Velké jazykové modely (LLM)

Úvod do ICT

Jiří Poláček

Text z pohledu LLM

Token:

- Fragment slova, často písmeno nebo slabika, ale patří sem i interpunkce či ustálená slovní spojení (např. „ty jsi“ apod.)
- Má tisíce atributů či vah, které pomáhají porozumět významu
  - citové zabarvení, odbornost, modernost apod.
- LLM modely se odlišují:
  - „slovníkem“ tokenů
  - velikostí tohoto slovníku (desítky až stovky tisíc tokenů)
  - kapacitou odpovědi, tj. kolik tokenů je schopen přečíst a vypsat

https://platform.openai.com/tokenizer

Jak LLM „píše“ své odpovědi

- LLM je v principu generátor náhodných slov
- Na základě vstupu a toho, co již bylo vypsáno, je pro každý token spočítána pravděpodobnost, s jakou bude vybrán a vypsán
  - Např. pro text „do těsta přidej“ to bude „vejce“ (20 %), „mléko“ (15 %), „olej“ (12 %), „prášek“ (5%) „pepř“ (1 %), „bagr“ (0 %) …
  - Přesněji řečeno, v příkladu výše je naznačena pravděpodobnost, s jakou bude vybrána sekvence tokenů, která dohromady dává dané slovo
- Náhodně vybrán může být také speciální token sdělující STOP
  - Jednou musí bát vybrán – čím delší odpověď, tím pravděpodobněji konec
  - Lze specifikovat délku odpovědi nebo nastavit „stop sekvence“

Rozmanitost odpovědí

- Teplota – ovlivňuje výpočet distribuce pravděpodobnosti
  - Nejnižší (nulová) teplota = vždy bude vybrán nejpravděpodobnější token
  - Vyšší teplota = méně pravděpodobné tokeny mají větší šanci na výběr
- TopP – omezuje počet tokenů, ze kterých se vybírá
  - Např. hodnota 0,5 znamená, že se bude vybírat jen z poloviny tokenů (těch více pravděpodobných)
  - Podobně TopK vybírá jenom z *k* nejvíce pravděpodobných tokenů
- Penalizace – zamezení opakování stejných slov
  - Snižuje pravděpodobnost modelu opakovaně psát stejné odpovědi

Asistenti

- Podobně jako se do komunikace s lidmi promítá jejich osobnost, může i AI asistent, se kterým si píšeme, být různě naladěn
- Výchozí asistent nejpoužívanějších modelů:
  - Zdvořilý, úslužný, přející, pomáhající, omlouvající se za chyby, na všechny naše prompty reaguje se samozřejmostí
- Možné redefinice asistentů:
  - Tón komunikace (věcný, sarkastický, …)
  - Kompetence (učitel matematiky, šachový velmistr, …)
  - Funkce (možnost práce se soubory, vyhledávání na webu, …)

https://character.ai/

Další zajímavé odkazy

- Hřiště pro vývojáře
  - https://platform.openai.com/playground
  - https://studio.ai21.com/
- Rozcestníky AI nástrojů
  - https://ejaj.cz/
  - https://allthingsai.com/
- Perličky
  - https://huggingface.co/ (AI komunita)
  - https://lmarena.ai/?arena (hodnotí se, který LLM dal lepší odpověď)