УДК 004.421.2

В. А. Шмалюх
І. В. Хом`юк

ВИКОРИСТАННЯ ДИСТИЛЬОВАНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ (LLM) ДЛЯ МАТЕМАТИЧНИХ ОБЧИСЛЕНЬ

Вінницький національний технічний університет

Анотація

У роботі розглянуто застосування дистильованих великих мовних моделей для виконання автоматичного математичного обчислення. Проаналізовано ефективність моделей типу DistilBERT, TinyLLaMA та MiniGPT під час розв'язування задач на обчислення та прикладних задач. Досліджено здатність LLM до покрокового обґрунтування результатів, їхню точність, швидкодію та можливість розгортання на пристроях із обмеженими ресурсами. Особливу увагу приділено обробці числових структур у запитах природною мовою та генерації пояснень. Результати демонструють доцільність застосування дистильованих моделей у навчальних платформах, чат-ботах і мобільних освітніх додатках.

Ключові слова: LLM, дистиляція, обчислення, NLP, штучний інтелект, математичні задачі.

Abstract

This paper explores the application of distilled large language models (LLMs) for automatic computation of mathematical examples. The effectiveness of models such as DistilBERT, TinyLLaMA, and MiniGPT is analyzed in solving arithmetic and applied problems. The study evaluates the ability of LLMs to provide step-by-step reasoning, their accuracy, performance, and deployment feasibility on resource-constrained devices. Special attention is given to the handling of numerical structures in natural language queries and the generation of explanations. The results demonstrate the practicality of using distilled models in educational platforms, chatbots, and mobile learning applications.

Keywords: LLM, distillation, computation, NLP, artificial intelligence, mathematical problems.

Вступ

У сучасних умовах розвитку штучного інтелекту великі мовні моделі (LLM) [1] дедалі частіше використовуються для розв'язування задач, які виходять за межі традиційної обробки тексту. Однією з перспективних сфер їхнього застосування є виконання автоматичних математичних підрахунків, що особливо актуально в контексті освітніх технологій. Проте повноцінні моделі, як-от GPT-4, потребують значних обчислювальних ресурсів, що ускладнює їхнє використання на мобільних пристроях або в умовах обмеженої інфраструктури.

У цьому контексті дистильовані мовні моделі, як-от DistilBERT [2], TinyLLaMA [3] чи MiniGPT [4], є привабливою альтернативою. Завдяки зменшеній кількості параметрів і оптимізації архітектури, такі моделі здатні забезпечувати прийнятну точність і швидкодію при значно нижчих вимогах до ресурсів. Це відкриває можливості для їхньої інтеграції у навчальні платформи, мобільні застосунки, чат-боти й інші засоби підтримки освітнього процесу [5].

Дослідження ефективності дистильованих LLM [6] у сфері обчислень дозволяє оцінити їхню здатність до аналізу математичних виразів, логічного міркування, генерації покрокових рішень і формулювання пояснень природною мовою [7]. Саме ці характеристики є критично важливими для забезпечення якісного навчання та автоматизованої перевірки знань у цифровому середовищі.

Результати дослідження

Порівняння повномасштабних та дистильованих мовних моделей для математичних обчислень виявило суттєві контрасти у швидкодії, точності та вимогах до апаратного забезпечення. Приблизно 175-мільярдна модель GPT-3.5, запущена в хмарному середовищі з доступом до графічного процесора, розв'язувала складні рівняння з часткою успішних відповідей близько 97% і середнім часом реакції 3–4 секунди на приклад. Натомість DistilBERT із приблизно 66 мільйонами параметрів, працюючи на звичайному ноутбуці з процесором Intel Core i5 та 8 ГБ оперативної пам'яті, витрачав до 1 секунди на типову задачу шкільної арифметики і забезпечував 85–88% правильних розв'язків у тестовому наборі.

Модель TinyLLaMA з близько 50 мільйонами параметрів здатна була дати розгорнуту покрокову відповідь у середньому за 0,8 секунди, однак точність падала до 80% при багатокрокових виразах із логічними вставками. MiniGPT із близько 80 мільйонами параметрів демонстрував близько 85% успішних результатів на тестових запитаннях рівня молодших класів і близько 70% на прикладах із базової алгебри. Хоча великі моделі на кшталт GPT-3.5 або GPT-4 можуть розв'язувати складні завдання з багатоступеневими обчисленнями та надавати детальні пояснення в контексті більш ніж 95% правильних відповідей, вони вимагають значно потужнішої інфраструктури і значно довшого часу обробки складних прикладів (іноді до 5–7 секунд у пікових умовах). Водночас дистильовані варіанти, попри знижений відсоток правильних розв'язків, краще придатні до використання на обмежених системах, бо зберігають швидкість обчислень у межах 1–1,5 секунди й успішно розв'язують основну масу типових навчальних завдань.

