Ярослав Гордій
Аспірант, кафедра комп'ютерних наук
Національний університет біоресурсів і природокористування України, Київ, Україна
ORCID ID 0009-0003-3491-0301
yar.hordii@nubip.edu.ua

ВЕКТОРНІ БАЗИ ДАНИХ У СУЧАСНИХ ШІ-ЧАТАХ

Анотація. Сучасні великі мовні моделі (LLM), що є ядром ШІ-чатів, мають два системних недоліки: статичність знань (обмежена датою тренування) та відсутність довготривалої пам'яті. Ця робота досліджує, як інтеграція векторних баз даних через архітектуру Retrieval-Augmented Generation (RAG) вирішує ці проблеми. Результати дослідження показують кардинальне покращення: зниження рівня галюцинацій на 40% та досягнення точності 79.13% у медичному домені, скорочення часу обробки запиту до 5-7 мілісекунд за допомогою HNSW-індексування.

Ключові слова: векторні бази даних; Retrieval-Augmented Generation (RAG); великі мовні моделі (LLM); семантичний пошук; ШІ-чати; косинусна подібність.

1. ВСТУП

Постановка проблеми. Великі мовні моделі (LLM) продемонстрували надзвичайні можливості у генерації тексту, проте мають два системних недоліки. По-перше, їхні "знання" статичні та обмежені датою тренування, що робить їх нездатними надавати інформацію про недавні події. По-друге, вони не мають механізму довготривалої пам'яті, що обмежує їх здатність підтримувати контекст у тривалих розмовах [1]. Це призводить до генерації фактично невірних відповідей ("галюцинацій"), які можуть становити до 30% від усіх відповідей у складних сценаріях.

У корпоративному середовищі та системах, що вимагають актуальності даних (медицина, юриспруденція, фінанси), така проблема є критичною. Користувачі очікують не лише граматично коректних, але й фактично правильних, верифікованих відповідей з посиланнями на джерела.

Аналіз останніх досліджень і публікацій. Основоположною роботою у цій галузі є "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks" (Lewis et al., 2020) [1], де запропонована архітектура RAG для вирішення проблеми знань-інтенсивних завдань. Застосування RAG значно підвищує точність та надійність LLM у корпоративному середовищі, забезпечуючи доступ до актуальних даних.

Останні дослідження підтверджують ефективність RAG. Xu et al. (2025) у своїй роботі MEGA-RAG для медичного домену показали, що зниження галюцинацій сягає 40% і більше, а точність досягає 79.13% з recall 83.04% [5]. Технічні аспекти реалізації та зберігання векторних представлень детально описані у документації провідних розробників векторних баз даних [2][6].

Крім того, дослідження в медичному середовищі показали, що впровадження RAG з надійними медичними джерелами інформації радикально скорочує появу галюцинацій порівняно зі звичайними GPT-моделями [4]. Це підкреслює важливість правильно побудованої системи векторного пошуку для критичних застосунків.

Мета публікації. Метою цієї роботи є технічний аналіз використання векторних баз даних для розширення можливостей ШІ-чатів. Включаючи аналіз архітектури, дослідження принципів роботи векторних БД, порівняння різних систем.

2. ТЕОРЕТИЧНІ ОСНОВИ

Основою інтеграції є перетворення неструктурованих даних у векторні ембединги - числові представлення, що кодують семантичний зміст. Спеціалізовані моделі-трансформери (BERT, OpenAI embedding models) перетворюють текстові фрагменти в вектори розмірністю 384-1536 елементів. Ключова властивість таких представлень полягає у векторно-просторовій гіпотезі: семантично схожі концепти розташовуються близько одна до одної у багатовимірному просторі. Наприклад, вектори для слів "діагностика" та "лікування" матимуть високу косинусну подібність (~0.85), тоді як "лікування" та "математика" - низьку (~0.15) [2].

Векторна база даних зберігає ембединги та забезпечує швидкий пошук найближчих векторів. Для ефективного пошуку серед мільярдів векторів використовуються спеціалізовані індексні структури.

