

УДК 004.8

Тягунова М.Ю.¹, Захарченко К.Р.²

¹ канд. техн. наук, доц. НУ «Запорізька Політехніка»

² студ. гр. КНТ-510 НУ «Запорізька Політехніка»

ВЕЛИКІ МОВНІ МОДЕЛІ ДЛЯ ДОМАШНІХ СИСТЕМ

У минулі роки сфера LLM (Large Language Model – велика мовна модель) зазнала стрімкого розвитку. Великі компанії розробили та розгорнули вебзастосунки для доступу до їх моделей. Розгортання власної (локальної) LLM надає ряд переваг над використанням віддалених рішень:

- контроль над даними – при роботі з конфіденційними або чутливими даними забезпечується, що вони не потраплять до третіх сторін;

- незалежність від хмарних сервісів – потрібні сервіси можуть бути недоступні через різні причини: несправність на сервері чи регіональні обмеження;

- економія ресурсів – при довготривалому використанні може мати економічну доцільність використовувати локальну модель задля відсутності потреби у сплаті за послуги.

Метою даної роботи стало виявлення параметрів для якісного дослідження LLM. Дослідження у сфері LLM в загальному зосереджувались на якості генерації тексту, зокрема шляхом збільшення об'ємів даних, та їх

можливостях. Одним з цікавих проєктів по темі є платформа «Ollama», яка має велику бібліотеку навчених LLM і надає безкоштовний та легкий доступ до них [1].

З аналізу бібліотеки платформи «Ollama» було виявлено, що багато моделей мають по декілька варіантів, з різною кількістю параметрів, що надає можливість запустити легші версії моделей навіть на не дуже потужних користувацьких комп'ютерах. Порівнювати різні моделі можна за наступними показниками:

- можливість відповідати на різні типи запитань;
- які мови підтримує модель;
- можливість вирішення завдань з математики та програмування [2].

Швидкодія моделей пропонується вимірювати за середнім значенням кількості токенів, які модель генерує за секунду, та об'ємом необхідних ресурсів. Серед найбільш впливових параметрів є обсяг оперативної пам'яті та навантаження на центральний і графічний процесори, необхідних для роботи моделі. Швидкодія моделей може залежати від операційної системи (ОС), тому під час дослідження цього питання, було прийнято рішення провести паралельне порівняння швидкодії на таких ОС як Windows та Linux. Результати експериментів дадуть змогу порівняти окремі моделі, сформувані рекомендації щодо оптимізації налаштувань та визначити більш ефективну модель з проаналізованих.

Таким чином, у роботі запропоновано дослідити різні моделі платформи «Ollama», оцінити їхню продуктивність на Windows і Linux, що дозволить сформувані рекомендації щодо оптимального вибору та налаштування моделей.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. GeeksforGeeks. Ollama explained: transforming AI accessibility and language processing. GeeksforGeeks. [Електронний ресурс]. — Режим доступу: <https://www.geeksforgeeks.org/ollama-explained-transforming-ai-accessibility-and-language-processing/>

2. 20 LLM evaluation benchmarks and how they work. Evidently AI - AI Testing & LLM Evaluation Platform. [Електронний ресурс]. — Режим доступу: <https://www.evidentlyai.com/llm-guide/llm-benchmarks>