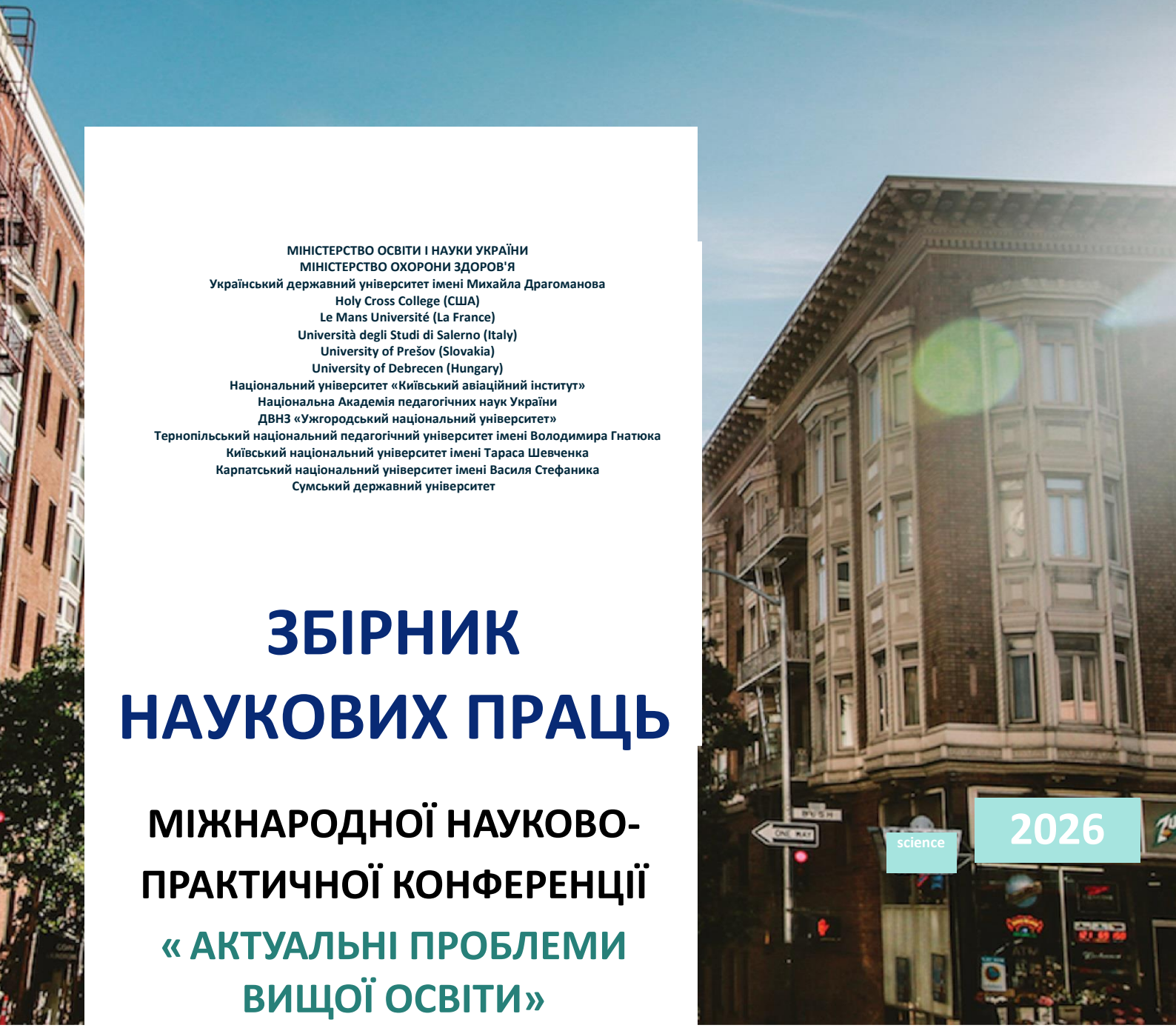

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
МІНІСТЕРСТВО ОХОРОНИ ЗДОРОВ'Я
Український державний університет імені Михайла Драгоманова
Holy Cross College (США)
Le Mans Université (La France)
Università degli Studi di Salerno (Italy)
University of Prešov (Slovakia)
University of Debrecen (Hungary)
Національний університет «Київський авіаційний інститут»
Національна Академія педагогічних наук України
ДВНЗ «Ужгородський національний університет»
Тернопільський національний педагогічний університет імені Володимира Гнатюка
Київський національний університет імені Тараса Шевченка
Карпатський національний університет імені Василя Стефаника
Сумський державний університет

