DOI https://doi.org/10.30525/978-9934-26-506-8-115

USE OF RAG SYSTEMS TO IMPROVE THE ACCURACY AND CONTENT OF QUERY RESULTS FOR LARGE LANGUAGE MODELS

ВИКОРИСТАННЯ RAG-СИСТЕМ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ ТА ЗМІСТОВНОСТІ РЕЗУЛЬТАТІВ ЗАПИТІВ ДО ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Plutalov Ya.A.,
Student (group 122-24-1м), LLC "Technical university "Metinvest polytechnic", Zaporizhzhia, Ukraine

Плуталов Я.А.,
студент гр. 122-24-1м, ТОВ «Технічний університет «Метінвест політехніка», м. Запоріжжя, Україна

Sahaida P.I.,
DSc (Engineering), Associate Professor, LLC "Technical university "Metinvest polytechnic", Zaporizhzhia, Ukraine

Сагайда П.І.,
д.т.н., доцент, ТОВ «Технічний університет «Метінвест політехніка», м. Запоріжжя, Україна

Поточні дослідження підходу Retrieval-Augmented Generation (RAG) спрямовані на покращення продуктивності великих мовних моделей (ВММ) через інтеграцію компонентів інформаційного пошуку на основі документів та ресурсів предметної області або підприємства (організації), для інформаційної підтримки фахівців якої призначена відповідна система. Зазвичай ВММ генерують відповіді на основі навчальних даних, але з RAG вони можуть звертатися до зовнішніх джерел для отримання актуальної або специфічної інформації, що підвищує точність і змістовність генерації. У цій роботі висвітлюються ключові переваги цієї технології, зокрема гнучкість та розширення знань моделі, особливо в контексті швидко мінливих або специфічних сфер, в тому числі для створення інтелектуального асистента інформаційного порталу Технічного університету «Метінвест Політехніка».

Робочий процес RAG складається з трьох етапів – індексування, пошук і генерація, які працюють разом, щоб підвищити точність відповіді. На етапі індексування текст кодується у вставки, які зберігаються у векторній базі даних із можливістю пошуку. Під час пошуку запит користувача також кодується, щоб знайти найбільш релевантні документи в цій базі даних. На етапі генерації модель поєднує запит користувача зі знайденими документами, що дає змогу системі генерувати більш точні результати. З використанням описаного вище структурованого підходу ВММ надає відповіді, які спираються на сучасні, релевантні джерела інформації.

Спосіб обробки даних, застосовуючи RAG системи, сприяє наданню більш чітких відповідей користувачеві та детальнішою роботою з певною документацією або необхідним джерелом даних, а не більш загальними знаннями ВММ. Сфера застосування RAG розширюється до мультимодальних галузей, адаптуючи його принципи до інтерпретації та обробки різноманітних форм даних, як-от зображення, відео та код. Подібне розширення підкреслює значне практичне значення RAG для розгортання штучного інтелекту, що може стати серйозним поштовхом для академічного та промислового секторів.

Попри значні переваги, RAG постає перед викликами: проблеми з ефективністю під час обробки великих наборів даних, залежність від якості зовнішніх джерел, а також ризики «галюцинацій» – ситуацій, коли модель може неправильно інтерпретувати або змішувати знайдену інформацію.

Використання RAG з інтеграцією зовнішніх джерел інформації підвищує точність та релевантність відповідей, надаючи можливість отримувати сучасні знання, що особливо корисно для напрямів з інтенсивним розвитком та унікальними вимогами. Проте виникає необхідність в подальших дослідженнях щодо підвищення ефективності генерування результатів та покращення якості джерела пошуку.

Перспективним напрямком даного дослідження є використання підходу RAG до створення інтелектуального асистента інформаційного порталу Технічного університету «Метінвест Політехніка», який на основі корпусу нормативних документів Університету буде формувати відповіді на питання студентів, викладачів та абітурієнтів по особливостям навчального процесу, наукових досліджень в Університеті, іншим аспектам функціонування закладу вищої освіти. При цьому відповіді повинні підкріплюватися посиланнями на релевантні абзаци або сторінки відповідних документів.

Висновки. Основна ідея RAG полягає у поєднанні великих мовних моделей з зовнішніми джерелами інформації для покращення якості відповідей. Організація таких систем повинна забезпечити виконання трьох основних етапів: індексацію документів підприємства або організації, пошук релевантних фрагментів тексту у відповідності до запиту користувача, генерацію відповідей з використанням великих мовних моделей. Превагами використання систем RAG для інформаційної підтримки користувачів є отримання більш точних, актуальних та деталізованих відповідей, які базуються на корпусі документів предметної області. Проблемами та викликами залишаються недостатня продуктивність таких систем, залежність від якості даних, що використовуються під час пошуку, ризики "галюцинацій". Перспективним напрямком даного дослідження є використання підходу RAG до створення інтелектуального асистента інформаційного порталу Технічного університету «Метінвест Політехніка».

Перелік використаних джерел

1. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Wang, H. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv.org e-Print archive. URL: https://arxiv.org/html/2312.10997v5#S2 (дата звернення: 25.10.2024).

2. Demir N. Hands-On with RAG: Step-by-Step Guide to Integrating Retrieval Augmented Generation in LLMs. Medium. URL: https://blog.demir.io/hands-on-with-rag-step-by-step-guide-to-integrating-retrieval-augmented-generation-in-llms-ac3cb075ab6f (дата звернення: 25.10.2024).

DOI https://doi.org/10.30525/978-9934-26-506-8-116

FUNDAMENTAL PRINCIPLES FOR DEVELOPING THE GEODATA ZONES SYSTEM FOR MONITORING THE STABILITY OF DEEP QUARRY SLOPES

ОСНОВНІ ЗАСАДИ РОЗРОБКИ СИСТЕМИ GEODATA ZONES ДЛЯ МОНІТОРИНГУ СТАНУ БОРТІВ ГЛИБОКИХ КАР'ЄРІВ

Romanenko A.O.,
PhD, Student (group 122-23-1M), LLC "Metinvest Polytechnic Technical University," Zaporizhzhia, Ukraine

Романенко А.О.,
к.т.н., студент гр. 122-23-1М, ТОВ «Технічний університет «Метінвест політехніка», м. Запоріжжя, Україна

Вступ. Промислові об'єкти, такі як кар'єри, вимагають постійного моніторингу стійкості, оскільки зміни в геологічних умовах або порушення цілісності бортів можуть призвести до небезпечних ситуацій, включаючи обвали та зсуви. Забезпечення стабільності гірничих робіт є ключовим фактором для безпеки працівників і збереження обладнання, а також для мінімізації негативного впливу на навколишнє середовище.

Система "GeoDATA zones" призначена для інтеграції даних з різних джерел, включаючи маркшейдерські і геофізичні вимірювання, щоб забезпечити комплексний підхід до аналізу стійкості бортів. Використання таких даних дозволяє отримувати актуальну інформацію про стан гірничих масивів у реальному часі, оперативно реагувати на зміни і вживати заходів для запобігання аварійним ситуаціям. Крім того, система інтегрує історичні дані, що дозволяє проводити довгостроковий аналіз та прогнозування ризиків, визначаючи зони підвищеної небезпеки.