

Umělá inteligence v e-INFRA CZ

Není to zvláštní, jak se to někdy semele?

Umělá inteligence

- Především velké jazykové modely (LLM)
 - “Velké” → nelze snadno provozovat “doma”
 - Provoz je **nákladný**: vyžaduje velké výpočetní zdroje
 - Jednotlivci mají obvykle **malé využití** oproti nákladům
 - Nejznámější podoba: **chat** (Chat GPT)
- Dnes rozšířené o **multimodalitu**
 - Generují obrázky, videa a zvuk (např. diffusion modely)
 - Rozpoznávají obsah v obrazech, videích a zvuku

Jazykové modely

A co s nimi?

- **Chat**

- Práce s texty: překlady, lepší formulace, analýza, shrnutí
- Generování programového kódu, vysvětlování chyb (ne vždy úspěšné)
- (Přes multimodalitu) také tvorba obrázků

- **Integrace do služeb**

- Chatbot pro dokumentaci či znalostní bázi
- Programovací asistenti v IDE (např. Jupyter Notebook, VS Code)

- **Nevhodné použití**

- Testování „znalostí“ modelu → rychlé zklamání

Jazykové modely

Jaký model?

- Různé typy modelů
 - Liší se znalostmi, zaměřením a velikostí (počet parametrů v miliardách)
 - **Coder** modely – zaměřené na programování
 - **Reasoning** modely – simulují „vnitřní dialog“
 - **Instruct** modely – optimalizované pro plnění příkazů
 - Obecně platí: větší model → většinou lepší výkon
- Použití nástrojů (**tools**)
 - Modely mohou integrovat externí nástroje, např. pro prohledávání internetu

Jazykové modely v e-INFRA CZ

- **Skladba modelů se mění**
 - Velmi rychlý vývoj → nové modely prakticky denně
 - Novější modely zpravidla nabízejí vyšší výkon
- **Hlavní aktuální modely**
 - **GPT-OSS-120B** – Reasoning model, ekvivalent známých mini modelů OpenAI
 - **DeepSeek R1 0528** (685B) – Reasoning model pro složité úkoly
 - Nejlepší pro český jazyk
 - Nemá přístup na internet → bezpečné použití
 - **Qwen3 Coder 480B Instruct** – Instruct model pro programování
- **Další dostupné modely**
 - Mistral Small 3.2, Aya, Phi-4
 - **Embedding modely:** Qwen3-Embedding-4B, Multilingual E5 Large, Nomic



AI služby e-INFRA CZ

- chat.ai.e-infra.cz
 - Obdoba ChatGPT, s dostupností různých modelů
 - **Nástroje**: vyhledávání na Internetu a GitHubu, stahování odkazovaných stránek
 - **Generování obrázků**
 - Podpora **Knowledge (RAG)** – omezená
 - **API integrace** – možnost napojení do vlastních aplikací/služeb

Tip dne: *Improve the text, return raw markdown, replace any internal triple backticks with \$\$\$, and wrap the output in triple backticks.*

AI služby e-INFRA CZ

- deepsite.ai.e-infra.cz
 - Generuje webové stránky a aplikace na základě textového popisu
 - Vytváří **HTML + CSS + JavaScript**
 - Ideální pro **designové návrhy** webů i aplikací

AI služby e-INFRA CZ

- **Dokumentační chatbot**

- Implementován ve všech dokumentačních částech e-INFRA CZ
- Odpovídá na otázky z dokumentace
 - Např.: „*How to run Matlab?*“
- Některé části mají **problem solver**
 - Např.: „*Create Deployment for Kubernetes with Nginx image and a proper Service*“

AI v Jupyter Noteboocích

- hub.cloud.e-infra.cz
 - Nabízí **Jupyter** instanci s Notebook Intelligence
 - Integrace **chatu** přímo v **Jupyter Lab**
 - Napojeno na naše modely
 - **Možnosti využití:**
 - Konverzace s modelem
 - Opravy existujícího kódu
 - Generování nového kódu

Co dál?

- Chat je podporovaná služba
 - Modely se budou **měnit v čase**
 - Některé modely mohou být **v noci nedostupné** kvůli údržbě či testování
 - Možnost **požádat** o zpřístupnění nového modelu
 - Některé modely jsou **on-demand** – první odpověď může trvat i pár minut
- **Očekávané novinky**
 - Další služby s podporou AI
 - Rozvoj **AI agentů**

Děkuji za pozornost

k8s@cerit-sc.cz
https://docs.cerit-sc.cz
https://blog.e-infra.cz