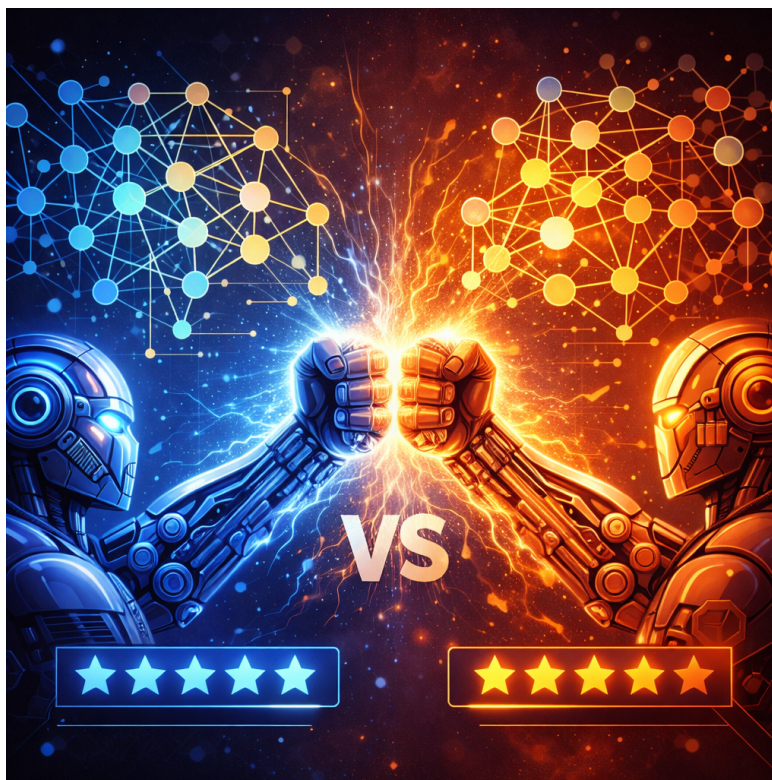




Moduly umělé inteligence



Západočeská Univerzita V Plzni
Katedra Kybernetiky
Semestrální práce - HKUI

Ondřej Knížek
1. ročník
31. ledna 2026

1 Úvod

V této semestrální práci srovnám nejvýznamnější komerční LLMs (Large Language Models) od různých společností, konkrétně od OpenAI (GPT-3.5, GPT-4, GPT-5), Googlu (Gemini 1.0 Pro, Gemini 2.0 Pro, Gemini 3 Pro) a od společnosti DeepSeek (DeepSeek-v3). V potaz vezmu především technické specifikace.

2 Specifikace

2.1 Počet parametrů

Parametry jsou jinými slovy proměnné neuronové sítě, které se během trénování upravují. Díky nim se modely dokáží učit.

$$y = f\left(\sum_{i=1}^n w_i x_i + b\right)$$

Parametr umělého neuronu má dvě části:

1. w - váhy jednotlivých vstupů (weights)
2. b - práh (bias).

Zbytek neuronu je tvořen vstupy x a aktivační funkcí f .

Důležitou informací je, že celkový počet parametrů nehraje hlavní roli v “chytrosti” modelu. Ty totiž využívají Mixture of Experts (MoE), což znamená, že model může obsahovat několik miliard parametrů, ale aktivuje se pouze zlomek specializovaných na konkrétní otázku.

Dále hraje zásadní roli počet trénovacích tokenů, tzn. množství dat, na kterých byl model vytrénován. Pokud má model méně parametrů, ale více kvalitních dat, může být chytřejší než větší model.

Název	Počet parametrů	Počet aktivních parametrů
GPT-3.5	175 miliard	Všechny
GPT-4	1,7 bilionu	280 miliard
GPT-5	2 až 5 bilionů	300 až 500 miliard
Gemini 1.0 Pro	100 až 170 miliard	Všechny
Gemini 2.0 Pro	1,2 až 2 biliony	300 až 450 miliard
Gemini 3 Pro	2 až 2,5 bilionu	300 až 500 miliard
DeepSeek-v3	671 miliard	37 miliard

Údaje v tabulce jsou jen přibližné - odhadnuté experty. Pouze u GPT-3.5 a DeepSeek-v3 byla tato informace zveřejněná. Společnosti tají specifikace jednotlivých modelů z důvodů velkého konkurenčního boje.

V raných stádiích vývoje umělé inteligence se nevyžívala modularita (MoE), proto modely jako GPT-3.5 a Gemini 1.0 Pro využívali všechny neurony najednou.

