УДК 004.8:378.147

РЕТРІВАЛ-АУГМЕНТОВАНА ГЕНЕРАЦІЯ ЯК ОСНОВА ІНТЕЛЕКТУАЛЬНОЇ СИСТЕМИ ДОВІДКОВОЇ ПІДТРИМКИ СТУДЕНТІВ

Клименко Д. В.
Національний університет «Одеська політехніка»
Україна, Одеса
7942968@stud.op.edu.ua

Анотація. *Розглянуто проблему достовірності відповідей великих мовних моделей у системах довідкової підтримки студентів. Запропоновано архітектуру на основі ретрівал-аугментованої генерації (RAG) з гібридним пошуком, обов'язковим цитуванням першоджерел та відмовою від відповіді за браку підтверджувальних фрагментів. Визначено заходи зниження ризиків і метрики оцінювання якості.*
Ключові слова: *ретрівал-аугментована генерація; велика мовна модель; гібридний пошук; галюцинації; довідкова система.*

Вступ. Студентам постійно потрібні оперативні відповіді щодо розкладу, вимог силабусів, порядку оформлення документів та інших нормативних положень. Ці відомості розпорошені між сайтами, файлами й месенджерами, тому пошук потрібного забирає час. Чатботи на основі великих мовних моделей (LLM) могли б розв'язати цю проблему, проте такі моделі схильні генерувати правдоподібні, але недостовірні твердження (галюцинувати), що знижує якість системи та довіру до неї [1]. Для довідкових систем, де хибна відповідь про строки чи правила має практичні наслідки, це особливо критично.

Один зі способів зменшення цього ризику – ретрівал-аугментована генерація (Retrieval-Augmented Generation, RAG), яка поєднує параметричну пам'ять попередньо навченої моделі та непараметричну пам'ять у вигляді індексу зовнішніх документів [2]. В огляді [3] розрізнено три етапи розвитку підходу – Naive RAG, Advanced RAG і Modular RAG – та три основні складники: пошукач, генератор і методи аугментації. Водночас базові реалізації залишають відкритими питання якості пошуку та узгодженості відповіді з наданим контекстом [3].

Мета роботи – обґрунтувати архітектуру RAG-системи довідкової підтримки студентів, яка зменшує ризик недостовірних відповідей і дає змогу оновлювати знання без перенавчання моделі.

Альтернативою є донавчання моделі на корпусі документів. Однак воно потребує повторного навчання після кожної зміни нормативних положень і не дає прямого посилання на джерело відповіді, тому для документів, що часто оновлюються, доцільнішим є підхід RAG [2], [3]. Для досягнення мети поставлено такі завдання: визначити склад і взаємодію модулів системи; запропонувати спосіб поєднання семантичного та лексичного пошуку; описати заходи зниження ключових ризиків; сформувати методику оцінювання якості.

Основна частина. Запропоновано архітектуру з чотирьох модулів. *Модуль індексації* збирає документи університету (положення, силабуси, розклади), очищає їх, ділить на фрагменти з перекриттям і зберігає разом із метаданими (джерело, дата, версія документа, підрозділ), після чого обчислює векторні представлення та будує індекс. *Модуль пошуку* для запиту q оцінює релевантність фрагмента d як зважену суму семантичної та лексичної оцінок:

$$s(q, d) = \alpha \cdot s_{dense}(q, d) + (1 - \alpha) \cdot s_{lex}(q, d) \qquad (1)$$

де s_{dense} – косинусна близькість векторних представлень; s_{lex} – лексична оцінка (наприклад, BM25); $\alpha \in [0; 1]$ – коефіцієнт, що підбирається на валідаційній вибірці. Лексична складова потрібна, бо запити студентів часто містять точні назви дисциплін, номери наказів і дати, які погано передаються лише семантичною близькістю. Кандидатів фільтрують за метаданими (актуальна версія документа) і переупорядковують окремою моделлю повторного ранжування.

Якість пошуку значною мірою залежить від способу розбиття документів. Надто короткі фрагменти втрачають контекст (наприклад, умови винятку з правила), надто довгі – містять сторонню інформацію. Тому фрагменти формуються за структурою документа (розділ, пункт, стаття), а за потреби довгі пункти ділять із частковим перекриттям. Початкові значення довжини фрагмента та перекриття розглядаються як параметри, що добираються експериментально на валідаційній вибірці.

Модуль генерації передає LLM лише відібрані фрагменти та інструкцію: відповідати виключно на їхній основі, вказувати джерело кожного твердження і повідомляти про відсутність відомостей, якщо контекст недостатній. *Модуль контролю* веде журнал запитів, збирає оцінки користувачів і запускає повторне індексування при зміні документа. Відповідність ризиків і заходів наведено в табл. 1.

