoda:** Obrovská knihovna předpřipravených modelů.

2. LM Studio (GUI aplikace)

Grafické rozhraní, které vám umožní stahovat konkrétní verze modelů z Hugging Face a testovat je v okně chatu.

- **Klíčová funkce:** Ukazuje v reálném čase, kolik procent modelu se vejde do vaší grafické paměti.

3. Open WebUI (Firemní standard)

Pokud chcete rozhraní, které vypadá a funguje přesně jako ChatGPT, ale běží u vás.

- **Nasazení:** Ideální přes Docker.
- **Funkce:** Podpora více uživatelů, historie chatů, nahrávání dokumentů (RAG).

Formáty a Kvantizace (GGUF)

Modely se málokdy stahují v plné velikosti (FP16). Používá se formát **GGUF**, který umožňuje snížit bitovou přesnost (kvantizaci) modelu:

Kvantizace	Kvalita	Nároky	Doporučení
Q8_0	99% originálu	Velmi vysoké	Pro kritické úlohy
Q4_K_M	95% originálu	Nízké	Zlatý standard (nejlepší poměr)
IQ2_S	70% originálu	Minimální	Pro slabé telefony/IoT

Rychlý návod: První spuštění (Ollama)

1. Stáhněte a nainstalujte **Ollama** z [[https://ollama.com|oficiálního webu]].

2. Otevřete terminál (PowerShell / Bash).

3. Spusťte svůj první model:

```
# Pro rychlý test (velmi malý model)
ollama run phi4:tiny
```

```
# Pro kvalitní chat (střední model)
ollama run llama3.1:8b
```

4. Model se automaticky stáhne a můžete začít psát.

TIP: Pokud chcete model využívat v jiných aplikacích (např. v MS Word nebo VS Code), Ollama automaticky běží na adrese <http://localhost:11434>.

— **Související dokumentace:**

- [Průvodce Small Language Models](#)
- [Aktualizace ovladačů pro AI](#)
- [Nastavení vlastní báze znalostí \(RAG\)](#)

— **Autor:** @IT_Admin **Verze dokumentu:** 1.2 (2026)

From:

<https://www.serviceit.cz/> - IT ENCYKLOPEDIE

Permanent link:

https://www.serviceit.cz/doku.php?id=it:hostovani_modelu

Last update:

2026/01/04 16:41

