Small Language Models (SLM)

Small Language Models (SLM) představují novou generaci modelů umělé inteligence, které se zaměřují na efektivitu, rychlost a specializaci. Na rozdíl od svých „velkých bratrů“ (LLM) mají výrazně méně parametrů, ale díky kvalitním datům dosahují překvapivých výsledků.

Hlavní charakteristika

Zatímco modely jako GPT-4 pracují se stovkami miliard až biliony parametrů, SLM se obvykle pohybují v rozmezí **1 až 10 miliard parametrů**.

Klíčové výhody SLM

- **Lokální běh (On-device AI):** Model lze spustit na běžném notebooku nebo mobilním telefonu bez internetu.
- **Ochrana soukromí:** Data neopouštějí zařízení, což je ideální pro bankovnictví nebo zdravotnictví.
- **Nízké náklady:** Provoz SLM vyžaduje zlomek elektrické energie a výpočetního výkonu oproti velkým modelům.
- **Rychlost:** Mají velmi nízkou latenci (okamžité generování textu).

Srovnání parametrů

Parametr	LLM (např. GPT-4)	SLM (např. Phi-3)
Velikost	Stovky GB / Terabajty	Jednotky GB
Hardware	GPU clustery (H100)	Běžné CPU / Mobilní čipy
Využití	Všeobecné znalosti, komplexní úvahy	Specializované úlohy, asistenti
Cena za dotaz	Vyšší (API poplatky)	Téměř nulová (vlastní HW)

Příklady moderních SLM

- **Microsoft Phi-3:** Jeden z nejvýkonnějších modelů ve své třídě (3.8B parametrů).
- **Google Gemma:** Otevřené modely postavené na stejné technologii jako Gemini.
- **Meta Llama 3 (8B):** Velmi populární model pro lokální nasazení.
- **Mistral 7B:** Francouzský model, který odstartoval trend efektivních menších modelů.

Praktické využití

> **Příklad:** Firma může nasadit SLM pro analýzu interních smluv. Model běží na firemním serveru, nikdo zvenčí k datům nemá přístup a odpovědi jsou generovány okamžitě v rámci interního systému.

— **Viz také:**

- Rozcestník AI
- Jak hostovat modely lokálně