

УДК 004.8

П. Містерман; М. Петрик, д. ф.-м. н., проф.

(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

**РОЗРОБКА МАЛОЇ МОВНОЇ МОДЕЛІ SLM ДЛЯ СИСТЕМ
З ОБМЕЖЕНИМИ ОБЧИСЛЮВАЛЬНИМИ РЕСУРСАМИ
НА БАЗІ AZURE ML ТА AI FOUNDRY**

UDC 004.8

P. Misterman; M. Petryk, DR., Prof.

**DEVELOPMENT OF A SMALL LANGUAGE MODEL (SLM)
FOR RESOURCE-CONSTRAINED SYSTEMS BASED ON
AZURE ML AND AI FOUNDRY**

У сучасних умовах стрімкого розвитку штучного інтелекту все більшої уваги набуває використання малих мовних моделей (SLM), які здатні забезпечувати високу ефективність при суттєво нижчих вимогах до обчислювальних ресурсів порівняно з великими мовними моделями (LLM). Зростаючий попит на локальні та економічно ефективні рішення зумовлює необхідність адаптації SLM до різних прикладних задач — класифікації, екстракції даних, діалогових систем та автоматизації бізнес-процесів.

Платформа Azure Machine Learning у поєднанні з Azure AI Foundry надає можливості для навчання, керування та розгортання моделей різного масштабу, включно з оптимізованими SLM, що дозволяє створювати продуктивні системи навіть у середовищах з обмеженими ресурсами. Одним із ключових аспектів роботи є дослідження можливостей використання малої моделі замість LLM у задачах, де традиційно очікується висока якість мовного розуміння, а також оцінювання продуктивності, швидкодії, вартості та точності.

Під час розробки рішення передбачено використання технік донавчання моделей, таких як LoRA та QLoRA, а також застосування форматів оптимізації типу GGUF та інструментів прискорення інференсу, включно з vLLM. Це дозволяє забезпечити високу швидкість відповіді, зменшити споживання пам'яті та адаптувати модель до вузької предметної області в умовах реальних бізнес-сценаріїв.

Практичне значення роботи полягає у створенні прототипу системи, здатної виконувати типові завдання штучного інтелекту без значних апаратних витрат. Це відкриває можливості для впровадження інтелектуальних рішень у малих організаціях, на периферійних пристроях або в середовищах з підвищеними вимогами до приватності та локальності обробки даних. Результати дослідження демонструють потенціал SLM як життєздатної альтернативи LLM для широкого спектра задач, забезпечуючи баланс між точністю, продуктивністю та доступністю.

Література

1. Azure Machine Learning documentation | Microsoft Learn. URL: <https://learn.microsoft.com/uk-ua/azure/machine-learning/?view=azureml-api-2>.
2. What is Microsoft Foundry? - Microsoft Foundry | Microsoft Learn. URL: <https://learn.microsoft.com/en-us/azure/ai-foundry/what-is-azure-ai-foundry?view=foundation-classic>.
3. Maninder S. *Practical Guide to Fine-tune LLMs with LoRA*. 13.10.2024. URL: <https://medium.com/@manindersingh120996/practical-guide-to-fine-tune-llms-with-lora-c835a99d7593>.