

**СЕРВІС ГЕНЕРАЦІЇ
СИНТЕТИЧНИХ E-COMMERCE ДАТАСЕТІВ НА ОСНОВІ LLM
ДЛЯ ПРАКТИЧНОГО НАВЧАННЯ ДАТА ІНЖЕНЕРІВ**

Голоха Н.Е. Решетняк А.О.

e-mail: nikita.holokha@nure.ua anton.reshetniak@nure.ua

Науковий керівник – ст. в. Широкопетлєва М.С.

Харківський національний університет радіоелектроніки, каф. ПІ
м. Харків, Україна

This work is devoted to the development of a platform for generating synthetic e-commerce datasets based on large language model simulations for practical training of data engineers. The platform provides access to generated data through a REST API and a webhook-based push system, enabling integration into existing training pipelines. A comparative analysis of existing dataset generation approaches was carried out. It has been found that the proposed LLM-based approach produces statistically realistic data with configurable complexity, making it suitable for practical data engineering training without violating data privacy requirements.

Дата інженерія є однією з найбільш затребуваних спеціальностей у сфері ІТ, попит на фахівців щороку зростає разом із обсягами даних що генеруються бізнесом. Підготовка таких спеціалістів вимагає роботи з реалістичними наборами даних у реальних умовах. Але ключова проблема полягає у критичному розриві між теоретичною базою та можливістю її апробації в умовах реальних промислових навантажень. На сьогодні єдиним із дієвих механізмів трансформації початківця у фахівця залишається залучення до корпоративних програм інтернатури, де студенти мають доступ до реальної інфраструктури. Такий стан справ формує деструктивну дистанцію між академічною підготовкою та індустріальними стандартами, де неможливість повноцінного навчання змушує бізнес йти на компроміс і наймати спеціалістів без необхідного практичного досвіду.

Сучасний інструментарій для підготовки дата-інженерів представлений декількома підходами, кожен із яких має критичні технологічні обмеження, що заважають повноцінній симуляції виробничих процесів. Використання бібліотек синтетичної генерації, як-от Faker чи Mimesis, дозволяє створювати лише статичні структури, позбавлені логічних взаємозв'язків та динамічного контексту поведінки користувачів. Складніші методи на основі генеративних моделей (GAN, VAE) здатні відтворювати статистичні розподіли, проте їхнє застосування замикається у «парадоксі першоджерела», оскільки вони самі потребують величезних масивів реальних даних для навчання [1]. Водночас перспективні LLM-симуляції (наприклад, RecAgent чи Agent4Rec) хоч і наближаються

до моделювання реальних сценаріїв взаємодії, все ще залишаються експериментальними рішеннями, які не забезпечують необхідної масштабованості та цілісності для побудови промислових пайплайнів обробки даних у навчальних цілях [2, 3].

З огляду на зазначене, виникає гостра потреба у створенні інструментарію, який би з'єднував реалізм промислових процесів із доступністю навчальних платформ. Відтак, метою даної роботи є розробка сервісу для генерації синтетичних e-commerce датасетів на основі LLM-симуляції користувацьких сесій. Це дає змогу подолати дефіцит якісних практичних матеріалів і сформувати автентичне середовище для підготовки дата-інженерів, нівелюючи розрив між теоретичним навчанням та вимогами сучасної IT-індустрії

Ключова особливість запропонованого сервісу полягає в підході до симуляції: завдяки використанню Headless-браузерів виконавці ініціюють сесії взаємодії на основі промптів, а LLM-агент, який виступає у ролі віртуального користувача, взаємодіє з реальним фронтом e-commerce застосунку або його mock-інтерфейсом імітуючи рухи мишки квадратичними кривими Безье з 15 – 30 кроками інтерполяції, набір тексту – посимвольний із гаусівським розподілом затримок та імітацією помилок друку з автоматичною корекцією через `backspace` [4]. Використовується агентне моделювання на основі персон, і на підставі цієї інформації приймається рішення про навігацію: які сторінки відвідувати, які товари переглядати, що додавати до кошика та чи оформлювати замовлення. Вебсервер перехоплює реальні події браузера та API-запити, що виникають у процесі такої сесії, і зберігає їх у базі даних. Такий підхід забезпечує природну реалістичність даних, оскільки події породжуються реальною взаємодією з інтерфейсом, а не генеруються штучно. Це дозволяє налаштовувати структуру подій та конкретно формувати дані для ETL/ELT-практики. Студент отримує датасет, що містить реальні виклики: неоднорідну частоту подій, залежності між діями в межах сесії та природні аномалії поведінки.