ЗБІРНИК НАУКОВИХ ПРАЦЬ

МІЖНАРОДНОЇ НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

« АКТУАЛЬНІ ПРОБЛЕМИ ВИЩОЇ ОСВІТИ»

23 квітня 2026 року

Київ - 2026

УДК 378.02:001.89(08)

Збірник наукових праць Міжнародної науково-практичної конференції «Актуальні проблеми вищої освіти» / Відповідальний редактор проф. Т.Ю. Дудка. – К., 2026. 342 с.

Матеріали збірника друкуються в авторській редакції. За науковий зміст і якість поданих матеріалів відповідальність несуть учасники конференції

RAG ПІДХІД : АРХІТЕКТУРА ТА ЗАСТОСУВАННЯ В ОСВІТНЬОМУ ПРОЦЕСІ

Шемет Даніїл Аркадійович
здобувач наукового ступеня

Умрик Марія Анаталоійвна
професор кафедри інформаційних технологій і програмування
Український державний університет імені Михайла Драгоманова

Аналіз наукових джерел показує, що дослідники по-різному підходять до трактування сутності RAG. У роботі Patrick Lewis він описується як модель, що поєднує параметричну та непараметричну пам’ять, тобто внутрішні знання моделі із зовнішніми джерелами даних [5, с. 2]. Водночас у прикладних та оглядових дослідженнях RAG розглядається як підхід, що інтегрує генерацію тексту з механізмами інформаційного пошуку [1], або як гібридна архітектура, що поєднує можливості великих мовних моделей із системами пошуку знань [2, с. 2]. Попри відмінності у формулюваннях, усі підходи відображають спільну ідею - розширення можливостей генеративних моделей за рахунок зовнішніх даних.

Узагальнюючи наведені підходи, доцільно визначити RAG як підхід у сфері штучного інтелекту, що поєднує генерацію тексту великими мовними моделями з пошуком релевантної інформації у зовнішніх джерелах, забезпечуючи використання як внутрішніх, так і зовнішніх знань під час формування відповіді.

Поява RAG зумовлена обмеженістю параметричних знань мовних моделей, складністю їх оперативного оновлення та відсутністю прозорої прив’язки відповідей до джерел. Це створює фундаментальну невідповідність між здатністю моделей генерувати переконливі відповіді та необхідністю забезпечення їх достовірності й перевірюваності [6, с. 151].

Саме в контексті подолання цих обмежень сформувалися різні підходи до реалізації мовних агентів, серед яких особливе місце займає RAG. До основних підходів належать LLM-only, retrieval-only, fine-tuning та Retrieval-Augmented Generation (RAG). Серед них RAG вирізняється як збалансований підхід, що поєднує семантичний пошук із генерацією відповіді на основі знайденого контексту, що дозволяє підвищити точність і зменшити кількість помилок.

У науковій літературі RAG швидко еволюціонував від «класичної» схеми (один запит → top-k фрагментів → відповідь) до модульних і керованих архітектур із повторним пошуком, перевіркою, ранжуванням та оцінюванням якості [3, с. 4].

Практична реалізація RAG зазвичай включає послідовність компонентів: підготовку корпусу даних, сегментацію та векторизацію тексту, індексацію, обробку запиту користувача, пошук і ранжування релевантних фрагментів, формування промпта та генерацію відповіді. Додатково передбачаються механізми оцінювання якості та оновлення бази знань (рис. 1) [4].



