УДК 004.89 DOI: https://doi.org/10.30837/IYF.IIS.2024.086

ТЕСТУВАННЯ ЛОГІЧНИХ ЗДІБНОСТЕЙ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Лавриненко С. Р.
Науковий керівник – к.т.н., доц. Рябова Н. В.
Харківський національний університет радіоелектроніки, каф. ШІ
м. Харків, Україна
e-mail: sofiia.lavrynenko@nure.ua

This work is devoted to investigating the performance of Large Language Models (LLMs) in solving logical problems in Ukrainian, where key words are replaced with nonsensical ones to challenge the models' reliance on prior knowledge. It highlights a study comparing the abilities of four models—ChatGPT 3.5, ChatGPT 4.0, Copilot, and Gemini—across different testing scenarios, including both isolated and contextual problem-solving. The findings reveal that all models significantly outperform random guessing, with ChatGPT 4.0 showing exceptionally high accuracy, suggesting its potential in applications requiring complex logical reasoning.

Генеративний штучний інтелект (Generative Artificial Intelligence, GenAI) зайняв передові позиції на ринку сучасних інтелектуальних інформаційних технологій. GenAI розвивається як під область Deep Learning та Machine Learning і має за мету розробку методів і моделей для генерації об’єктів заданого типу, перетворення об’єктів одного типу до іншого (generative modeling, text-to-image synthesis, image-to-image translation). Завдяки стрімкому розвитку технологій глибоких нейронних мереж GANs (Generative Adversarial Networks, тобто генеративних змагальних мереж) та найбільш перспективному типу моделей глибоких нейронних мереж типу Transformers, поширеним лозунгом GenAI сьогодні став “anything-to anything transformation”, тобто генерація будь-яких об’єктів на основі будь-яких об’єктів заданого типу [1]. Найбільш вражаючих результатів досягнуто в обробці людської мови.

Великі мовні моделі (Large language models, LLM) отримали широке розповсюдження за останні півтора роки. Складається враження, що LLM мають надзвичайні можливості для розуміння, створення та інтерпретації людської мови. Однак, коли ми говоримо, що LLM «розуміють», маємо на увазі, що вони можуть обробляти та генерувати текст у спосіб, який виглядає зв’язним і релевантним контексту, а не те, що вони володіють людською свідомістю чи розумінням [2].

Зараз кожному користувачеві доступні запити до декількох LLM, і може навіть здаватися що ці моделі мають ознаки наявності свідомості. В ЗМІ повідомляють що наступні версії цих моделей можуть стати сильним штучним інтелектом. Не дивно що науковці вже тестують, наскільки мовні моделі справляються зі складним плануванням [3] та порівнюють їх

відповіді з відповіддю людей-експертів [4]. Ми провели дослідження, чи справляються сучасні логічні моделі з задачами на логіку українською мовою, де іменники та / або дієслова замінені на безглузді слова, які не дають моделі асоціювати їх з певними попередніми знаннями.

Такий тест зокрема використовувався на кафедрі штучного інтелекту Харківського національного університету радіоелектроніки для тестування студентів з дисципліни «Дискретна математика». Складається з 30 питань з трьома варіантами відповідей, з яких лише одна вірна. Для уникнення асоціації з предметами реального світу іменниками є безглузді видумані слова. Оскільки даний тест популярний в інтернеті та міг бути використаний при навчанні мовних моделей, то ми повторно змінили іменники на інші безглузді слова, а також провели ще декілька варіантів щодо ускладнення задачі, наприклад замінивши всі іменники одним словом. При цьому всі 30 питань задавалися в одному чаті (зберігався контекст, що ускладнювало задачу).

Одним з досліджень було порівняння успішності моделей для випадку коли кожна задача – окремий чат (контекст інших задач недоступний), і випадку де всі задачі задаються послідовно в одному чаті.

Тестування проведено з чотирма моделями: ChatGPT 3.5, ChatGPT 4.0, Copilot та Gemini [5]. Таблиця відповідей доступна за посиланням docs.google.com/spreadsheets/d/15skQkVHV4ozDr_hC6GNpVaCBtxGcRCRO Приклад запитів та відповідей наведено в таблиці 1. Зведені результати моделей в розрізі експериментів наведені в таблиці 2.

Таблиця 1 – Приклад запитів та відповідей моделей

Запит	ChatGPT 3.5	ChatGPT 4.0	Copilot	Gemini
Шашкин боїться як качей, і паганів. a) Шашкин не боїться паганів; b) Шашкин боїться качей; c) Шашкин боїться качей більше, ніж паганів, але й паганів боїться також.	+	+	–	–
Відомо, що карнатик обов'язково або синій, або солом'яний, або те й інше разом. a) карнатик не може бути солом'яним; b) карнатик не може бути червоним і несолом'яним одночасно; c) карнатик не може бути синім і солом'яним одночасно.	+	+	+	+
Існують жешки з хворою індою. a) не всяка жешка може похвалитися здоровою індою; b) не всяка жешка може похвалитися хворою індою; c) існують жешки зі здоровою індою.	–	+	–	–

Таблиця 2 – Успішність моделей в експериментах

Запит	ChatGPT 3.5	ChatGPT 4.0	Copilot	Gemini
Спільний чат, версія 1 питань	18/30	28/30	25/30	19/30
Спільний чат, версія 2 питань	15/30	25/30	20/30	18/30
Спільний чат, версія 3 (всі іменники одинакові)	16/30	24/30	20/30	19/30
Окремі чати, версія 3 (всі іменники одинакові)	18/30	23/30	23/30	18/30
Загальні підсумки	67/120	100/120	88/120	74/120

Отримані результати показують, що всі чотири моделі справляються з логічними задачами значимо краще ніж при випадковому вгадуванні. Треба відмітити, що в експериментах Copilot використовувався в режимі «Точний», оскільки при випробуванні режимів «Творчий» та «Збалансований» результат був на рівні випадкового вгадування.

Проведене дослідження показує, що не всі моделі одинаково успішні у вирішенні задач, де корпус текстів не допомагає. При цьому можемо впевнено сказати що ChatGPT 4.0 дає надзвичайно високий навіть для людини результат. Помічена відмінність ChatGPT 4.0 при наданні відповідей: спочатку задача перетворювалася в логічний вираз, а потім по ній виводилася відповідь. Це вирізняло від прийому використаного Gemini, де здебільшого спочатку надана відповідь пізніше пояснювалася. Таке пізніше пояснення вже зробленого вибору звичайно ж не приносило жодного покращення, тому що обґрунтовувало вже прийняту неправильну відповідь. Аналізуючи вплив попереднього контексту можна сказати, що задавати задачу в окремому чаті чи всі задачі в одному чаті суттєвої різниці немає.

Наявність таких логічних здібностей у мовних моделей підтверджує можливість використовувати мовні моделі як складові в майбутніх моделях сильного штучного інтелекту.

Список використаних джерел:

1. Tauli T. Generative AI: How ChatGPT and Other AI Tools Will Revolutionize Business. 1st ed. Apress, 2023. 216 p.

2. Raschka S. Build a Large Language Model. Manning, 2024. 400 p.

3. Shen Y., Song K., Tan X., Li D.S., Lu W., Zhuang Y.T. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face : ArXiv, abs/2303.17580, 2023.

4. Guo B., Zhang X., Wang Z., Jiang M., Nie J., Ding Y., Yue J., Wu Y. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection : ArXiv, abs/2301.07597, 2023.

5. Team, Gemini, et al. Gemini: A Family of Highly Capable Multimodal Models : ArXiv abs/2312.11805, 2023.