Доступ до згенерованих даних реалізовано двома способами, що відповідають різним навчальним сценаріям. Перший – через DataAPI, REST-інтерфейс, що дозволяє надсилати запит на отримання події у форматі JSON для подальшої пакетної обробки. Це дозволяє практикувати побудову повноцінних ETL/ELT-пайплайнів засобами Apache Spark або dbt від завантаження сирих даних до формування аналітичних шарів. Другий – система подієвих сповіщень на вебхуки, яка надсилає потік подій на вказану адресу у реальному часі під час симуляції. Такий режим відтворює умови роботи з виробничими стрімінговими системами та дозволяє практикувати обробку даних засобами Apache Kafka [5] або аналогічних платформ. Поєднання обох способів

доступу охоплює два ключові напрямки підготовки дата інженера – пакетну та потокову обробку даних.

Запропонований сервіс заснований на масштабованій архітектурі, яка складається із 8 Docker-сервісів: Webserver (FastAPI та React.ts), який поєднує в собі фронтенд для керування сервісом та бекенд, PostgreSQL як реляційна база для зберігання даних сервісу, Celery Beat та Celery Worker, для оркестрації та масштабування запитів до LLM, Redis для Celery, Ollama (LLM-сервер), Ollama Keep-Alive (допоміжний процес для працездатності Ollama), та фронтенд з яким взаємодіє LLM.

Запропоновано метод агентної симуляції поведінкових даних на основі LLM-персон, який, на відміну від статичної генерації чи статистичних моделей, дозволяє відтворювати складні каузальні зв'язки в користувацьких сесіях. Це забезпечує формування датасетів з високим рівнем логічної цілісності, що критично для тестування систем аналітики. На відміну від статичних файлових сховищ, запропонований підхід дозволяє одночасно моделювати як пакетну обробку (DataAPI), так і реагування у реальному часі (Webhook-стримінг), що забезпечує комплексне покриття компетенцій спеціаліста відповідно до концепції Modern Data Stack.

Список використаних джерел:

1. Jordon J., Szpruch L., Houssiau F. et al. Synthetic Data – what, why and how? // arXiv preprint arXiv:2205.03257. 2022. DOI: <https://doi.org/10.48550/arXiv.2205.03257>
2. Wang L., Zhang J., Yang H. et al. User Behavior Simulation with Large Language Model based Agents // arXiv preprint arXiv:2306.02552. 2024. DOI: <https://doi.org/10.48550/arXiv.2306.02552>
3. An Zhang, Yuxin Chen, Leheng Sheng et al. On Generative Agents in Recommendation // arXiv preprint arXiv:2310.10108. 2023. DOI: <https://doi.org/10.48550/arXiv.2310.10108>
4. Park J. S., O'Brien J. C., Cai C. J. et al. Generative Agents: Interactive Simulacra of Human Behavior // ACM Transactions on Computer-Human Interaction. 2023. Vol. 28, no. 4. P. 1–22. DOI: <https://doi.org/10.1145/3586183.3606763>
5. Apukhtin V., Shirokopetleva M., Skovorodnikova V. The Relevance of Using Message Brokers in Robust Enterprise Applications. 2019 IEEE International Scientific-Practical Conference Problems of Infocommunications, Science and Technology (PIC S&T), Kyiv, Ukraine, 8–11 October 2019. 2019. DOI: <https://doi.org/10.1109/picst47496.2019.9061224>