Token

Token je atomická jednotka dat, se kterou pracují modely jako GPT-4, Llama nebo Claude. Může se jednat o celé slovo, část slova, jednotlivé písmeno nebo dokonce interpunkční znaménko. Tokenizace je proces, při kterém se lidský text převádí na číselný formát, kterému rozumí matematické operace uvnitř neuronové sítě.

1. Jak funguje tokenizace?

Modely používají pokročilé algoritmy (např. **Byte Pair Encoding - BPE**), aby text efektivně rozdělily:

- **Běžná slova:** Často tvoří jeden token (např. „apple“).
- **Dlouhá nebo neobvyklá slova:** Jsou rozdělena na více částí (např. „tokenizace“ $\rightarrow$ „token“ + „izace“).
- **Mezery a znaky:** Mezera před slovem je obvykle součástí tokenu.

Pravidlo palce: V angličtině platí hrubý odhad, že **1000 tokenů odpovídá přibližně 750 slovům**. V češtině je kvůli složitější gramatice a diakritice poměr tokenů ke slovům o něco vyšší (slova se častěji dělí na více částí).

2. Proč jsou tokeny důležité?

A. Kontextové okno (Context Window)

Každý model má pevně danou kapacitu paměti, tzv. kontextové okno, definované v tokenech.

- Pokud má model limit 128k tokenů, znamená to, že „vidí“ zhruba 300 stránek textu najednou.
- Jakmile je limit překročen, nejstarší tokeny z paměti vypadávají.

B. Cena a rychlost

Většina poskytovatelů AI služeb (OpenAI, Anthropic) účtuje poplatky za používání API na základě počtu zpracovaných tokenů (vstupních i výstupních). Zároveň počet tokenů určuje rychlost generování – čím více tokenů musí model vytvořit, tím déle odpověď trvá.

3. Tokeny a různé jazyky

Efektivita tokenizace se liší podle jazyka:

- **Angličtina:** Je nejefektivnější, většina slov = 1 token.
- **Čeština/Slovenština:** Méně efektivní, diakritika (háčky, čárky) může někdy způsobit, že jedno písmeno spotřebuje více tokenů.

- **Programovací kódy:** Jsou velmi efektivně tokenizovány, protože obsahují mnoho opakujících se klíčových slov.

4. Speciální tokeny

Modely používají i skryté tokeny pro řízení konverzace:

- **<|endoftext|>:** Označuje konec dokumentu nebo odpovědi.
- **<|system|>:** Označuje instrukce pro systémové nastavení modelu.
- **<|user|>:** Označuje začátek vstupu od uživatele.

5. Vizuální představa: Embeddings

Po rozdělení na tokeny model přiřadí každému tokenu unikátní číslo a následně jej převede na **Embedding** – vektor čísel v mnohorozměrném prostoru. V tomto prostoru mají sémanticky podobné tokeny (např. „král“ a „panovník“) souřadnice blízko sebe.

> **Zajímavost:** Zkuste si v ChatGPT nebo jiném modelu nechat napsat slovo pozpátku. Modely s tím mají často problém, protože slovo nevidí jako písmena, ale jako celistvé tokeny, a „neví“, z jakých písmen se token přesně skládá bez dodatečného uvažování.

Zpět na AI rozcestník