УДК 004.8:004.9

ВПЛИВ МЕХАНІЗМУ RAG НА ЯКІСТЬ ВІДПОВІДЕЙ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ У СИСТЕМАХ АВТОМАТИЗОВАНОЇ ПІДТРИМКИ КОРИСТУВАЧІВ

Косяк М.В.
Національний університет «Одеська юридична академія»
Україна, Одеса
mykolakosiakv@gmail.com

Анотація. *У роботі досліджено вплив механізму генерації, доповненої пошуком (RAG), на якість відповідей великих мовних моделей у системах автоматизованої підтримки користувачів. Розглянуто взаємозв'язок між релевантністю знайдених фрагментів бази знань і повнотою та достовірністю сформованих відповідей. Запропоновано порівняння трьох режимів генерації: без контексту, з еталонним фрагментом і зі знайденим фрагментом. Для оцінювання передбачено використання показника hit@k, повноти обов'язкових фактів та інтегральної оцінки якості відповіді.*
Ключові слова: *великі мовні моделі; генерація, доповнена пошуком; RAG; база знань; автоматизація підтримки користувачів; якість відповідей.*

Вступ. Автоматизація відповідей на типові запити користувачів за допомогою великих мовних моделей (Large Language Models, LLM) дає змогу формувати природномовні відповіді на основі запитів, сформульованих у довільній формі. Однак використання таких моделей у системах підтримки користувачів пов'язане з проблемою достовірності інформації. Знання, закладені в параметрах моделі під час навчання, не обов'язково відповідають актуальним правилам, тарифам, процедурам та умовам обслуговування конкретної організації. Крім того, модель може генерувати правдоподібні, але фактично необґрунтовані твердження, що є особливо небажаним у фінансовій, телекомунікаційній та інших предметних сферах, де помилкова відповідь може призвести до неправильних дій користувача.

Одним із підходів до зменшення зазначених ризиків є генерація, доповнена пошуком (Retrieval-Augmented Generation, RAG). Її принцип полягає в тому, що перед формуванням відповіді система знаходить у зовнішньому корпусі знань фрагменти, релевантні запиту, та передає їх мовній моделі як додатковий контекст [1, 2]. Таким чином, модель отримує можливість спиратися не лише на параметричні знання, а й на інформацію з визначеного джерела, яку можна оновлювати незалежно від самої моделі.

Водночас наявність механізму пошуку не гарантує правильності кінцевої відповіді. Якщо пошуковий компонент не знаходить потрібного фрагмента, повертає інформацію з подібною тематикою, але іншого змісту, або передає надмірний за обсягом контекст, модель може сформувати неповну чи помилкову відповідь. Навіть за умови отримання правильного фрагмента залишається ризик його некоректного інтерпретування або доповнення твердженнями, яких немає у джерелі. Отже, ефективність RAG необхідно оцінювати не лише за точністю пошуку, а й за тим, наскільки знайдений контекст впливає на фактичну повноту та достовірність згенерованої відповіді

Метою роботи є визначення впливу механізму пошуку зовнішніх знань на якість відповідей великих мовних моделей у системах автоматизованої підтримки користувачів. Для досягнення поставленої мети необхідно: проаналізувати основні етапи взаємодії пошукового та генеративного компонентів RAG; сформувати умови порівняння відповідей за різних ре-

жимів використання контексту; визначити показники оцінювання якості пошуку та сформованих відповідей; передбачити розмежування помилок, спричинених недостатньою релевантністю знайденої інформації, і помилок її опрацювання мовною моделлю..

Основна частина. У системі автоматизованої підтримки користувачів механізм RAG доцільно розглядати як послідовність двох взаємопов'язаних процесів: пошуку інформації в зовнішній базі знань та генерації відповіді на основі отриманих фрагментів [1, 3]. На першому етапі система аналізує текст запиту й визначає фрагменти корпусу знань, які мають найбільшу відповідність його змісту. На другому етапі відібрані фрагменти включаються до контексту мовної моделі разом із запитом та інструкціями щодо формування відповіді.

Якість роботи такої системи залежить від кожного з цих етапів. Пошуковий компонент має забезпечити виявлення потрібної інформації та її належне ранжування. Генеративний компонент, своєю чергою, повинен використати знайдені відомості без втрати важливих умов, числових значень і обмежень. Якщо хоча б один із компонентів працює неналежним чином, це може погіршити кінцеву відповідь. Наприклад, відсутність у контексті конкретного строку виконання операції може призвести до неповної відповіді, навіть якщо мовна модель правильно опрацювала всі передані їй відомості.

