УДК 00.451(045)
DOI: 10.18372/2073-4751.77.18666

Холявкіна Т.В., к.т.н.,
orcid.org/0000-0003-2595-9405,
e-mail: holyavkina.t@gmail.com,
Чуба І.В., к.т.н.,
orcid.org/0000-0003-3336-5105,
e-mail: irishachuba@gmail.com,
Шолупата А.Ю.,
orcid.org/0009-0006-0919-6902,
e-mail: a.y.sholupata@gmail.com

СИНЕРГІЯ КОСИНУС-ПОДІБНОСТІ ТА ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ

Національний авіаційний університет

Вступ

Штучний інтелект (ШІ) з'явився як галузь комп'ютерних наук у середині 20 століття. Перші спроби створення штучного інтелекту були пов'язані з ідеєю створення машин, які могли б відтворювати розумові функції людини. Важливу роль у розвитку ШІ відіграє застосування методів машинного навчання, таких як нейронні мережі, що дозволяє системам пошуку вдосконалюватися та адаптуватися до змінних потреб користувачів.

Штучний інтелект змінив спосіб пошуку інформації. Поки більшість пошукових систем використовує пошук на основі ключових слів, штучний інтелект пропонує новий спосіб. Нейронний і семантичний пошук, який покладається не на самі ключові слова, а зосереджується на їх значенні.

З розвитком технологій змінюються і запити користувачів. Згідно до результатів опитування 2021 року проведеного *McKinsey & Company*, 71% користувачів очікують персоналізованих результатів пошуку та часто розчаровуються, коли ці очікування не виправдовуються [1].

Для того, щоб задовольнити потреби користувачів створюють чат-ботів. Чат-боти на основі штучного інтелекту з'явились завдяки поєднанню ряду технологій. Перші етапи включали в себе використання правил та програм для відповіді на конкретні запитання.

Сучасні чат-боти використовують алгоритми машинного навчання, зокрема нейронні мережі, для розпізнавання намірів користувачів, контекст їх запитів та навіть можуть підтримувати діалог.

Чат-боти використовуються в різних сферах, включаючи обслуговування клієнтів, консультування, підтримку користувачів, тощо. Їх популярність постійно зростає відповідно до попиту від користувачів, більшість з яких регулярно використовує чат-ботів в повсякденному житті.

Аналіз останніх досліджень і публікацій

Ринок штучного інтелекту переживає феноменальний розквіт, його поточна вартість наближена до 100 мільярдів доларів США (рис.1). Якщо у 2021 році вартість світового ринку була 93,5 мільярди, то вже в 2022 році вона збільшилась до 136,6 мільярдів доларів. За прогнозами *Next Move Strategy Consulting* ринок штучного інтелекту очікує двадцятикратний зріст до 2030 року, зростаючи з рівнем *CAGR* у 38,1% протягом прогнозованого періоду, що приблизить вартість до майже двох трильйонів доларів США [2].

Особливою популярність користується так званий *Generative AI*(генеративний ШІ) – тип штучного інтелекту, що має здатність вивчати та повторно застосовувати властивості та закономірності даних для широкого діапазону дій, від створення тексту, зображень та відео в різних стилях до генерації персоналізованого контенту.

Це дозволяє машинам виконувати творчі завдання, які раніше були виключно для людей.

Згідно до опитування *Capgemini Research Institute* майже всі керівники (96 %) в опитуванні вказали, що генеративний ШІ – важлива тема обговорення на засіданнях їхніх рад директорів (рис. 2), що підтверджує великий потенціал розвитку штучного інтелекту [3].

