Лазарович Юлія Ігорівна
студентка 4 курсу, групи ІПЗ-42
Карпатського національного університету
імені Василя Стефаника
м. Івано-Франківськ
+380 (342) 59-60-58
yuliia.lazarovych.22@pnu.edu.ua
Науковий керівник: Козленко Микола Іванович
кандидат технічних наук
доцент кафедри інформаційних технологій
Карпатського національного університету
імені Василя Стефаника
м. Івано-Франківськ

АНАЛІЗ СУЧАСНИХ ПІДХОДІВ ДО РОЗУМІННЯ ВІЗУАЛЬНИХ СЦЕН ДЛЯ РЕАЛІЗАЦІЇ МУЛЬТИМОДАЛЬНОГО RAG-ПІДХОДУ

Постановка задачі. Функціонування сучасних смарт-асистентів обмежене статичною природою навчальних наборів даних, що спричиняє неактуальність інформації та виникнення «галюцинацій» при обробці рідкісних запитів [1]. Проблема критична для задач візуального розуміння, де інтерпретація зображень вимагає точних знань про об’єкти, відсутні в тренувальній вибірці. Ефективним шляхом вирішення зазначеної проблеми є застосування підходу Retrieval-Augmented Generation (RAG), який інтегрує генеративні можливості великих мовних моделей із зовнішніми базами знань.

Мета дослідження. Метою роботи є аналіз існуючих методів інтеграції знань та формулювання вимог до мультимодальної RAG-системи, спрямованої на забезпечення високої семантичної точності та бажаної швидкодії. Дослідження фокусується на розробці підходу, що дозволив би смарт-асистенту надавати обґрунтовані відповіді на візуальні запитання, використовуючи динамічні бази знань без необхідності постійного перенавчання моделі, при цьому дотримуючись обмежень щодо часу реакції системи.

Результати дослідження. У ході роботи проаналізовано ключові архітектури RAG для визначення оптимальної конфігурації системи. Зокрема, метод REVEAL реалізує концепцію наскрізного навчання, де ретривер і генератор оптимізуються спільно з використанням гібридних даних та механізму attentive fusion, що забезпечує високу семантичну узгодженість, проте його ефективність прямо залежить від повноти бази знань, а компресія даних може призводити до втрати деталей [2]. Інший підхід, Visual-RAG, базується на text-to-image пошуку і доводить, що для відповідей на питання про специфічні візуальні деталі пошук схожих зображень є ефективнішим за текстовий, хоча цей метод обмежений вузькою доменною спеціалізацією [3]. Альтернативна архітектура RAP фокусується на перцептивній точності при обробці зображень високої роздільності, застосовуючи динамічний пошук фрагментів зі

збереженням просторових відношень, однак це супроводжується значними обчислювальними витратами [4]. Також розглянуто напрям mKG-RAG, що використовує мультимодальні графи знань для покращення структурного розуміння та логічного міркування, мінімізуючи інформаційний шум, але залишаючись вразливим до помилок на етапі побудови графа [5].

На основі проведеного аналізу сформульовано вимоги до проектованої системи (Рис. 1) . Для забезпечення балансу між точністю та швидкодією доцільним визначено використання комбінованого запиту, що поєднує зображення та текст, оскільки така конфігурація сприяє кращому узгодженню семантичних ознак. Важливим аспектом є впровадження reasoning-шару з listwise-ранжуванням, що дозволяє оцінювати список кандидатів одночасно, підвищуючи релевантність відбору порівняно з поточковими методами. Також пропонується мінімізувати контекст, що передається генератору до 1–3 найбільш релевантних документів, щоб запобігти розмиванню уваги моделі та підвищити семантичну узгодженість відповіді. Технічна реалізація системи передбачає використання векторних індексах та механізмів cross-attention, забезпечуючи роботу в режимі zero-shot без додаткового навчання параметрів

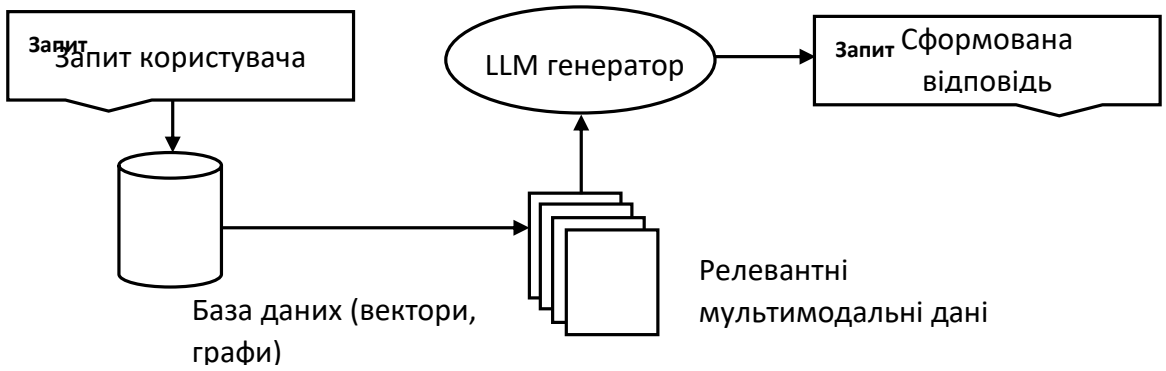

Рисунок 1 — Схема архітектури Retrieval-Augmented Generation (RAG).

Висновки та перспективи. Запропонована архітектура mRAG розглядається як перспективний підхід до забезпечення актуальності знань у смарт-асистентах. Поєднання векторного пошуку, структурованого ранжування та мінімізації контексту дозволяє досягти необхідного балансу між семантичною точністю відповіді та обчислювальною ефективністю системи.

Список використаних джерел

1. C.-W. Hu et al., “mRAG: Elucidating the design space of multi-modal retrieval-augmented generation,” arXiv preprint arXiv:2505.24073, 2025, doi: 10.48550/arXiv.2505.24073.

2. Z. Hu et al., “Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory,” in 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 2023, pp. 23369–23379, doi: 10.1109/CVPR52729.2023.02238.

3. Y. Wu et al., “Visual-RAG: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries,” arXiv preprint arXiv:2502.16636, 2025, doi: 10.48550/arXiv.2502.16636.

4. W. Wang et al., “Retrieval-augmented perception: High-resolution image perception meets visual RAG,” in Proc. 42nd Int. Conf. Mach. Learn. (ICML), 2025, doi: 10.48550/arXiv.2503.01222.

5. X. Yuan et al., “mKG-RAG: Multimodal knowledge graph-enhanced RAG for visual question answering,” arXiv preprint arXiv:2508.05318, 2025, doi: 10.48550/arXiv.2508.05318.