УДК 004.048
А.Луцків канд. техн. наук, С. Андруньків
Тернопільський національний технічний університет імені Івана Пулюя

РОЗРОБКА СИСТЕМИ МОНІТОРИНГУ СЕМАНТИЧНОЇ ЯКОСТІ ТА ГАЛЮЦИНАЦІЙ ДЛЯ МОДЕЛЕЙ LLM В MLOPS

Andriy Lutskiv PhD., Assoc. Prof., Serhii Andrunkiv

DEVELOPMENT OF A SYSTEM FOR MONITORING SEMANTIC QUALITY AND HALLUCINATIONS FOR LLM MODELS IN MLOPS

Однією із ключових проблем MLOps для великих мовних моделей є відсутність надійних засобів автоматизованого моніторингу семантичної якості відповідей. На відміну від класичних моделей машинного навчання, де моніторинг фокусується на дрифті даних та зміні точності, LLM схильні до так званих "галюцинацій", при яких генерується фактологічно неправдива інформація, а також семантичної деградації, при якій проявляється токсичність та втрата тональності. Така поведінка створює значні ризики при їхньому використанні у виробничому середовищі.

В основі роботи лежить система (див. рис. 1), інтегрована в MLOps-цикл, що призначена для безперервного моніторингу LLM в режимі реального часу. Архітектурно вона складається з таких взаємопов'язаних компонентів:

– асинхронний логгер, який перехоплює та кешує пари "запит-відповідь" безпосередньо з виробничого сервісу LLM, не впливаючи на затримку для кінцевого користувача;

– підсистема оцінки якості, що асинхронно обробляє збережені дані і складається з детектора галюцинацій для валідації фактологічної коректності відповідей, класифікатора семантичної якості для оцінки токсичності та аналізатора безпеки для виявлення зловмисних запитів;

– експортер метрик, який агрегує результати роботи евалюаторів у кількісні показники, такі як відсоток галюцинацій, середній бал токсичності, після чого експортує їх у часову базу даних.

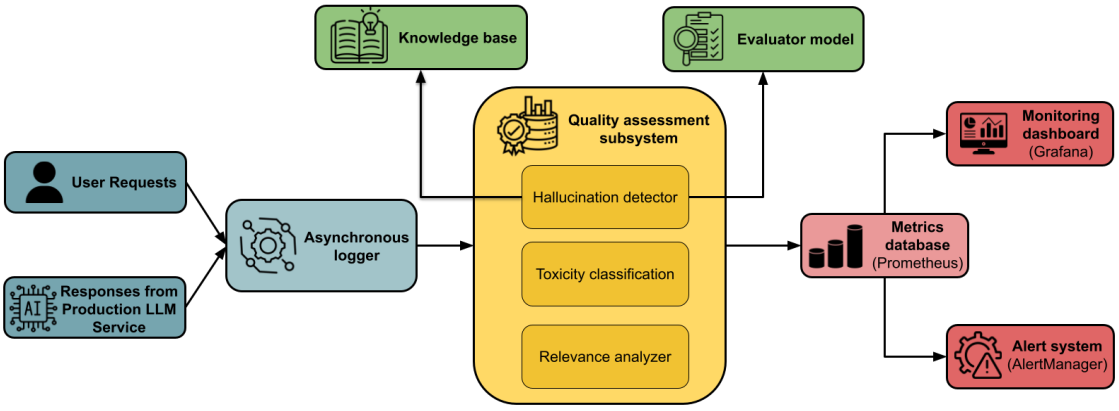

Рис. 1. Архітектурна схема системи моніторингу якості LLM

На основі зібраних даних формуються дашборди, що візуалізують тренди семантичної якості, безпеки та рівня галюцинацій. Система також інтегрована з механізмами сповіщень, які миттєво реагують на раптову деградацію поведінки моделі та виявляють проблеми до того, як вони вплинуть на користувачів.