ПІДВИЩЕННЯ РЕЛЕВАНТНОСТІ ШТУЧНОГО ІНТЕЛЕКТУ В ГАЛУЗІ РАДІОЕЛЕКТРОНІКИ НА ОСНОВІ ТЕХНОЛОГІЇ RAG

Под'ячий Ю. І., Синенко Б. М.
Національний технічний університет «Харківський політехнічний інститут», м.Харків

Сучасний розвиток радіоелектроніки та інтелектуальних систем взаємодії вимагає впровадження новітніх технологій для забезпечення високої продуктивності, точності та надійності роботи різноманітних пристроїв. Одним із ключових напрямів, який активно досліджується в галузі радіоелектроніки, є застосування штучного інтелекту (ШІ) для підвищення релевантності та ефективності функціонування систем.

Для вирішення поставленої перед ШІ проблеми використовуються великі мовні моделі (Large Language Models, LLM), які налічують у своєму складі мільярди параметрів. Оскільки LLM не мають можливості постійного оновлення, їхні знання "заморожуються" на момент завершення їх "навчання". Це призводить до того, що моделі можуть надавати застарілі або неточні відповіді на запити, які потребують актуальної інформації. Це особливо критично в таких галузях, які бурхливо розвиваються, зокрема в радіоелектроніці.

Щоб розширити інформаційну базу, технологія RAG (Retrieval Augmented Generation – генерація доповненим пошуком) пропонує низку ефективних рішень. Одне з них є механізм пошуку релевантної інформації в широко доступних зовнішніх джерелах – енциклопедіях, підручниках, монографіях та ін. Це сприяє суттєвому підвищенню точності і релевантності відповідей, особливо для задач, що вимагають фактичної або спеціалізованої інформації.

Основна ідея представленої розробки полягає у створенні чат-боту, який відшукує в мережі і працює з PDF-документами як джерелами корисної інформації. Застосунок автоматично індексує завантажені документи, зберігає їх у векторному сховищі, а також дозволяє користувачеві отримувати точні й контекстуально релевантні відповіді на основі його запитів. Цей підхід може суттєво скоротити час, витрачений на вирішення поставленої перед ШІ проблеми, зменшити помилки, пов'язані з людським фактором, і звільнити ресурси для виконання більш складних завдань.

Запропонований в цій роботі чат-бот реалізовано на платформі Java із використанням універсального фреймворка Spring Framework, векторної бази даних Milvus та великої мовної моделі GPT API, розробленої OpenAI. Для інтеграції цих інструментів створене відкрите рішення YDA Framework, яке дозволяє інкапсулювати функціонал RAG і значно спрощує розробку подібних застосунків. У межах YDA всі компоненти – індексація, пошук, ранжування, генерація та обробка запитів – інтегровані в модульну структуру, що забезпечує гнучкість і масштабованість.

Наведено схему архітектури розробленого застосунку з детальними поясненнями і демонструється приклад його роботи. Також надається порівняльна таблиця можливостей застосунку з існуючими рішеннями.