

Maksym PetrovNational Technical University of Ukraine “Igor
Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine

RETRIEVAL-AUGMENTED GENERATION ARCHITECTURE FOR IMPROVING THE RELIABILITY OF GENERATIVE INFORMATION SYSTEMS

Abstract. The paper substantiates the use of retrieval-augmented generation as an architectural approach for generative information systems operating on local, regulatory, or frequently updated data. The workflow is considered as a sequence of corpus preparation, chunk indexing, semantic retrieval, context construction, and answer generation. A practical set of reliability criteria is proposed, including retrieval relevance, evidence coverage, source traceability, and resistance to stale data. Retrieval augmentation does not automatically eliminate erroneous outputs; however, it provides a technical basis for controlling information provenance and updating knowledge without full model retraining.

Keywords: retrieval-augmented generation, large language models, information systems, semantic search, reliability, artificial intelligence.

Максим ПетровНаціональний технічний університет України «Київський
політехнічний інститут ім. Ігоря Сікорського», Київ, Україна

АРХІТЕКТУРА ПОШУКОВО-ДОПОВНЕНОЇ ГЕНЕРАЦІЇ ДЛЯ ПІДВИЩЕННЯ НАДІЙНОСТІ ГЕНЕРАТИВНИХ ІНФОРМАЦІЙНИХ СИСТЕМ

Анотація. У роботі обґрунтовано доцільність використання пошуково-доповненої генерації як архітектурного підходу до побудови генеративних інформаційних систем, що працюють із локальними, нормативними або часто оновлюваними масивами даних. Розглянуто послідовність обробки запиту: підготовка корпусу, індексація фрагментів, семантичний пошук, формування контексту та генерація відповіді. Запропоновано набір практичних критеріїв оцінювання надійності такої системи: релевантність знайдених фрагментів, повнота доказової бази, простежуваність джерел і стійкість до застарілих даних. Показано, що пошуково-доповнена генерація не усуває ризик помилкових відповідей автоматично, але створює технічну основу для контролю походження інформації та регулярного оновлення знань без повного перенавчання моделі.

Ключові слова: пошуково-доповнена генерація, великі мовні моделі, інформаційні системи, семантичний пошук, надійність, штучний інтелект.

Генеративні інформаційні системи на основі великих мовних моделей швидко переходять від універсальних чат-інтерфейсів до спеціалізованих рішень для освіти, науки, бізнесу та внутрішніх корпоративних сервісів. Ключова проблема такого переходу полягає в тому, що параметричні знання моделі є фіксованими на момент навчання та не гарантують ані актуальності, ані коректного відтворення вузькоспеціалізованих фактів. Для систем, які повинні відповідати за внутрішніми регламентами, методичними матеріалами, нормативними документами або базами знань, цього обмеження достатньо, щоб навіть стилістично переконлива відповідь була непридатною для практичного використання.

Одним із базових способів подолання цієї проблеми є архітектура retrieval-augmented generation, запропонована як поєднання параметричної пам'яті генеративної моделі та зовнішньої непараметричної пам'яті, з якої перед формуванням відповіді добираються релевантні фрагменти (Lewis et al., 2020). У прикладній системі цей принцип доцільно реалізовувати не як просте «прикріплення документа до запиту», а як окремий конвеєр роботи з даними. Документи очищаються, діляться на логічні фрагменти, доповнюються метаданими, перетворюються на векторні подання та індексуються. На етапі запиту система знаходить найближчі за змістом фрагменти, після чого передає їх мовній моделі разом з інструкцією використовувати саме наданий контекст.

Для надійності особливе значення має якість розбиття документів. Надто великі фрагменти містять багато зайвої інформації та погіршують точність пошуку; надто короткі можуть втрачати смислову цілісність. Практичним рішенням є поєднання структурного поділу за заголовками, абзацами або пунктами документа з обмеженням максимальної довжини фрагмента та невеликим перекриттям сусідніх частин. Для нормативних і навчально-методичних документів доцільно зберігати разом із фрагментом назву документа, розділ, дату, версію та посилання на первинне джерело. Такі метадані використовуються не лише для пошуку, а й для відображення доказової бази користувачеві.

Таблиця 1. Контрольні точки пошуково-доповненої генерації

Етап	Основний ризик	Практичний контроль
Підготовка даних	Застарілі або дубльовані документи	Версіонування, дата чинності, дедуплікація
Пошук	Нерелевантні фрагменти	Тестові запити, поріг релевантності, метадані
Генерація	Непідтверджені твердження	Відповідь лише за контекстом, цитування джерел
Експлуатація	Непомітна деградація якості	Журнал запитів, контрольний набір, періодична перевірка

Джерело: складено автором на основі Lewis et al. (2020) та Tabassi (2023).

Найслабшою ланкою системи часто є не генерація, а пошук. Якщо релевантний фрагмент не потрапив у контекст, мовна модель може сформулювати правдоподібну відповідь на основі загальних знань. Тому оцінювання має розділяти щонайменше два рівні: якість retrieval-компонента та якість кінцевої відповіді. Для першого рівня перевіряють, чи потрапляє потрібний документ або фрагмент у перші результати пошуку; для другого — чи відповідає висновок наведеним джерелам і чи не додає модель непідтверджених тверджень. В освітньому середовищі такий поділ особливо важливий, оскільки огляди застосування генеративних моделей вказують одночасно на їхню корисність як навчального асистента і на ризик отримання неправильної або вигаданої інформації (Lo, 2023).

Окремим механізмом контролю є вимога до моделі відмовлятися від відповіді, якщо в отриманому контексті немає достатньої інформації. Це не повинно бути лише текстовою інструкцією. На рівні системи можна встановлювати мінімальний поріг релевантності пошуку, вимагати наявності двох незалежних фрагментів для критичних тверджень або повертати користувачеві список джерел замість синтезованої відповіді. Для документів, що часто змінюються, важливо також урахувувати дату та статус версії: архівний наказ чи застарілий регламент не повинен мати таку саму вагу, як чинний документ.

Підхід до управління ризиками доцільно узгоджувати з загальною логікою NIST AI Risk Management Framework, де робота з ризиками штучного інтелекту організовується через функції *govern*, *map*, *measure* і *manage* (Tabassi, 2023). Для пошуково-доповненої системи це означає визначення відповідальних за джерела даних, опис сценаріїв використання, регулярні тестові набори запитань, вимірювання частки відповідей із коректними посиланнями та процедуру реагування на виявлені помилки. Особливо важливо вести журнал версій індексу, оскільки зміна корпусу даних здатна суттєво змінити поведінку системи без зміни самої мовної моделі.

Таким чином, пошуково-доповнена генерація є доцільною не як універсальний засіб «усунення галюцинацій», а як контрольована архітектура доступу генеративної моделі до перевірених зовнішніх знань. Її практична цінність визначається якістю даних, пошуку, метаданих і процедур оцінювання. Для організаційних інформаційних систем найбільш раціонально починати з вузького корпусу документів, чітко визначених типів запитань і вимоги показувати джерела кожного фактологічного висновку. Саме така конфігурація дає змогу перетворити генеративну модель з автономного «джерела відповідей» на інтерфейс до керованої бази знань.

REFERENCES / СПИСОК ДЖЕРЕЛ

1. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
2. Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences*, 13(4), 410. <https://doi.org/10.3390/educsci13040410>
3. Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>