РЕЗУЛЬТАТИ ТЕСТУВАННЯ ЛОКАЛЬНОЇ МОВНОЇ МОДЕЛІ GPT-OSS-20B

В.І. Слюсар, д.т.н., професор, ГНС.

За кілька днів до офіційного представлення GPT-5 компанія OpenAI поширила у вільний доступ свої відкриті великі мовні моделі (LLM) gpt-oss-20b та gpt-oss-120b. Їх поява в черговий раз актуалізувала проблему вибору LLM для локального застосування, без доступу до Інтернет. Метою доповіді є узагальнення кращої практики застосування LLM gpt-oss-20b для обробки україномовних текстових документів в конфіденційному режимі.

На відміну від LLM MamayLM-Gemma-2-9B-IT-v0.1, яку фахівці спеціально донавчали на україномовних датасетах, модель gpt-oss-20b була використана за принципом “як є”. Для її тестування було задіяно типовий ноутбук з відеокартою NVIDIA RTX 3060, що мала 16 ГБ відеопам’яті. Для локального запуску gpt-oss-20b використовувався фреймворк Ollama, за допомогою якого досліджувана LLM була завантажена з сайту фреймворку https://ollama.com/library/gpt-oss:20b та змонтована на вінчестер/SSD ноутбука. При цьому загальний розмір моделі, що мала квантизацію вагових коефіцієнтів MXFP4 (4,25 біт), становив 12,83 ГБ.

В ході дослідження в черговий раз було підтверджено, що продуктивність і потреби в ресурсах однієї й тієї ж локальної LLM залежать від встановленого розміру контекстного вікна. У рамках фреймворку Ollama для gpt-oss-20b це вікно могло дискретно обмежуватися рівнем 4K, 8K, 16K, 32K, 64K та 128K. При цьому автором вивчалася можливість застосування різних за розміром контекстних вікон в залежності від наявного після встановлення LLM вільного простору на вінчестері. Крім того, важливим питанням було з’ясування максимального контекстного вікна, при якому gpt-oss-20b могла повністю розміститись і здійснювати інференс у пам’яті відеокарти.

В результаті експериментів вдалося з’ясувати, що контекстні вікна 4K, 8K та 16K дозволяють повністю змонтувати gpt-oss-20b на зазначену відеокарту RTX 3060. При цьому запуск LLM не вимагав залучення додаткового простору на SSD. Такий результат дозволяє висунути гіпотезу, що для обробки контекстного запиту обсягом 1K необхідно приблизно 0,25 ГБ відеопам’яті.

В якості тестового завдання при експериментах використовувався переклад на українську мову тексту англомовного стандарту НАТО обсягом 2600 слів. При цьому було виявлено, що такий обсяг слів

перевищував контекстне вікно 8K. Недоліком фреймворку Ollama є неможливість попередньо визначити, чи виходить кількість токенів у текстовому запиті за межі встановленого ліміту контексту. Тому при запуску зазначеного тестового завдання мовна модель його виконувала, але при досягненні верхньої межі контекстного обсягу починались галюцинації з зацикленням у часі генерації тексту.

У разі збільшення розмірів контекстного вікна до 32K та 64K модель gpt-oss-20b залучала для роботи додатково відповідно приблизно 6 та 7 ГБ наявного на вінчестері вільного простору. Це дозволяє зробити припущення, що для таких налаштувань розмірів контексту запуск gpt-oss-20b можливий на відеокарті з 24 ГБ відеопам'яті. При граничній межі контекстного обсягу 128K запуск LLM залучав додатково до наявних 16 ГБ відеопам'яті ще приблизно 29 ГБ на вінчестері. У разі їх відсутності в процесі інференсу фреймворк Ollama видавав помилку: "500 Internal Server Error: llama runner process has terminated: exit status 2 ". Таким чином, для запуску gpt-oss-20b з максимальним контекстним вікном цілковито на відеокарті необхідно мати відеопам'ять обсягом 48 ГБ.

Стосовно якості перекладу слід зазначити, що вона цілком очікувано поступалася результатам GPT-4o. Разом з тим, у переважній більшості випадків отримані україномовні тексти були краще структуровані за змістом у порівнянні з LLM MamayLM-Gemma-2-9B-IT-v0.1. При цьому запуск gpt-oss-20b за замовчуванням завжди розпочинався режимом ланцюга роздумів. Тривалість його фази під час інференсу досить сильно залежала від розмірів текстового запиту. Наприклад, при обробці зазначеного тестового завдання обсягом 2600 слів до появи першого слова перекладеного тексту могло пройти від 92 до 112 сек, залежно від обсягів контекстного вікна та інших випадкових впливів. В той же час, відповідь на звернення "Привіт" могла зайняти від 0,9 сек при контекстному вікні 4K до 6,8 – 9,7 сек - при вікні 128K.

Таким чином, слід зробити висновок, що мовна модель gpt-oss-20b без будь яких донавчань українській мові може бути використана для вирішення завдань обробки конфіденційних документів. При цьому за якістю перекладів стандартів НАТО вона не тільки не поступається LLM MamayLM-Gemma-2-9B-IT-v0.1, яку навмисно донавчали на україномовних датасетах, а й у багатьох випадках дає дещо кращі результати. Подальші дослідження будуть зосереджені на використанні інших фреймворків, призначених для локального розгортання LLM, а також тестуванні більш потужної моделі gpt-oss-120b та розширенні кола тестових завдань.