Zadání pro KIV/TSP1

Název zadání: Benchmark tokenové efektivity a syntaktické hustoty programovacích jazyků pro LLM

Krátký popis zadání:

Cílem projektu je navrhnout, implementovat a vyhodnotit automatizovaný benchmark, který systematicky změřit tokenovou efektivitu různých programovacích jazyků a paradigmat napříč moderními AI modely. Projekt prozkoumá poměr informační hustoty vůči počtu tokenů a posoudí dopad moderních syntaktických konstrukcí či různých tokenizérů na finanční náklady, latenci a využití kontextového okna LLM.

Podrobnosti zadání:

S rozvojem integrace velkých jazykových modelů (LLM) do vývojářských nástrojů představuje spotřeba tokenů klíčovou metriku ovlivňující nejen finanční náklady na API volání, ale také využití kontextového okna a latenci odezvy. Tým studentů bude řešit zejména:

- Výběr reprezentativní sady testovaných jazyků a paradigmat s různou úrovní abstrakce (např. Java/C++ vs. Python/Kotlin/Rust/Haskell) a srovnání se specializovanými DSL (např. Vanilla Python vs. Pandas/Polars vs. SQL).
- Sestavení datasetu úloh různé obtížnosti (Easy, Medium, Hard) vycházejícího z benchmarků jako HumanEval, MBPP, Rosetta Code či LeetCode.
- Přípravu více variant implementací pro každou úlohu (Baseline vs. High-Level / Compact s využitím moderních konstrukcí a knihoven).
- Měření a analýzu na různých modelech a tokenizérech (OpenAI, Anthropic Claude, Google Gemini, Meta Llama, Mistral, DeepSeek a jejich tokenizérech typu tiktoken či BPE).
- Vyhodnocení vlivu syntaxe na tokenovou stopu, přehlednost a chování odlišných tokenizérů.

Projekt studentům umožní získat zkušenosti s:

- návrhem a realizací automatizovaných testovacích a benchmarkingových rozhraní,
- hlubším porozuměním architektuře LLM, konkrétně fungování tokenizérů a zpracování kontextu,
- srovnávací analýzou syntaxe a paradigmat napříč širokou škálou programovacích jazyků,
- zpracováním, statistickou analýzou a vizualizací naměřených dat.

Časová náročnost [člověko-hodin]: 1500

Doporučená velikost týmu: 6 – 7

Požadovaný profil týmu:

Projekt vyžaduje tým se zájmem o velké jazykové modely (LLM), teorii programovacích jazyků a datovou analýzu. Vhodné jsou následující znalosti a dovednosti:

- znalost více programovacích jazyků a jejich paradigmat (procedurální, objektové, funkcionální),

- základní orientace v problematice LLM (tokenizace, kontextové okno, API rozhraní modelů),
- zkušenost s vývojem v Pythonu pro automatizaci, sběr a zpracování dat,
- schopnost návrhu benchmarkingových metodik a statistického vyhodnocení výsledků,
- práce s verzovacími systémy a automatizovaným spouštěním testů (CI/CD, skriptování),
- analytické myšlení a schopnost interpretace naměřených dat.