УДК 004.89:004.912

ДОСЛІДЖЕННЯ ПИТАННЯ ВИКОРИСТАННЯ LLMS ДЛЯ ГЕНЕРАЦІЇ ТЕСТІВ З МЕТОЮ МОНІТОРИНГУ РОЗУМІННЯ ПРОЧИТАНОГО МАТЕРІАЛУ

Талах В.О.
email: vladyslav.talakh@nure.ua
Науковий керівник – к.т.н., доц. Яковлева О.В.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

This paper is devoted to the study of using LLMs to generate tests based on book texts to monitor reading understanding. Five models were chosen to select the best one, tests were generated, and given to students to take. After passing each test, students left their scores for this test. Evaluation results led to the selection of the optimal model, considering scores, price, and generation speed. Based on the selected model, further work will be done to develop a reading application with tests generation, tests execution and control of reading understanding.

У сучасному світі спостерігається тенденція на зменшення кількості читачів серед молоді. Молодь, яка виросла в епоху комп'ютерних ігор та швидкого медіаконтенту, надає меншу перевагу читанню, що погано впливає на когнітивні навички. Для підвищення інтересу до читання та покращення розуміння тексту можуть допомогти додатки, що автоматично генерують інтерактивний контент на основі прочитаного: тести, ілюстрації, квізи, тощо. Сфера застосування подібних додатків виходить за межі особистого користування. Вони можуть бути інтегровані в навчальний процес, де вчителі стикаються з проблемою заохочення учнів до читання, або у сім'ї, де батьки прагнуть розвинути в дітях любов до книг.

Останні роки показують значний прогрес у сфері штучного інтелекту. Великі мовні моделі, системи комп'ютерного зору та мультимодальні моделі досягли вражаючих результатів у різноманітних завданнях [1, 2]. Ці технології знаходять застосування в різних сферах. Перспективним є застосування LLM для аналізу текстів та генерації запитань, що дозволяє створювати інструменти для оцінювання розуміння прочитаного.

Аналіз додатків для читання виявив значні обмеження в їхній функціональності щодо моніторингу розуміння прочитаного. Більшість популярних додатків фокусуються на базових функціях читання. Існують додатки, які включають елементи тестування, проте вони обмежені попередньо створеним контентом і не дозволяють генерувати тести для довільних текстів, завантажених користувачем.

Таким чином, на ринку спостерігається відсутність сервісів, що дозволяють автоматично генерувати тести на основі завантажених користувачем

книг. Такий додаток мав би значний потенціал для використання як у освітньому процесі, так і для індивідуальних користувачів чи сімей.

Важливою частиною створення подібного застосунку є вибір моделі, яка буде показувати найкращі показники генерації питань. Для проведення тестування було залучено учнів, які проходили попередньо згенеровані тести на основі однакових фрагментів тексту книги та залишали відгуки за 5-бальною шкалою від -2 до 2 за показниками: коректність питань, коректність варіантів відповідей та цікавість питань. Для проведення тестування було відібрано 5 популярних моделей за статистикою з відкритих джерел [3, 4], які за попередніми тестами показали кращі результати генерації тестів, а саме: GPT 4o, Gemini 1.5 Pro, Gemini 2.0 Flash, Claude 3.5 Sonnet, Claude 3.7 Sonnet. Для простоти та швидкості розробки платформи для проведення тестування було використано Python бібліотеку Gradio для створення застосунку, де залучені користувачі змогли б проходити тести і залишати відгуки, та платформу HuggingFace для його хостингу [5].

