Секція 2.

**ВИКОРИСТАННЯ ЛОКАЛЬНИХ МОВНИХ МОДЕЛЕЙ**

Попов Р. О.

Дніпровський національний університет ім. Олеся Гончара

Науковий керівник Карпенко Н. В.

**Актуальність**. Вплив великих мовних моделей (англ. Large Language Model, LLM) очевидний: для користувачів – це інструмент для навчання та аналізу, а дослідники постійно розробляють нові архітектури, що демонструють кращі результати. LLM також стали важливими для бізнесу, оскільки багато додатків використовують API від OpenAI, Anthropic, Mistral та інших компаній. Це зручно і ефективно, оскільки для запуску моделей потрібні великі ресурси.

Однак, використання віддалених моделей через API має недоліки, такі як проблеми з конфіденційністю, обмежена налаштовуваність та інші. Виникає питання: чи можна запускати LLM локально і які переваги це дасть порівняно з віддаленими моделями? Як це вплине на ефективність AI-додатків?

**Мета**. Розглянути інструменти для локального запуску мовних моделей та проаналізувати їхні можливості порівняно з віддаленими моделями IT-компаній.

**Осовні положення**. Для запуску великих мовних моделей необхідне коштовне обладнання та великі витрати електроенергії. Натомість, зараз активно набирає популярність тема малих мовних моделей (англ. Small Language Models, SLM), які потребують менше пам’яті та мають велику швидкість генерації виводу [1].

На обсяг ресурсів для запуску SLM впливає три основні фактори: кількість параметрів, квантизація та архітектура. Кількість параметрів зазвичай вимірюється в мільярдах, і компанії випускають моделі з різною кількістю параметрів, як-от Google Gemma 2 з 2 млрд., 9 млрд. та 27 млрд. параметрів. За допомогою квантизації можна зменшити розмір моделі та пришвидчити виконання арифметичних операцій, що досягається спрощеним представленням чисел (8, 4 або навіть 2 біти) [2].

Згідно з результатами тестувань SLM на різних збірках даних, саме архітектура та дата випуску моделі найбільше впливають на якість генерації. Різні архітектури адаптовані для різних задач, а моделі поточного року мають якість в декілька разів кращу за моделі попереднього року [1]. Внутрішній експеримент Microsoft показав, що SLM мають досить високу ефективність та в десятки разів меншу вартість обслуговування, аніж LLM [3].

Популярними SLM зараз є такі моделі: Microsoft Phi3.5, Alibaba Qwen2.5, Google Gemma 2, Meta Llama 3.2, та інші. Зазвичай більшість SLM та LLM зберігаються на сервісі Hugging Face, а стандартом де-факто для локального запуску моделей є программа llama.cpp. Інтеграція локальних моделей в існуючий AI-додаток може бути абсолютно безшовною, оскільки llama.cpp (та інші програми) можуть надавати свої послуги через HTTP API, сумісний з API OpenAI.

**Висновки**. Розглянуто питання локального запуску мовних моделей. Перевагою мовних моделей, які можна запустити локально, є повне збереження конфіденційності даних, на відміну від використання через API віддалених мовних моделей.

Існує досить багато SLM для вирішення різних задач, які знаходяться у вільному доступі, їх можна завантажити з Hugging Face та запускати локально через llama.cpp.

Проведений аналіз показує, що для задач обробки інформації SLM мають ефективність порівняну з LLM, і до того ж потребують набагато менше обчислювальних ресурсів.

**Список літератури**

1. Lu Z., Li X., Cai D. et al. Small Language Models: Survey, Measurements, and Insights. 09.2024. DOI:10.48550/ARXIV.2409.15790.

2. Li S., Ning X., Wang L. et al. Evaluating Quantized Large Language Models. 28.02.2024. DOI:10.48550/ARXIV.2402.18158.

3. Li B., Zhang Y., Bubeck S. et al. Small Language Models for Application Interactions: A Case Study. 05.2024. DOI:10.48550/ARXIV.2405.20347.

**Відомості про авторів**

Попов Руслан Олександрович, студент кафедри електронних обчислювальних машин, м.т. +380959241172, [popov\_r@365.dnu.edu.ua](mailto:popov_r@365.dnu.edu.ua)

Карпенко Надія Валеріївна, доцент кафедри електронних обчислювальних машин, ДНУ ім. Олеся Гончара, к.ф.-м.н., доцент, м.т. +380963155031, [karpenko\_n@365.dnu.edu.ua](mailto:karpenko_n@365.dnu.edu.ua)