

ВИКОРИСТАННЯ ТА НАВЧАННЯ ГОТОВИХ МОВНИХ МОДЕЛЕЙ У СУЧАСНИХ СИСТЕМАХ ШТУЧНОГО ІНТЕЛЕКТУ

К. Буравченко, доцент

Н. Козірова, викладач

Р. Ткачук, асистент

Центральноукраїнський національний технічний університет

З огляду на швидке поширення та зростаючий потенціал великих мовних моделей (LLMs), постає потреба у дослідженні способів їх ефективного використання в умовах обмежених ресурсів, зокрема в локальному середовищі. Актуальним є вивчення інструментів, що забезпечують запуск таких моделей без хмарних обчислень, а також методів донавчання, які дозволяють адаптувати загальні моделі під конкретні прикладні завдання. Метою цієї роботи є аналіз можливостей використання відкритих мовних моделей, таких як LLaMA та Mistral, із застосуванням середовища LM Studio, а також дослідження сучасних підходів до їх донавчання, включаючи instruction tuning та LoRA.

Сучасні досягнення у сфері штучного інтелекту (ШІ) значною мірою ґрунтуються на використанні великих мовних моделей (LLMs), які вже були попередньо навчені на масивних текстових корпусах. Такий підхід дає змогу швидко створювати прикладні рішення без потреби навчання моделі з нуля. У роботі розглядаються приклади використання відкритих моделей, зокрема **Meta LLaMA** та **Mistral**, які демонструють високу ефективність навіть при локальному запуску без підключення до хмарних обчислювальних платформ.

Окрему увагу приділено інструменту **LM Studio**, що забезпечує інтерфейс для використання моделей LLaMA, GPT, Mistral та інших у десктопному середовищі. Його перевагою є підтримка інструкційного режиму, простота запуску моделей через формат GGUF, та можливість роботи без підключення до інтернету, що забезпечує конфіденційність та автономність застосунків.

Особливий акцент зроблено на **процесі донавчання (fine-tuning)** моделей:

- Розглянуто **методи інструкційного донавчання (instruction tuning)**, де модель адаптується до конкретних шаблонів завдань або стилів відповіді;
- Пояснено застосування техніки **LoRA (Low-Rank Adaptation)** — підходу, що дозволяє донавчати лише частину параметрів моделі;
- Описано використання бібліотек **Hugging Face Transformers**, які підтримують адаптацію моделей як у хмарному середовищі, так і локально.

Fine-tuning дозволяє значно зменшити витрати на навчання та зберігає основні властивості базової моделі, водночас підвищуючи її релевантність у конкретному контексті.

Висновок

Використання попередньо навчених мовних моделей значно спрощує створення систем штучного інтелекту, знижуючи витрати на ресурси та час розробки. Інструменти, такі як LM Studio, забезпечують зручне середовище для локального запуску та експериментів з моделями без потреби підключення до інтернету. Методи навчання, зокрема instruction tuning та LoRA, дозволяють ефективно адаптувати моделі до конкретних завдань без повного перенавчання, що робить такі підходи доцільними для широкого кола користувачів – від ентузіастів до професійних розробників. Таким чином, комбіноване використання готових рішень і легковагового навчання відкриває нові можливості для створення конфіденційних, гнучких та автономних ШІ-систем.

Список літератури

1. Основний курс Huggingface щодо бібліотеки transformers [Електронний ресурс] <https://huggingface.co/learn/llm-course/chapter1/1>
2. Документація з використання Llama [Електронний ресурс] <https://github.com/meta-llama/llama3>
3. Документація з використання Mistral [Електронний ресурс] <https://docs.mistral.ai/getting-started/quickstart/>