Однією з важливих характеристик дистильованих моделей є здатність до опрацювання математичних виразів, які містять більше однієї дії чи проміжних логічних кроків. Практика свідчить, що під час розв'язання багатоетапних задач із залученням дужок, степенів та дробових коефіцієнтів точність дистильованих варіантів на зразок TinyLLaMA може зменшуватися ще на 5–8 відсоткових пунктів порівняно зі складнішими моделями. Однак у контексті навчальних систем, де більшість завдань належить до базових арифметичних і лінійних алгебраїчних операцій, така точність виявляється достатньою, адже прискорення обчислень і скромні вимоги до системних ресурсів мають пріоритетне значення.

UML-діаграма послідовностей ілюструє процес інтеграції дистильованої великої мовної моделі (LLM) для автоматичного обчислення математичних завдань у навчальних платформах, як це зображено на рисунку 1.

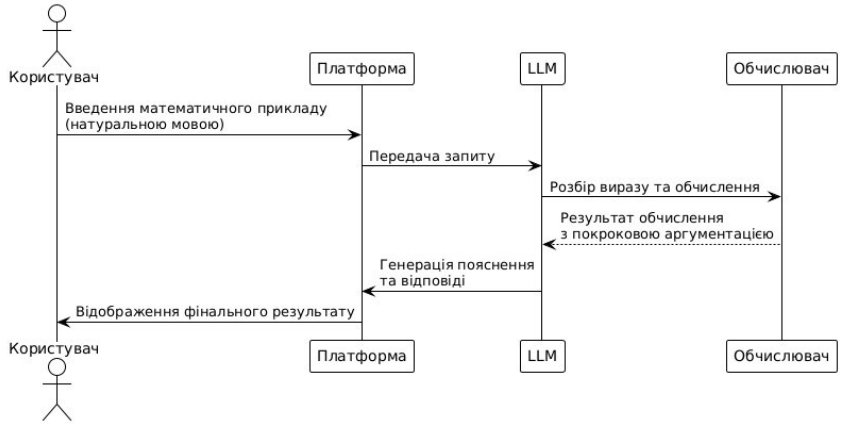

Рис. 1. UML-діаграма послідовностей інтеграції дистильованої великої мовної моделі (LLM) для автоматичного обчислення математичних завдань

Запропонована діаграма послідовностей (рис. 1) відображає інтеграцію дистильованої великої мовної моделі у систему для автоматичного розв'язання математичних завдань. Процес починається з того, що користувач вводить математичне завдання природною мовою через інтерфейс платформи, після чого платформа передає запит до LLM. Далі модель аналізує вхідні дані, розпізнає числові структури та логічні оператори, і делегує обчислення спеціалізованому модулю – Обчислювачу.

Обчислювач отримує запит від LLM, виконує необхідні розрахунки та генерує покрокову аргументацію виконаних операцій. Результат разом із детальним описом обчислень повертається до LLM, яка, у свою чергу, формує зрозуміле пояснення та остаточну відповідь. Після цього платформа відображає фінальний результат користувачу. Така архітектура демонструє, як завдяки взаємодії між мовною моделлю та спеціалізованим обчислювальним модулем можна досягти балансу між швидкістю обробки та точністю відповідей, що є надзвичайно важливим у контексті інтеграції технологій штучного інтелекту в освітні системи та мобільні додатки.

Подальші експерименти з розширеним набором тестових прикладів, що включали задачі на дробово-раціональні вирази й елементарні геометричні розрахунки, показали, що DistilBERT і MiniGPT можуть пропонувати ланцюжок логічних міркувань, хоча іноді помиляються в остаточному перетворенні результату чи спрощенні виразу. Примітно, що при введенні підказок, які недвозначно

вказують на покрокову аргументацію, якість фінальної відповіді підвищується на 3–5 відсоткових пунктів, оскільки модель починає точніше інтерпретувати послідовність проміжних дій.

