Лазарович Юлія Ігорівна
студентка 4 курсу, групи ІПЗ-42
Карпатський національний університет імені Василя Стефаника
(034) 259-60-58
yuliia.lazarovych.22@pnu.edu.ua

# АРХІТЕКТУРА ГІБРИДНОЇ RAG-СИСТЕМИ ДЛЯ МУЛЬТИМОДАЛЬНОГО СЕМАНТИЧНОГО АНАЛІЗУ І ГЕНЕРАЦІЇ

**Постановка задачі.** Проблема ефективної обробки мультимодальних даних ускладнюється необхідністю глибокого розуміння візуальних сцен, де класичні методи пошуку демонструють суттєві обмеження. Алгоритми точного пошуку (KNN) погано масштабуються, а наближені методи (ANN) та векторні представлення часто втрачають специфічні деталі сцени через надмірне узагальнення. Додатковим викликом є схильність мовних моделей до галюцинацій при відсутності чіткого фактологічного підґрунтя, що вимагає зміщення акценту з чистої швидкодії на семантичну точність та логічну узгодженість результатів.

**Мета дослідження.** Метою роботи є розробка архітектури гібридної RAG-системи, що інтегрує векторні бази даних та графи знань для підвищення точності обробки запитів. Дослідження спрямоване на створення єдиного семантичного простору для різнорідних даних, реалізацію механізму пріоритету логічного міркування (reasoning-first) та забезпечення мінімалістичного контексту для генеративної моделі задля зменшення ймовірності помилок.

**Результати дослідження.** Розроблена архітектура базується на використанні датасету Visual Genome, який, на відміну від наборів COCO або OK-VQA, містить явні анотації відношень у форматі "суб'єкт-предикат-об'єкт", що дозволяє побудувати еталонний граф знань для валідації результатів. Гібридне сховище поєднує Weaviate та Neo4j 5.x, що інтегруються шляхом передачі ідентифікаторів об'єктів для графового аналізу. Векторизація мультимодальних даних виконується моделлю CLIP, що дозволяє сформувати єдиний семантичний простір для даних різної модальності. Ефективність первинного відбору забезпечується використанням індексу HNSW, для якого встановлено залежність між точністю та затримкою: при значенні параметра efSearch рівному 40 показник Recall@1 становить близько 0.73 із затримкою 11.5 мс, тоді як збільшення параметра до 640 підвищує Recall@1 до 0.97, але призводить до зростання затримки до 68 мс. З огляду на вимоги до швидкодії системи, оптимальним визначено діапазон efSearch у межах 100–200, що забезпечує Recall на рівні 0.85–0.9.

Структура модуля Re-ranking реалізує принцип reasoning-first та передбачає виконання multi-hop traversal у графі знань на глибину 1-2 кроки для виявлення семантично пов'язаних сутностей (Рис. 1). Для забезпечення логічної узгодженості застосовується механізм семантичного злиття, що об'єднує результати векторного пошуку та графові зв'язки через зважену суму, після чого

reasoning-агент у режимі zero-shot оцінює релевантність кандидатів, що звужує вибірку до 1-3 найбільш значущих кандидатів, формуючи мінімалістичний контекст для генерації. Генеративна підсистема побудована на базі моделі Gemma 3 (27B Модель Gemma 3 (27B) використовує cross-attention для поєднання мультимодальних ознак із структурованими знаннями графа.

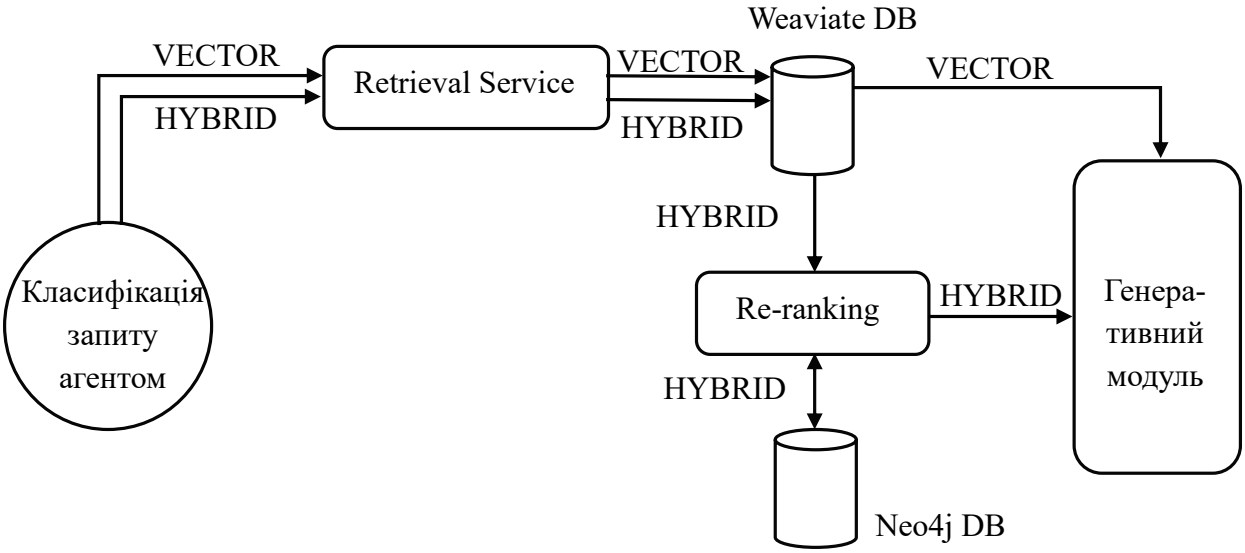


Рис. 1 Роль модуля Re-ranking.