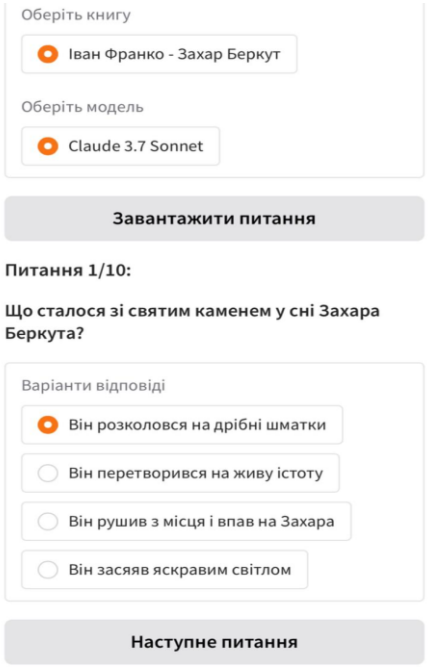

Рисунок 1 – Процес проведення тестування у застосунку

У тестуванні взяло участь близько 50 експертів (учні та вчителі школи) (талб.1). За результатами оцінювання моделі було відсортувано за якістю: Claude 3.7 Sonnet > Claude 3.5 Sonnet > Gemini 2.0 Flash > GPT 4o > Gemini 1.5 Pro; за часом: Gemini 2.0 Flash > Gemini 1.5 Pro > Claude 3.5 Sonnet > Claude 3.7 Sonnet > GPT 4o; за ціною: Gemini 2.0 Flash >> Gemini 1.5 Pro >> GPT 4o > Claude 3.5 Sonnet > Claude 3.7 Sonnet (перша зліва модель – краща).

Як бачимо, Claude 3.7 Sonnet має значну перевагу за результатами оцінювання, проте вона виявилась найдорожчою, та передостанньою за швидкістю. Показник швидкості в даній задачі не є настільки значним, оскільки питання будуть генеруватися у фоновому режимі. Показник ціни, в свою

чергу, є доволі значним. Тому Claude 3.5 Sonnet, є гарним варіантом для цієї задачі. Вона не сильно відстає від найкращої моделі, при цьому все ще має перевагу над іншими конкурентами, та має прийнятну ціну.

Таблиця 1 – Результати оцінювання моделей LLMs

Модель	Коректність питань	Коректність варіантів відповідей	Цікавість питань	Середня оцінка	Час (хв:с)	Ціна ($)
GPT 4o	0.79	0.67	0.71	0.72	04:02	0.127
Gemini 1.5 Pro	0.77	0.73	0.62	0.71	02:14	0.064
Gemini 2.0 Flash	0.77	0.87	0.64	0.76	01:17	0.006
Claude 3.5 Sonnet	0.87	0.85	0.91	0.88	02:47	0.154
Claude 3.7 Sonnet	0.91	0.98	1.00	0.96	03:15	0.259

Генерація тестів є основним функціоналом майбутньому застосунку, тому даний дослідницький етап є ключовим. Обрана модель буде використана у застосунку, який дозволить користувачам завантажувати свої книги, проходити тести, слідкувати за статистикою читання та проходження тестів, а також може бути інтегрований в освітній процес.

Список використаних джерел:

1. Application a Committee of Kohonen Neural Networks to Training of Image Classifier Based on Description of Descriptors Set / V. Gorokhovatskyi et al. *IEEE Access*. 2024. P. 1. URL: https://doi.org/10.1109/access.2024.3404371.

2. Gorokhovatskyi O., Yakovleva O. Medoids as a packing of ORB image descriptors. *Advanced Information Systems*. 2024. Vol. 8, no. 2. P. 5–11. URL: https://doi.org/10.20998/2522-9052.2024.2.01.

3. Artificial Analysis. URL: https://artificialanalysis.ai/guide (дата звернення: 28.02.2025).

4. Chatbot Arena. URL: https://lmarena.ai/ (дата звернення: 28.02.2025).

5. Yakovleva, O., Matúšová S., Talakh V. Gradio and Hugging capabilities for developing research AI applications. *Scientific practice: modern and classical research methods:* Collection of scientific papers «ΛΌГОΣ» with Proceedings of the VII International Scientific and Practical Conference, Boston, February 14, 2025. Boston-Vinnytsia: Primedia eLaunch& UKRLOGOS Group LLC, 2025. P. 202-205. URL: https://doi.org/10.36074/logos-14.02.2025.043.