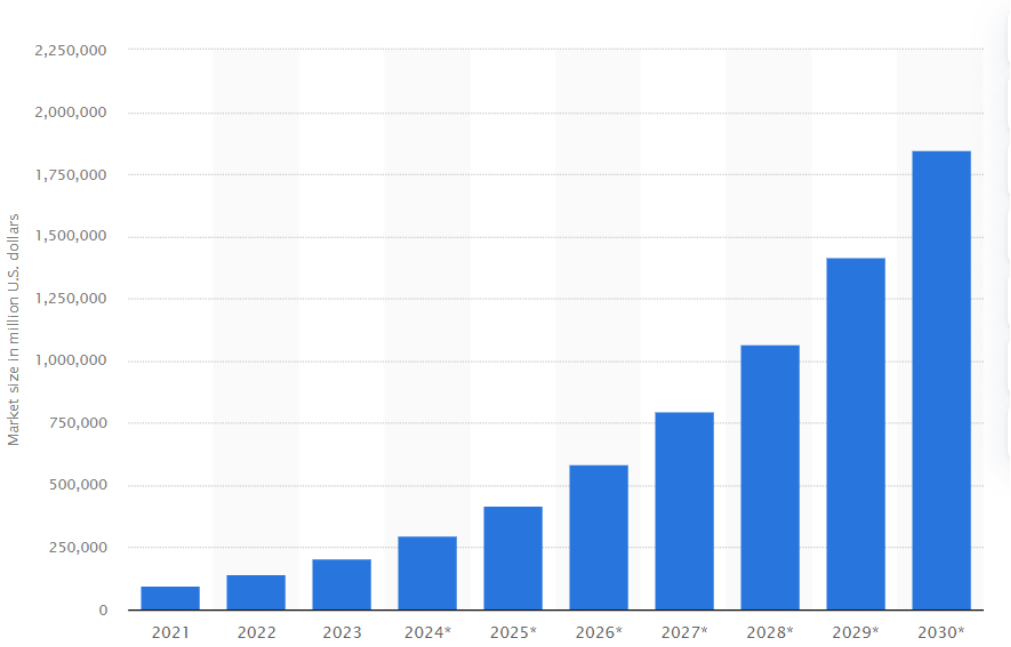

Рис. 1. Прогноз зростання вартості ринку штучного інтелекту

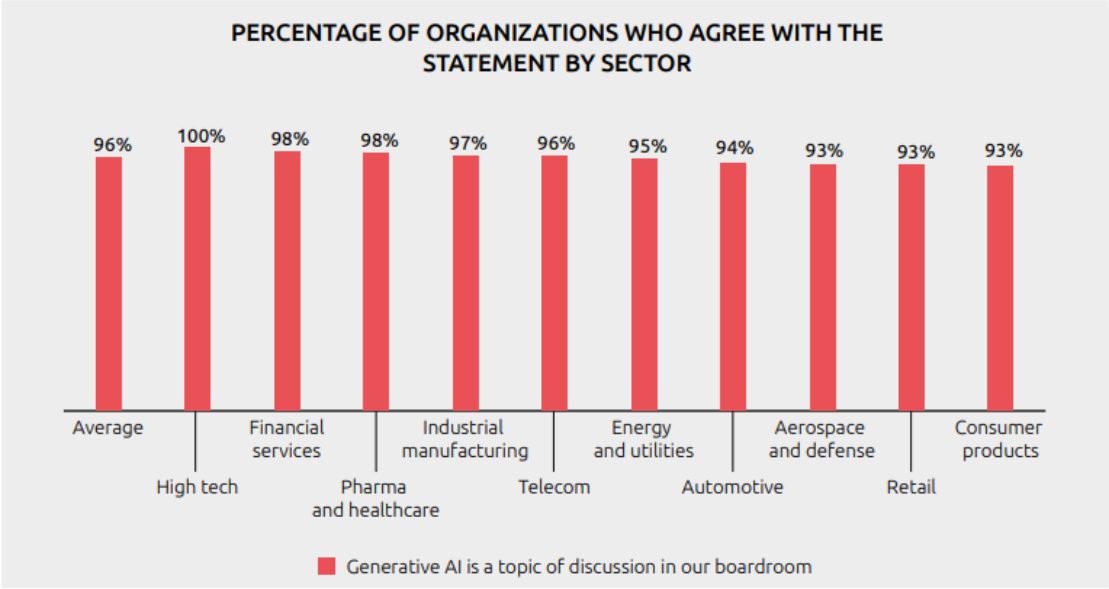

Рис. 2. Статистика обговорень штучного інтелекту серед керівників

У березні 2023 року *Aberdeen Strategy & Research* провели опитування серед 642 спеціалістів з різних галузей (рис. 3). На запитання «Як вони будуть знаходити інформацію в інтернеті в майбутньому», 42% опитуваних обрало чат-боти зі штучним інтелектом [4].