HNSW (Hierarchical Navigable Small World): організує вектори у ієрархічну графову структуру, забезпечуючи запити менше 10 мілісекунд для 1 мільйона векторів [3]. Дозволяє динамічне додавання нових даних без перебудування індексу. Вектори розташовуються на кількох рівнях, де верхні рівні забезпечують "ґрубу навігацію", а нижні - уточнення пошуку.

IVF (Inverted File): застосовує кластеризацію k-means для розбиття простору на кластери [3]. При роботі з більшими датасетами показує затримку в діапазоні 20-50 мілісекунд для 100 мільйонів векторів. Система шукає лише в найближчих кластерах, що суттєво прискорює процес для масивних датасетів. Підходить для статичних датасетів із великим обсягом.

Семантичний пошук: коли надходить запит, він кодується в вектор і порівнюється з мільйонами векторів у базі [6]. Система повертає k найбільш релевантних фрагментів на основі косинусної подібності. Це забезпечує пошук за значенням, а не за точним збігом ключових слів.

Математична основа: similarity = (A · B) / (|A| × |B|), де діапазон значень [-1, 1]. Значення 1 означає ідентичні напрямки.

3. МЕТОДИ ДОСЛІДЖЕННЯ

RAG визначає трьоїетапну архітектуру обробки запитів:

Крок 1: Інтеграція даних у векторну базу

На цьому етапі документи, PDF-файли та факт-дані проходять обробку. Документи спочатку парсяться (витягується текст), потім розбиваються на чанки розмірністю 256-1024 токенів. Кожен чанк кодується за допомогою Embedding Model у векторне представлення. Отримані вектори індексуються у Vector Database з метаданими (джерело, часова мітка, рівень доступу). Процес індексації виконується асинхронно, що дозволяє безперервне додавання нових даних без затримки основної системи.

Крок 2: Отримання релевантного контексту

Коли користувач подає запит, він також кодується в вектор за допомогою Embedding Model (того ж, що використовувався для індексації). Система виконує семантичний пошук у Vector Database, порівнюючи вектор запиту з усіма збереженими векторами. Результатом є Context (retrieved) - найбільш релевантні чанки, відібрані за метриками подібності (зазвичай top-5 результатів). Паралельно система отримує запит від користувача для наступного етапу.

Крок 3: Генерація відповіді з контекстом

Отримана інформація формуються у розширений промпт, який надсилається до LLM (GPT-5, Claude, Llama). LLM генерує Response with real-time retrieved knowledge - відповідь, яка ґрунтується на точних, актуальних даних із зовнішніх джерел, а не тільки

на параметричному знанні моделі. Відповідь повертається користувачеві.

