УДК 004.7

Башманівський М. О., магістрант,
Чижмотря О. В., ст. викладач,
Дмитренко І. А., ст. викладач
Державний університет «Житомирська політехніка»

АРХІТЕКТУРА FULL-STACK ЗАСТОСУНКУ ДЛЯ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ REDDIT З ВИКОРИСТАННЯМ ЛОКАЛЬНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Стрімкий розвиток великих мовних моделей (LLM) відкриває нові можливості для обробки неструктурованого контенту, однак покладання на комерційні хмарні API створює суттєві недоліки, зокрема високу вартість, затримки та ризики конфіденційності. Ці обмеження формують нагальну проблему створення архітектур, здатних безпечно та автономно інтегрувати локально розгорнуті LLM, що особливо актуально для аналізу динамічних платформ, як-от Reddit. Метою даної роботи є представлення багаторівневої архітектури full-stack застосунку, розробленого для інтелектуального аналізу контенту Reddit. Запропонована система реалізує повний цикл оброблення даних — від ініціації запиту та збору до аналітичної інтерпретації та візуалізації, — спираючись на локально розгорнуту велику мовну модель. Такий підхід дозволяє мінімізувати залежність від зовнішніх сервісів і забезпечує контроль над усім процесом обробки даних, що є критично важливим у контексті захисту інформації. Крім того, він відкриває можливості для адаптації моделі під специфіку конкретного домену або завдання.

Для представлення фізичної архітектури та взаємодії компонентів системи була розроблена діаграма розгортання (рис. 1). Вона ілюструє розміщення програмних артефактів на фізичних або віртуальних вузлах. На самому WebServer одночасно розгорнуто декілька ключових процесів, що забезпечують його функціональність. Середовище виконання Node.js / Express API реалізує всю бізнес-логіку, обробку запитів, інтеграцію з API та оркестрацію аналізу. Поруч із ним функціонує локальний AI-сервер LM Studio, що надає доступ до моделі openai/gpt-oss-20b. Важливо, що зв'язок з ним відбувається через локальний інтерфейс http://127.0.0.1:1234, гарантуючи, що дані аналізу не залишають межі сервера. Також на сервері розгорнуто in-memory сховище Redis, яке використовується для кешування проміжних результатів NLP-обробки та керування чергами завдань.

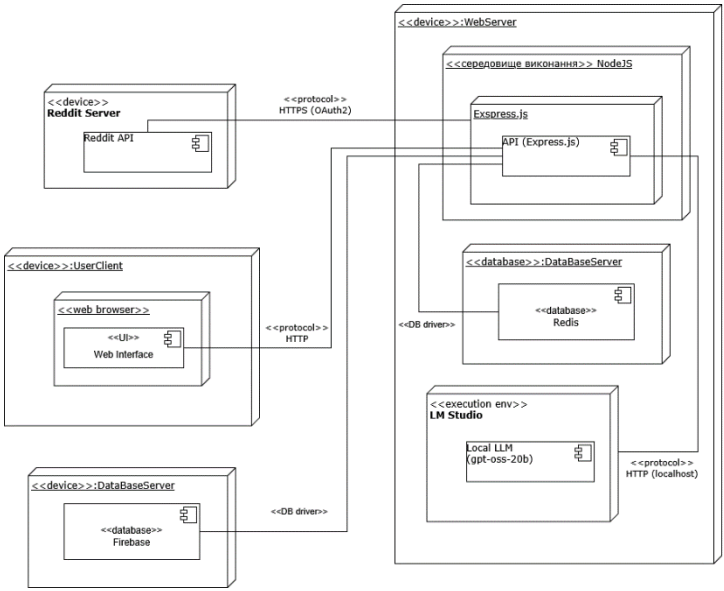

Рисунок 1 – Діаграма розгортання застосунку для інтелектуального аналізу

Зовнішній вузол Firebase Cloud надає два критичні артефакти: Firebase Authentication для безпечної автентифікації користувачів та Firestore для зберігання результатів аналізу. Інший зовнішній вузол, Reddit API, виступає джерелом даних, до якого сервер звертається через протокол OAuth2.

Запропонована багаторівнева архітектура є гнучким рішенням для інтеграції локальних LLM в full-stack застосунки. Ключова перевага полягає у використанні локального AI-двигуна, що гарантує повну конфіденційність даних на відміну від хмарних API [1]. Серверний рівень на Node.js виступає оркестратором повного циклу обробки, від збору даних з Reddit API до збереження результатів у Firebase. Розроблений підхід є життєздатною, безпечною та відмовостійкою моделлю для інтелектуальних систем, орієнтованих на приватність.

Список використаних джерел:

1. Zeng A., Liu J., Zhang H., et al. LLM Application Architectures: A Survey and Design Pattern Analysis. *arXiv preprint,* arXiv:2404.02852, 2024. URL: https://arxiv.org/abs/2404.02852