Для мінімізації галюцинацій впроваджено контракт взаємодії через Prompt-Engine, який трансформує JSON-дані у промпт, що вимагає від моделі розставлення цитувань (node_id) для кожного згенерованого твердження (Рис. 2). Програмна верифікація відповідей здійснюється компонентом Output Formatter, який відхиляє непідтверджені цитати, забезпечуючи деградацію до безпечного опису сцени у разі виявлення розбіжностей.

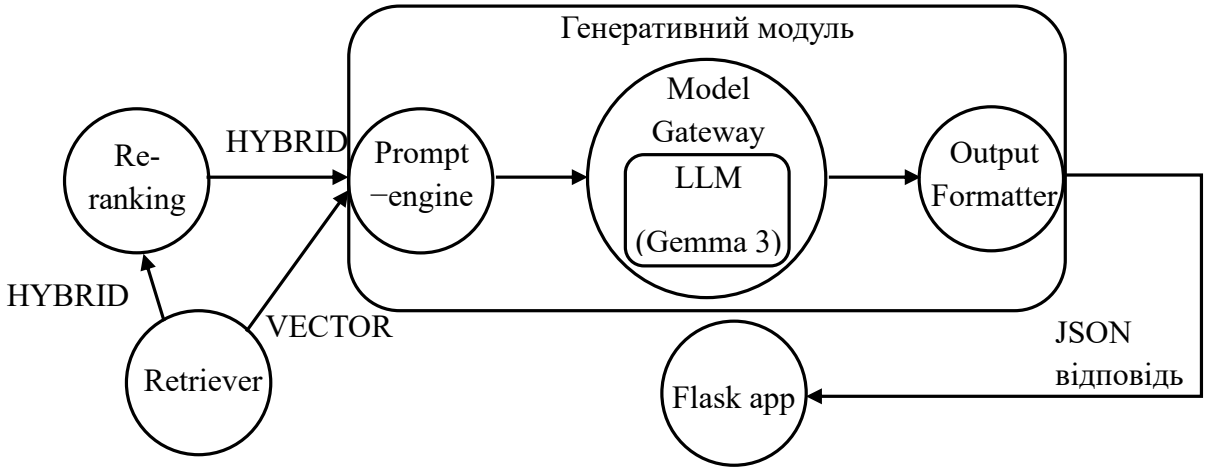


Рис. 2 Ланцюг генерації після reasoning-етапу.

Технічна реалізація системи виконана у вигляді асинхронного веб-додатка на базі Flask із використанням черги завдань Celery та брокера Redis, що дозволяє розподілити навантаження повного циклу RAG-пайплайну, який включає етапи

векторизації, оркестрації агентів та генерації, та є критичним для обробки складних логічних запитів.

Комплексна валідація архітектури на наборі Visual Genome передбачає оцінку модуля пошуку за метриками Recall@k, Precision@k та MRR з урахуванням вимог до затримки (Latency). Ефективність логічного переранжування визначається точністю формування малого контексту та помилкою калібрування (ECE), тоді як якість фінальної генерації перевіряється через порівняння з еталонними відповідями за допомогою лексичних метрик BLEU та ROUGE.

**Висновки та перспективи.** Запропонована гібридна архітектура має на меті поєднати швидкість векторного пошуку з точністю графового аналізу, що є критичним для глибокого розуміння візуального контексту. Впровадження етапу логічного переранжування (reasoning-first) є доцільним для фільтрації семантично близьких, але логічно невідповідних результатів, а використання механізмів верифікації цитувань сприяє підвищенню надійності генерації. Подальші дослідження можуть бути спрямовані на емпіричне визначення оптимального балансу між векторною та графовою складовими при формуванні контексту. Перспективним є впровадження багаторівневого кешування для зменшення навантаження на базу знань та оптимізації параметрів індексації.

**Список використаних джерел**

1. Radford A. et al. Learning Transferable Visual Models From Natural Language Supervision [Електронний ресурс] // arXiv. 2021. Preprint arXiv:2103.00020. Режим доступу: https://arxiv.org/abs/2103.00020 (дата звернення: 19.11.2025).

2. Choudhary M. Hybrid Agentic RAG with Neo4j and Weaviate using LlamaAgent [Електронний ресурс] // Medium. 2024. Режим доступу: https://medium.com/@milind.choudhary.42/hybrid-agentic-rag-with-neo4j-and-weaviate-using-llamaagent-b09a517e949e (дата звернення: 19.11.2025).

3. Briggs J. Hierarchical Navigable Small Worlds (HNSW) [Електронний ресурс] // Pinecone Learn Series. б. р. Режим доступу: https://www.pinecone.io/learn/series/faiss/hnsw/ (дата звернення: 19.11.2025).

4. Vector Search and Knowledge Graphs: Better Together [Електронний ресурс] // Neo4j Blog. б. р. Режим доступу: https://neo4j.com/blog/developer/vectors-graphs-better-together/ (дата звернення: 19.11.2025).

5. What are the key configuration parameters for an HNSW index [Електронний ресурс] // Milvus Documentation. б. р. Режим доступу: https://milvus.io/ai-quick-reference/what-are-the-key-configuration-parameters-for-an-hnsw-index-such-as-m-and-efconstructionefsearch-and-how-does-each-influence-the-tradeoff-between-index-size-build-time-query-speed-and-recall (дата звернення: 19.11.2025).