Для дослідження впливу контексту доцільно використовувати три режими формування відповіді. Перший режим передбачає генерацію без зовнішнього контексту: модель отримує лише запит і формує відповідь на основі власних параметричних знань. Він використовується як базовий рівень для визначення результату без залучення пошукового компонента. Другий режим передбачає передавання моделі еталонного фрагмента бази знань, який містить необхідну для відповіді інформацію. Такий режим дає змогу оцінити якість генерації за умови надання правильного контексту та фактично визначає орієнтир для результату за ідеального пошуку. Третій режим передбачає використання фрагмента, фактично знайденого пошуковим компонентом. Він відображає роботу повного конвеєра RAG і дає змогу оцінити вплив реальної якості пошуку на кінцеву відповідь.

Порівняння зазначених режимів має важливе методичне значення. Різниця між відповідями без контексту та з еталонним фрагментом характеризує потенційний вплив зовнішніх знань на результат генерації. Відмінність між відповідями з еталонним і знайденим фрагментами дає змогу оцінити втрати якості, пов'язані з роботою пошуку. Якщо відповідь із еталонним контекстом є якісною, а зі знайденим – ні, це може свідчити про недостатню релевантність або повноту результатів пошуку. Якщо ж обидві відповіді мають низькі оцінки, проблема може бути пов'язана з інтерпретацією контексту або способом формування відповіді самою моделлю.

Для реалізації пошуку може використовуватися алгоритм BM25, який ранжує фрагменти корпусу за відповідністю термінів запиту тексту документа з урахуванням частоти термінів та їхньої інформативності в корпусі. Такий підхід дає змогу отримувати впорядкований список фрагментів без попереднього навчання окремої пошукової моделі. Для оцінювання результатів пошуку доцільно застосовувати показник hit@k, який визначає частку запитів, для яких хоча б один правильний фрагмент потрапив до перших k позицій результатів:

$$\text{hit@k} = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{1}\left[G(q) \cap R_k(q) \neq \emptyset\right]$$

де: Q – множина тестових запитів; $G(q)$ – множина правильних фрагментів для запиту q; $R_k(q)$ – множина перших k фрагментів, знайдених пошуковою системою; $\mathbb{1}[\cdot]$ – індикаторна функція, що дорівнює 1, якщо умова виконується, і 0 – інакше. У межах дослідження доціль-

но розглядати hit@1 і hit@3, що дасть змогу оцінити як точність вибору першого результату, так і здатність пошуку включити правильний фрагмент до невеликого набору кандидатів.

Водночас показник hit@k характеризує лише результат пошукового етапу. Він не дає відповіді на питання, чи використала мовна модель знайдену інформацію належним чином. Тому оцінювання кінцевої відповіді має здійснюватися окремо. Для цього можна використати повноту обов'язкових фактів (Fact Coverage, FC), яка визначає частку передбачених еталоном фактів, відображених у відповіді:

$$FC(q) = \frac{|F_{\text{found}}(q)|}{|F(q)|} \cdot 100\%$$

де $F(q)$ – множина обов'язкових фактів для запиту, а $F_{\text{found}}(q)$ – множина фактів, які фактично присутні у сформованій відповіді. Такий показник дає змогу перевірити, чи не втрачено під час генерації важливі умови, строки, числові значення та інші відомості, необхідні для повноцінного розв'язання запиту.

Для комплексного порівняння відповідей також може застосовуватися інтегральна оцінка Q_{det}, яка поєднує повноту фактів, лексичну близькість до еталонної відповіді та штраф за виявлені ознаки галюцинацій. Її використання дає змогу враховувати не лише наявність необхідної інформації, а й відповідність сформульованої відповіді еталону та наявність потенційно необґрунтованих тверджень. При цьому автоматично виявлені ознаки галюцинацій слід трактувати як індикатор ризику, а не як вичерпне підтвердження фактичної неправильності відповіді.

Важливо, що оцінки пошукового та генеративного компонентів не слід підміняти одна одною. Високе значення hit@k не гарантує високої якості відповіді, оскільки модель може неповно використати навіть правильний фрагмент. Аналогічно, якісна відповідь без контексту не доводить ефективності RAG, адже модель могла сформувати її на основі власних знань. Саме тому результати трьох режимів необхідно зіставляти за однаковим набором запитів і єдиними правилами оцінювання.