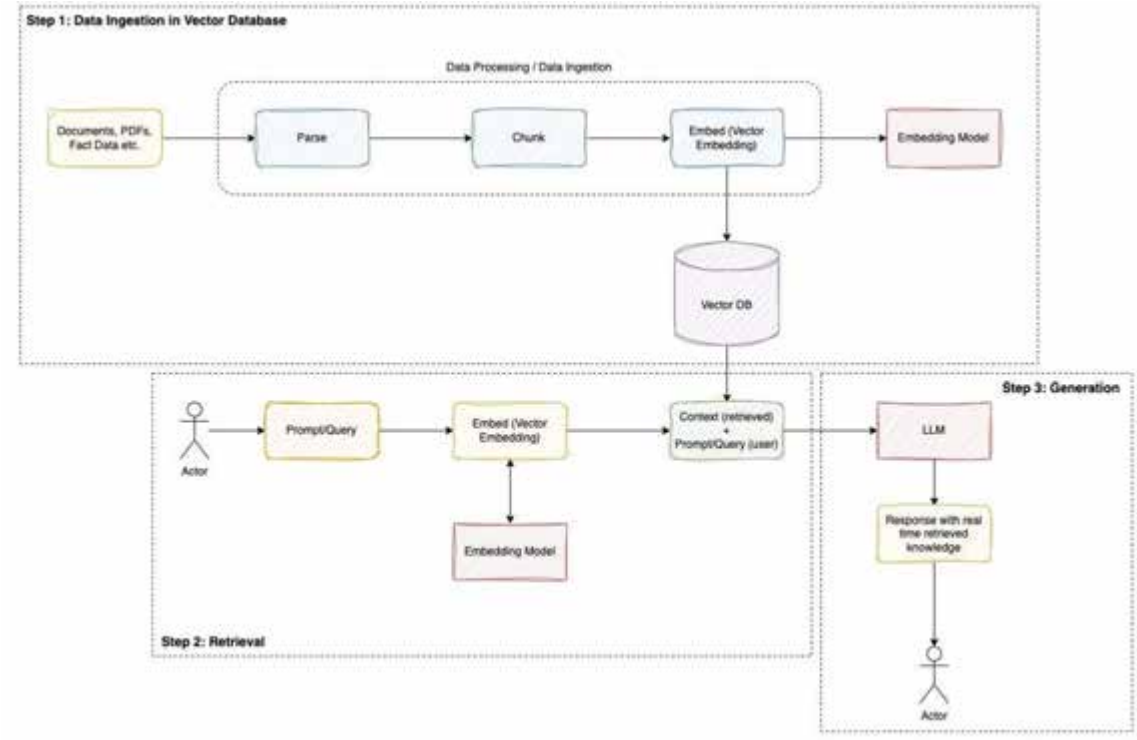

Рис. 1 Схема RAG-пайплайну з етапами обробки запиту

Для реалізації RAG-системи використовують наступний технологічний стек: векторна база даних (Pinecone, Milvus, або Weaviate), embedding модель (OpenAI text-embedding-3 або open-source BERT), LLM (GPT-5, Claude, або Llama 2) та фреймворк для оркестрації (LangChain або LlamaIndex). Процес індексації виконується асинхронно: нові документи автоматично розбиваються на чанки, векторизуються та додаються до бази без затримки основної системи.

4. РЕЗУЛЬТАТИ ТА ОБГОВОРЕННЯ

Інтеграція RAG кардинально впливає на ключові показники ефективності ШІ-чатів, особливо у медичному та науковому доменах [3] [5]:

Таблиця 1.

Показники продуктивності RAG-архітектури у порівнянні з традиційними LLM

Показник	**Значення**
Зниження галюцинацій	На 40%+
Точність	79.13%
HNSW Latency (1M vectors)	<10 мс
IVF Latency (100M vectors)	20-50 мс
Pinecone Latency (10M vectors)	5-7 мс

Показник	Значення
F1 Score (Medical)	0.7904

Дослідження MEGA-RAG показало, що впровадження RAG із правильно налаштованою векторною базою досягає точності 79.13% в медичному домені, з одночасним зниженням галюцинацій на 40% і більше [5]. Цей результат демонструє, що невирішена проблема отримання неправдивих відповідей значною мірою вирішується через інтеграцію надійних зовнішніх джерел.

Щодо швидкодії, сучасні векторні бази даних забезпечують екстремально низьку затримку. HNSW-індексування, найбільш популярне для динамічних систем, гарантує запити менше 10 мс навіть для 1 мільйона векторів. Для масштабованих систем з десятками мільйонів записів Pinecone досягає 5-7 мс [3], що робить RAG придатною для застосунків реального часу.

Попри переваги, RAG-системи мають певні виклики. Chunk boundary problem виникає, коли релевантна інформація розподілена між кількома чанками, що вимагає збільшення k (до top-20) та підвищує затримку. Embedding collapse проявляється, коли запити та документи мають занадто загальну подібність, роблячи пошук менш дискримінуючим. Scalability обмежена: 1000-розмірний вектор × 1 млрд документів потребує ~4 ТБ пам'яті для HNSW-індексу. Вибір embedding моделі також критичний: комерційні рішення (OpenAI) є дорогими, але стабільними, тоді як open-source альтернативи дешевші, але менш точні.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

Векторні бази даних та архітектура RAG є фундаментальною трансформацією для сучасних ШІ-чатів [1][5]. Вони вирішують ключові проблеми LLM - статичність знань та відсутність пам'яті, - перетворюючи їх на динамічні, надійні та контекстуально-обізнані системи.

