АНАЛІЗ ПРОДУКТИВНОСТІ ЛОКАЛЬНО РОЗГОРНУТИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

В.І. Слюсар к.т.н., професор, ГНС;
І.П. Гусаковський, СНС.

Зростання ролі штучного інтелекту ставить на порядок денний питання розгортання та ефективного використання великих мовних моделей в ізольованих локальних мережах. Це є критично важливим для забезпечення конфіденційності даних та автономності роботи. Метою даного дослідження є оцінка продуктивності та практичної придатності популярних та кастомізованих LLM для виконання різних завдань на українській мові.

Для проведення експериментів було розгорнуто цифровий полігон на базі двох серверів інференсу:

1.Сервер на основі Ollama, де тестувалися моделі: llama3.1 (8b), llama3.2 (3b), Mistral (7b), llama3 (8b), Llava (7b), Llava-llama3 (8b), Deepseek-R1 (7b).

2.Сервер на основі KoboldCpp, де тестувалися кастомізовані моделі: Mistral (7b), MamayLM-Gemma-2-9B-IT-v0.1.Q4_K_M, MamayLM-Gemma-2-9B-IT-v0.1.Q4_K_S та MamayLM-Gemma-2-9B-IT-v0.1.Q5_K_S.

Оцінка моделей проводилася за такими критеріями:

1. Рівень володіння українською мовою (здатність генерувати граматично та стилістично коректний текст).
2. Точність перекладу (якість перекладу з англійської мови на українську у порівнянні з еталонними перекладами від Google Translate, Google Gemini та OpenAI ChatGPT).
3. Швидкість інференсу (продуктивність моделі, виміряна у символах (токенах) за секунду).
4. Розуміння контексту (здатність підтримувати логічний та зв'язний діалог).
5. Підтримка мультимодальності (здатність обробляти зображення).

У ході випробувань було встановлено, що більшість протестованих моделей демонструють задовільний рівень володіння українською мовою та розуміння контексту. Однак, за сукупністю показників, найкращі результати продемонструвала кастомізована модель MamayLM-Gemma-2-9B-IT-v0.1.Q5_K_S, запущена на сервері KoboldCpp. Її продуктивність та якість генерації тексту, зокрема у завданнях перекладу, виявилися релевантними та співставними з результатами,

отриманими від провідних хмарних сервісів, таких як Google Gemini та ChatGPT.

Проведене дослідження доводить, що локально розгорнуті, зокрема кастомізовані, великі мовні моделі здатні досягати високого рівня продуктивності для специфічних мовних завдань. Модель MamayLM-Gemma-2-9B-IT-v0.1.Q5_K_S у поєднанні з оптимізованим сервером генерації KoboldCpp є ефективним рішенням, що може слугувати конкурентоспроможною альтернативою популярним онлайн-сервісам з обробки текстів, забезпечуючи при цьому повний контроль над даними.