Для забезпечення коректності порівняння в експерименті потрібно зберігати незмінними модель, параметри генерації, інструкції промпту та склад тестового набору. Змінною має бути саме режим подання контексту. Окремо слід фіксувати час і вартість обробки запиту, оскільки використання пошукового компонента та передавання додаткових фрагментів збільшує обсяг операцій і може впливати на затримку відповіді. Таким чином, оцінювання RAG має враховувати не лише приріст якості, а й ресурсні витрати, необхідні для його досягнення.

Особливу увагу необхідно приділити складу корпусу знань. Якщо кожному запиту відповідає лише один правильний фрагмент, а тематично схожі, але нерелевантні документи відсутні, пошуковий компонент працює в надто спрощених умовах. У такому разі показник hit@k може бути завищеним і не відображати складності пошуку в реальній базі знань служби підтримки. Для наближення експерименту до практичних умов корпус доцільно доповнювати дистракторами – фрагментами, які тематично або лексично подібні до правильних, але не містять потрібної відповіді. Додатково необхідно використовувати перефразовані запити, щоб перевірити стійкість пошуку до відмінностей у формулюваннях користувачів. Згенеровані автоматично дистрактори й перефрази мають проходити перевірку перед включенням до експериментального набору.

Якщо потрібна інформація відсутня в корпусі або пошук не знаходить достатньо релевантного фрагмента, система не повинна автоматично формувати впевнену відповідь на основі непідтверджених припущень. У таких випадках доцільно передбачити відмову від авто-

матичної відповіді або передавання запиту оператору. Це особливо важливо для запитів, де помилкова інформація може мати суттєві наслідки для користувача. Отже, RAG доцільно розглядати не як самостійний механізм гарантування достовірності, а як компонент конвеєра автоматизованої підтримки, ефективність якого залежить від якості корпусу знань, пошуку, генерації та правил обробки невизначених ситуацій.

Висновки. Розглянуто особливості використання механізму RAG для автоматизації відповідей на типові запити користувачів. Обґрунтовано, що вплив зовнішнього контексту на якість відповіді визначається не лише здатністю пошукового компонента знаходити релевантні фрагменти, а й тим, наскільки повно та коректно мовна модель використовує отриману інформацію. Запропоновано порівнювати три режими генерації: без контексту, з еталонним фрагментом і зі знайденим фрагментом. Такий підхід дає змогу розмежувати потенційні можливості генерації за наявності правильних знань і втрати якості, зумовлені недоліками пошуку.

Визначено доцільність спільного використання показника hit@k для оцінювання пошукового компонента та метрик повноти фактів і комплексної якості для оцінювання сформованих відповідей. Показано, що високі результати пошуку самі по собі не є достатньою умовою достовірної відповіді, а застосування RAG не усуває потреби в контролі необґрунтованих тверджень і механізмі передавання запитів оператору. Практична цінність запропонованого підходу полягає в можливості визначати, на якому етапі конвеєра виникають помилки, та обґрунтовано обирати напрями вдосконалення систем автоматизованої підтримки користувачів.

Використання інструментів ШІ. Під час підготовки тексту тез використано ChatGPT (GPT-6) для граматичного редагування матеріалу, формулювання анотацій.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Lewis, P., Perez, E., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems, 33*. https://papers.nips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[2] Gao Y. et al. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv, 2312.10997*. https://doi.org/10.48550/arXiv.2312.10997

[3] Es, S., James, J., Espinosa-Anke, L., Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 150–158*. https://doi.org/10.18653/v1/2024.eacl-demo.16

Kosiak M.V.

The impact of retrieval-augmented generation on the quality of large language model responses in automated user support systems

Abstract. *The paper examines the impact of Retrieval-Augmented Generation (RAG) on the quality of large language model responses in automated user support systems. The relationship between the relevance of retrieved knowledge base passages and the completeness and reliability of generated responses is considered. Three generation modes are proposed for comparison: without context, with a reference passage, and with a retrieved passage. The evaluation approach includes the hit@k metric, fact coverage, and an integrated response quality score.*

Keywords*: large language models; retrieval-augmented generation; RAG; knowledge base; automated user support; response quality.*