Обробка запиту виконується послідовно: (1) нормалізація запиту та визначення мови; (2) гібридний пошук і фільтрація за метаданими; (3) повторне ранжування та відбір обмеженої кількості фрагментів; (4) формування запиту до LLM з інструкцією та відібраним контекстом; (5) перевірка відповіді на наявність посилань на надані фрагменти; (6) видача відповіді користувачу разом із переліком джерел. Якщо після кроку (3) жоден фрагмент не перевищує встановленого порога релевантності, система не звертається до моделі, а повідомляє, що відомостей недостатньо, і пропонує звернутися до відповідного підрозділу. Це знижує ризик заповнення прогалин непідтвердженими припущеннями моделі [1].

Таблиця 1 – Ризики та заходи їх зниження

Ризик	Захід в архітектурі
Галюцинації моделі	Відповідь лише за наданими фрагментами; обов’язкове цитування джерела; відмова за недостатнього контексту
Неактуальна інформація	Версіонування документів; фільтрація за метаданими; автоматичне переіндексування при змінах
Пропуск релевантного фрагмента	Гібридний пошук за формулою (1); повторне ранжування кандидатів
Витік персональних даних	В індекс потрапляють лише відкриті нормативні документи; журнали запитів знеособлюються

Оцінювання проводиться окремо для пошуку та для генерації, оскільки помилка на першому етапі неминуче позначається на другому. Для пошуку використовуються точність і повнота на глибині k:

$$Precision@k = |R_k \cap G| / k, \quad (2)$$

$$Recall@k = |R_k \cap G| / |G|, \quad (3)$$

де R_k – множина перших k знайдених фрагментів; G – множина фрагментів, що містять відповідь за експертною розміткою. Якість генерації оцінюється за вірністю відповіді наданому контексту (частка тверджень, підтверджених фрагментами) та релевантністю відповіді запи-

ту; ці показники узгоджуються з аспектами якості RAG, наведеними в [3]. Склад показників подано в табл. 2.

Таблиця 2 – Показники оцінювання системи

Показник	Що характеризує	Спосіб визначення
Precision@k, Recall@k	Якість пошуку фрагментів	За формулами (2), (3) на тестовому наборі запитів
Вірність контексту	Відсутність тверджень без підтвердження	Частка тверджень відповіді, підтверджених фрагментами; експертна перевірка
Релевантність відповіді	Відповідність відповіді запиту	Експертна оцінка за шкалою
Частка коректних відмов	Поведінка за браку відомостей	Частка запитів без відповіді в корпусі, на які система відмовила

Експериментальна перевірка передбачає формування тестового набору запитів, типових для студентів (розклад, оцінювання, академічна відпустка, поновлення, оформлення документів), із еталонними відповідями, підготовленими експертами. Порівнюються три конфігурації: LLM без пошуку; базовий (Naive) RAG; запропонована архітектура з гібридним пошуком і повторним ранжуванням. Окремо досліджується вплив коефіцієнта α у формулі (1) і розміру фрагментів на показники пошуку.

Варто врахувати обмеження підходу. Якість відповіді залежить від повноти та коректності самих документів: суперечності між версіями положень призводять до суперечливих відповідей, тому потрібні процедури контролю версій. Відповіді системи мають довідковий характер і не замінюють консультацію відповідальної особи. Для персональних запитів доцільно застосовувати локально розгорнуті моделі.

Висновки. Запропоновано архітектуру RAG-системи довідкової підтримки студентів, що поєднує гібридний пошук, фільтрацію за метаданими, повторне ранжування та генерацію з обов'язковим цитуванням джерел. Визначено заходи зниження ключових ризиків і метрики оцінювання якості. Практична цінність полягає в тому, що знання системи оновлюються зміною індексу без перенавчання моделі. Запропоновано методику оцінювання, що розділяє показники пошуку та генерації. Експериментальну перевірку на корпусі документів університету (добір α, розміру фрагментів, порівняння з базовим RAG) заплановано як наступний етап.

Використання інструментів ШІ. Інструмент ШІ Claude 5.5 Sonnet (Anthropic) було використано для аналізу наукових джерел за темою RAG, стилістичного редагування тексту та перевірки бібліографічних описів.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). *Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12),* Article 248. https://doi.org/10.1145/3571730

[2] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). *Retrieval-augmented generation for knowledge-intensive NLP tasks*. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. *33*, pp. 9459–9474). Curran

Associates.
https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf
[3] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Guo, Q., Wang, M., & Wang, H. (2023). *Retrieval-augmented generation for large language models*: A survey. arXiv. https://doi.org/10.48550/arXiv.2312.10997

Klymenko D. V.

Retrieval-augmented generation as the basis of an intelligent student information support system

Abstract. *The paper addresses the reliability of large language model answers in student information support systems. An architecture based on retrieval-augmented generation (RAG) is proposed, featuring hybrid retrieval, mandatory source citation, and refusal to answer when supporting evidence is missing. Risk mitigation measures and quality metrics are defined.*

Keywords: *retrieval-augmented generation; large language model; hybrid retrieval; hallucinations; information support system.*