Рис. 1. Схема базової RAG-архітектури.

У контексті освітніх застосувань RAG має низку суттєвих переваг. Він забезпечує прозорість відповідей завдяки можливості прив’язки до конкретних джерел, що сприяє дотриманню принципів академічної доброчесності. Архітектура RAG також дозволяє працювати з обмеженою базою знань, що є важливим для контролю змісту навчальних матеріалів і запобігання відхиленню від теми. Важливою перевагою є можливість оновлення знань без повторного

навчання моделі, що забезпечує гнучкість і економічну ефективність системи. Крім того, використання RAG у навчальному процесі сприяє формуванню у студентів розуміння ключових компонентів сучасних систем ШІ, зокрема embedding (векторні представлення), механізмів пошуку та оцінювання релевантності інформації. Водночас такі системи залишаються залежними від якості пошукового механізму та можуть містити залишкові помилки у відповідях.

Отже, проведений аналіз свідчить, що RAG сьогодні розглядається не як одна фіксована модель, а як підхід, який поєднує можливості великих мовних моделей із зовнішнім пошуком інформації. У різних дослідженнях акцент робиться або на поєднанні внутрішніх і зовнішніх знань, або на інтеграції генерації з механізмами retrieval, однак спільною ідеєю залишається розширення можливостей мовних моделей за рахунок актуальних джерел даних.

У цьому контексті підхід Retrieval-Augmented Generation (RAG) виступає як ефективне поєднання генеративних і пошукових механізмів, що забезпечує більш обґрунтоване формування відповідей.

З практичної точки зору, RAG підхід є особливо корисним там, де важлива актуальність і контрольованість інформації. В освітніх системах це проявляється у можливості прив’язувати відповіді до конкретних джерел, підвищувати прозорість результатів і зменшувати ризик появи недостовірної інформації.

Разом з тим, ефективність таких систем значною мірою залежить від організації процесів пошуку та відбору інформації, що актуалізує потребу у подальшому розвитку методів оптимізації RAG-архітектур і їх адаптації до освітніх завдань.

Література

1. Gillis A. S. Retrieval-Augmented Generation (RAG) [Електронний ресурс] // TechTarget SearchEnterpriseAI. – 2024. – Режим доступу: https://www.techtarget.com/searchenterpriseai/definition/retrieval-augmented-generation

2. Liu C. Retrieval-Augmented Generation: A Survey of Methodologies, Techniques, Applications, and Future Directions [Електронний ресурс] // ResearchGate. – 2025. – Режим доступу: https://www.researchgate.net/publication/397462545_Retrieval-Augmented_Generation_A_Survey_of_Methodologies_Techniques_Applications_and_Future_Directions

3. Guu K., Lee K., Tung Z., Pasupat P., Chang M. Retrieval Augmented Language Model Pre-Training [Електронний ресурс] // Proceedings of the 37th International Conference on Machine Learning (ICML). – 2020. – С. 3929–3938. – Режим доступу: https://proceedings.mlr.press/v119/guu20a/guu20a.pdf

4. Retrieval – Docs by LangChain [Електронний ресурс] // LangChain Documentation. – 2026. – Режим доступу: https://docs.langchain.com/oss/python/langchain/retrieval

5. 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. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks [Електронний ресурс] // Advances in Neural Information Processing Systems. – 2020. – Т. 33. – С. 9459–9474. – Режим доступу: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

6. Es S., James J., Espinosa-Anke L., Schockaert S. RAGAs: Automated Evaluation of Retrieval Augmented Generation [Електронний ресурс] // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. – 2024. – С. 150–158. – Режим доступу: https://aclanthology.org/2024.eacl-demo.16.pdf

7. Умрик М. А., Шемет Д. А., Струтинська О. В. Політика використання штучного інтелекту у процесі написання наукових робіт для студентів та освітян-науковців // Вісник науки та освіти. – 2025. – № 10(40). – С. 2410–2424. DOI: 10.52058/2786-6165-2025-10(40)-2410-2424.