LLM (Large Language Models)

Large Language Models (LLM), neboli velké jazykové modely, jsou typem umělé inteligence trénované na obrovském množství textových dat. Jsou schopny generovat text, překládat jazyky, psát kód a odpovídat na dotazy způsobem, který připomíná lidskou komunikaci.

Základem moderních LLM (jako jsou GPT-4, Claude nebo Llama) je architektura neurálních sítí zvaná **Transformer**.

1. Architektura Transformer Představená společností Google v roce 2017 (v článku „Attention Is All You Need“), tato architektura nahradila starší sekvenční modely (RNN, LSTM).

Klíčovým prvkem je **Self-Attention mechanismus** (sebepozornost), který umožňuje modelu:

- Vyhodnocovat váhu (důležitost) různých slov ve větě bez ohledu na jejich vzdálenost.
- Zpracovávat data paralelně (nikoliv slovo po slově), což umožnilo trénink na masivních datasetech.

2. Jak funguje generování (Tokenizace) Modely nepracují s celými slovy, ale s tzv. **tokeny** (části slov, znaky nebo celá slova).

- **Tokenizace:** Text „DokuWiki je super“ se rozdělí na [Doku, Wiki, je, super].
- **Embeddings:** Každý token je převeden do vícerozměrného vektoru čísel, který reprezentuje jeho sémantický význam.
- **Predikce:** Model vypočítává pravděpodobnost, jaký token by měl následovat po zadaném vstupu (Context Window).

3. Fáze vývoje modelu Proces vytvoření použitelného AI asistenta probíhá v několika krocích:

Fáze	Název	Popis
1.	**Pre-training**	Učení se surovým datům z internetu (predikce dalšího slova).
2.	**Fine-tuning**	Ladění na specifických sadách otázek a odpovědí.
3.	**RLHF**	*Reinforcement Learning from Human Feedback* – ladění podle lidských preferencí (bezpečnost, užitečnost).

4. Klíčové parametry a pojmy

- **Parametry:** Váhy v neurální síti, které se model učí (např. Llama-3-70B má 70 miliard parametrů).
- **Context Window:** Maximální počet tokenů, které model „udrží v paměti“ při jedné konverzaci.
- **Hallucination (Halucinace):** Jev, kdy model sebejistě generuje fakticky nesprávné informace.
- **Inference:** Proces, kdy již vytrénovaný model odpovídá na dotaz uživatele.

5. RAG (Retrieval-Augmented Generation) Protože LLM mají znalosti omezené datem ukončení tréninku (knowledge cutoff), používá se v podnicích technika **RAG**.

RAG umožňuje modelu nahlížet do externích databází (např. vaší DokuWiki) a odpovídat na základě aktuálních interních dokumentů, aniž by se musel znovu trénovat.

6. Příklady implementace (Open-Source) Pro lokální běh LLM na vlastním hardwaru se často používají nástroje:

- **Ollama:** Jednoduché rozhraní pro spouštění modelů.
- **vLLM:** Vysoce výkonný engine pro obsluhu modelů.
- **Hugging Face:** „GitHub“ pro AI modely a datasety.

> **Tip pro administrátory:** Pro provoz 7B (7 miliard) parametrů modelu v plné přesnosti je potřeba cca 14-28 GB VRAM, při použití kvantizace (komprese) postačí i 8 GB.

Zpět na AI rozcestník