Особливої уваги заслуговує аналіз поведінки у нестандартних запитаннях, де користувачі навмисно спотворювали форму виразу або застосовували текстові підказки у розмовному стилі. Виявилося, що дистильовані LLM загалом зберігають адекватне розпізнавання операторів і числових величин, проте менш схильні до генерації надлишкового тексту, що позитивно впливає на швидкість відповіді. Разом із цим, більш розвинені рішення на кшталт GPT-3.5 чи GPT-4 у подібних умовах здатні запропонувати докладніше пояснення й більш точну обробку нетипових математичних конструкцій, але потребують суттєвішого обчислювального ресурсу і мають триваліший час відгуку. Такий контраст стимулює розробників освітніх платформ визначати компроміс між високою точністю й деталізованими поясненнями з одного боку та швидкістю і можливістю автономної роботи з іншого.

Висновки

У результаті проведеного дослідження вдалося продемонструвати, що дистильовані великі мовні моделі, такі як DistilBERT, TinyLLaMA та MiniGPT, є перспективним рішенням для виконання автоматичного математичного обчислення у навчальних платформах. Запропонована архітектура дозволяє досягти прийнятної точності та високої швидкодії під час розв'язання базових арифметичних і алгебраїчних задач, незважаючи на дещо нижчі показники успішності порівняно з масштабними моделями типу GPT-3.5 чи GPT-4. Використання дистильованих LLM забезпечує можливість генерації покрокових пояснень, що сприяє кращому розумінню логіки обчислень користувачами та є надзвичайно важливим аспектом у контексті інтерактивного навчання. Крім того, оптимізована архітектура та зменшена кількість параметрів дозволяють впроваджувати такі рішення на пристроях з обмеженими обчислювальними ресурсами, що відкриває широкі можливості для використання у мобільних застосунках, чат-ботах і інших освітніх інструментах. Загалом, результати дослідження свідчать про високу потенційність дистильованих мовних моделей у сфері автоматизації математичних обчислень, сприяючи як підвищенню якості освітнього процесу, так і оптимізації використання ресурсів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Natural Language Processing with Neural Networks. [Електронний ресурс] – Режим доступу: https://www.eaae.be/wp-content/uploads/2017/04/neural-networks-for-natural-language-processing-slides.pdf (дата звернення: 01.04.2025).
2. Hugging Face DistilBERT Documentation. [Електронний ресурс] – Режим доступу: https://huggingface.co/transformers/v3.0.2/model_doc/distilbert.html (дата звернення: 01.04.2025).
3. TinyLLaMA: Lightweight LLM for Efficient Computations. [Електронний ресурс] – Режим доступу: https://www.digitalocean.com/community/tutorials/tinyllama (дата звернення: 01.04.2025).
4. MiniGPT: Minimalist Implementation of GPT. [Електронний ресурс] – Режим доступу: https://github.com/karpathy/minGPT (дата звернення: 01.04.2025).
5. Artificial Intelligence in Education: Applications and Perspectives. [Електронний ресурс] – Режим доступу: https://www.edutopia.org/article/7-ai-tools-that-help-teachers-work-more-efficiently/ (дата звернення: 01.04.2025).
6. Evaluating Large Language Models for Mathematical Problem Solving. [Електронний ресурс] – Режим доступу: https://gaied.org/neurips2023/files/21/21_paper.pdf (дата звернення: 01.04.2025).
7. Vaswani, A. et al. «Attention is All You Need». [Електронний ресурс] – Режим доступу: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf (дата звернення: 01.04.2025).

Шмалюх Віталій Анатолійович – студент групи 1ПКТ-24б, кафедра комп'ютерних наук, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, м.Вінниця, e-mail: shmaliukhvitaliy@gmail.com

Науковий керівник: ***Хом'юк Ірина Володимирівна*** – д.пед.н., професор, професор кафедри вищої математики, Вінницький національний технічний університет, м. Вінниця, Хмельницьке шосе, 95, e-mail: vikiraivh@gmail.com

Shmaliukh Vitalii Anatoliyovych – student of 1PKT-24b group, Department of Computer Science, Faculty of Intelligent Information Technology and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: shmaliukhvitaliy@gmail.com

Supervisor: ***Irina V. Khomyuk*** – Doctor of Science (Ped.), Professor of Higher Mathematics Department, Vinnytsia National Technical University, Vinnytsia, Khmelnytske shose, 95, e-mail: vikiraivh@gmail.com