Gemini 3 Pro byl první model, který využíval dynamické škálování. To znamená, že při jednoduchých úlohách aktivuje méně neuronů, zatímco u složitějšího uvažování (např. režim Deep Think) zapojuje maximum své kapacity.

2.2 Velikost kontextového okna

Tato vlastnost určuje, kolik textu si model v rámci jedné konverzace zapamatuje. Pokud dosáhneme limitu, model začne postupně zapomínat komunikaci od začátku konverzace.

Délka konverzace je měřená tzv. tokeny. To jsou nejmenší jednotky textu, které model zpracovává. Můžou představovat část slova, celé slovo, interpunkci, speciální znak nebo mezeru. Text o 1000 slovech obsahuje 1500-2000 tokenů. Nevýhoda češtiny oproti např. angličtině je taková, že používá cca dvojnásobně více tokenů. Diakritika bývá totiž často interpretována jako samostatný token.

Slovní zásoba přirozeného jazyka je extrémně rozsáhlá - miliony slov. Tokenizace umožňuje reprezentovat jazyk pomocí menší množiny základních jednotek - 30-50 tisíc. Každý token má vektorovou reprezentaci, což zefektivňuje výpočty.

Název	Počet tokenů	Přepočteno na strany textu
GPT-3.5	4096	15
GPT-4	8192	30
GPT-5	512 000	1 800
Gemini 1.0 Pro	32 768	110
Gemini 2.0 Pro	2 000 000	7 000
Gemini 3 Pro	5 000 000	17 500
DeepSeek-v3	128 000	450

Hodnoty jsou opět pouze orientační.

2.3 Multimodalita

Multimodalita je schopnost modelu pracovat v reálném čase nejen s textem, ale i s obrázky, zvukem a videem.

Název	Multimodalita
GPT-3.5	pouze text
GPT-4	text, obraz
GPT-5	text, obraz, audio
Gemini 1.0 Pro	text, obraz
Gemini 2.0 Pro	text, obraz, video
Gemini 3 Pro	text, obraz, video
DeepSeek-v3	text, obraz

GPT-4 byl první model, který dokázal pracovat s obrázkama. Dokázal je popsat nebo z nich vypsát text.

3 Závěr

Ze srovnání technickým specifikací vyplývá, že ačkoliv se počet parametrů u nejvýkonnějších modelů jako GPT-5 či Gemini 3 pohybuje v řádech bilionů, nejdůležitějším zůstává modularita (MoE) a efektivita aktivních parametrů.

Také se ukázalo, že modely od společnosti Google vynikají v extrémně velkém kontextovém okně, které umožňuje zpracování rozsáhlých dokumentů a videí. Na druhou stranu modely jako DeepSeek-v3 ukazují, že vysokého výkonu lze dosáhnout i s výrazně nižším počtem aktivních parametrů při zachování konkurenceschopné chytrosti.

Lze očekávat, že dalším milníkem ve vývoji nebude pouze navyšování výkonu, ale především hlubší využití multimodality a schopnost modelů provádět komplexní logické operace v reálném čase. Umělá inteligence se tak stává čím dál více univerzálním nástrojem, který překračuje hranice pouhého generování textu.

Zdroje

Online. Dostupné z: <https://www.leadmachine.cz/proc-vas-ma-nebo-nema-zajimat-kontextove-okno-u-llm/>. [cit. 2026-01-31].

Online. Dostupné z: <https://www.marigold.cz/ai/tokeny-versus-slova/>. [cit. 2026-01-31].

Online. Dostupné z: <https://www.aidatabase.cz/slovník/paramet/>. [cit. 2026-01-31].

Online. Dostupné z: <https://platform.openai.com/docs/models>. [cit. 2026-01-31].

Online. Dostupné z: <https://openai.com/cs-CZ/news/>. [cit. 2026-01-31].

Online. Dostupné z: <https://arxiv.org/abs/2412.19437>. [cit. 2026-01-31].

Online. Dostupné z: <https://artificialanalysis.ai/models>. [cit. 2026-01-31].

Online. Dostupné z: <https://blog.google/innovation-and-ai/technology/ai/>. [cit. 2026-01-31].

obrázek: Online. In: . Dostupné z: <https://chatgpt.com/g/g-p-6962ca2015ec8191afd484ca4cf7d881-mat-analyza/c/697e09cc-a248-8396-9190-9b6dce727dc3>. [cit. 2026-01-31].