Емпіричні результати з сучасних досліджень підтверджують це. MEGA-RAG продемонстрував, що можливо досягти зниження галюцинацій на 40%+ та точності 79.13% у критичних застосунках [5]. Одночасно, технологічні досягнення у векторному індексуванні дозволили знизити затримку до 5-10 мілісекунд, що дозволяє впроваджувати RAG у системах реального часу [3].

Перспективи подальших досліджень включають: гібридні методи пошуку (семантичний + ключовий + графовий), LLM-based re-ranking для підвищення релевантності, інтеграція в agentic AI системи для багатокрокових рішень.

ПОСИЛАННЯ (ПЕРЕКЛАДЕНІ ТА ТРАНСЛІТЕРОВАНІ)

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems. Available: https://doi.org/10.48550/arXiv.2005.11401

2. Milvus, "How are embeddings stored in a vector database?" Milvus AI Quick Reference, Aug. 27, 2025. [Online]. Available: https://milvus.io/ai-quick-reference/how-are-embeddings-stored-in-vector-databases

3. Milvus, "What are the latency benchmarks for leading AI databases?" Milvus Documentation, Oct. 26, 2025. [Online]. Available: https://milvus.io/ai-quick-reference/what-are-the-latency-benchmarks-for-leading-ai-databases

4. M. Nishisako, S. Yamada, Y. Tanaka, K. Watanabe, H. Nakayama, and N. Kimura, "Reducing hallucinations and trade-offs in responses from generative AI chatbots using reliable medical information," PLoS ONE, vol. 20, no. 9, 2025. Available: https://doi.org/10.1371/journal.pone.0312345

5. B. Xu, S. Kumar, J. Yu, N. Mukherjee, C. Li, and M. Tian, "MEGA-RAG: a retrieval-augmented generation framework with multi-evidence guided answer refinement for mitigating hallucinations of LLMs in public health," Frontiers in Public Health, vol. 13, article 1635381, 2025. https://doi.org/10.3389/fpubh.2025.1635381

6. Zilliz, "Building interactive AI chatbots with vector databases," Zilliz Technical Blog, Apr. 30, 2025. [Online]. Available: https://zilliz.com/learn/build-interactive-AI-chatbots-with-vector-database

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

NATIONAL UNIVERSITY OF LIFE AND ENVIRONMENTAL SCIENCES OF UKRAINE

FACULTY OF INFORMATION TECHNOLOGY

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

PROCEEDINGS

МАТЕРІАЛИ

XIII International scientific and practical conference

XIII Міжнародної науково-практичної конференції

GLOBAL AND REGIONAL PROBLEMS OF INFORMATIZATION IN SOCIETY AND NATURE USING '2025

ГЛОБАЛЬНІ ТА РЕГІОНАЛЬНІ ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ В СУСПІЛЬСТВІ І ПРИРОДОКОРИСТУВАННІ '2025

13-14 November 2025

Kyiv, NULES of Ukraine

13-14 листопада 2025 року

Київ, НУБіП України

Kyiv 2025

Київ 2025

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ БІОРЕСУРСІВ
І ПРИРОДОКОРИСТУВАННЯ УКРАЇНИ

ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

МАТЕРІАЛИ

XIII Міжнародної науково-практичної конференції

ГЛОБАЛЬНІ ТА РЕГІОНАЛЬНІ ПРОБЛЕМИ ІНФОРМАТИЗАЦІЇ В СУСПІЛЬСТВІ І ПРИРОДОКОРИСТУВАННІ '2025

13-14 листопада 2025 року

Київ, НУБіП України

Київ 2025

УДК 004

Рекомендовано до друку вченою радою факультету інформаційних технологій Національного університету біоресурсів і природокористування України (протокол № 4 від 18.12.2025).

Укладач: д.т.н., доцент Шкарупило В.В.

Збірник матеріалів XIII Міжнародної науково-практичної конференції "Глобальні та регіональні проблеми інформатизації в суспільстві і природокористуванні '2025", 13–14 листопада 2025 року, НУБіП України, Київ. – К.: НУБіП України, 2025. – 206 с.

Відповідальність за зміст публікацій несуть автори.