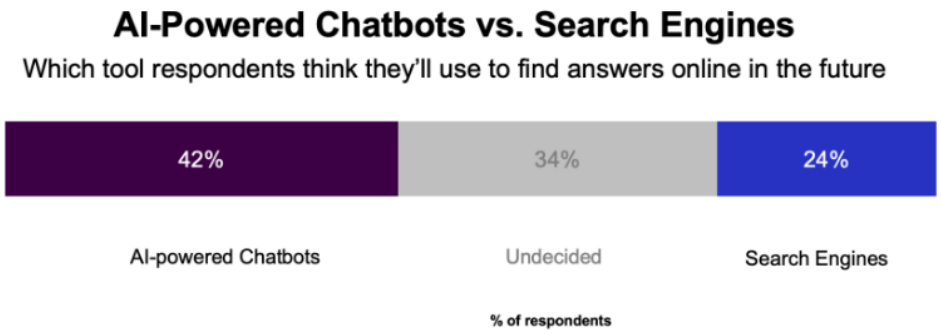

Рис. 3.Опитування, щодо використання чат-ботів для інтернет пошуку

За опитуванням компанії *Verint*, щодо використання чат-ботів для надання послуг, більшість опитуваних позитивно ставиться до взаємодії з чат-ботами на базі штучного інтелекту (рис. 4). Серед переваг клієнти обирають «економію часу та швидке вирішення проблеми» (47%), «легкість переходу від чат-боту до живої людини»(33%) та «полегшення обслуговування»(33%) [5].

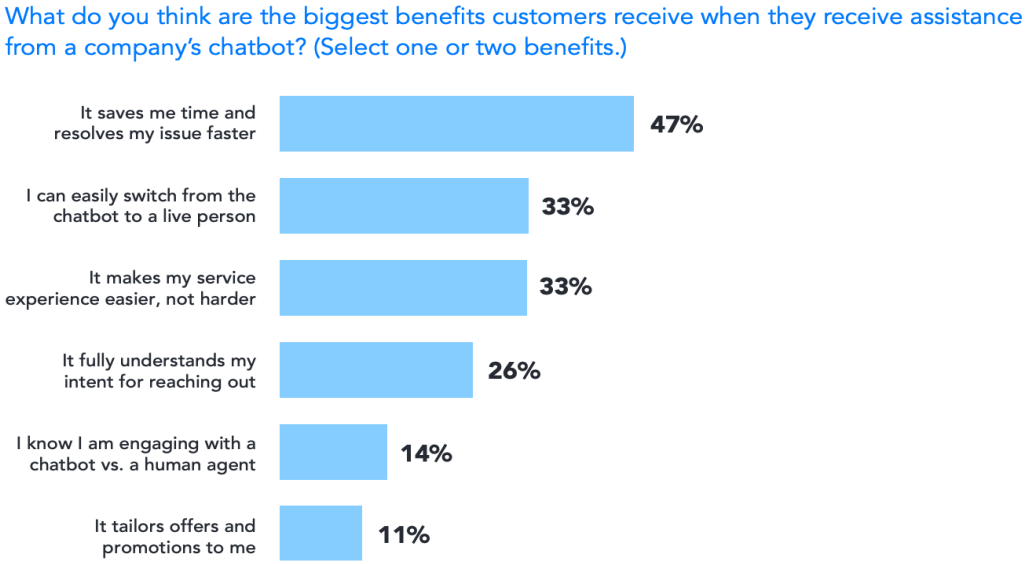

Рис. 4. Переваги, які знаходять користувачі при використанні чат-ботів

Мета статті (постановка завдання)

Мета цієї статті дослідити та проаналізувати використання синергії косинус-подібності та технологій штучного інтелекту для розробки додатку контекстного пошуку.

Основна частина

Основний стек технологій, які використовуються в додатку є наступним: *Java*, *SpringBoot*, *PostgreSQL* та *pgvector*. Додаток пошуку є масштабованим, пристосований до мікро-сервісної архітектури. Генерація векторів для бази знань відбувається за допомогою *OpenAI API*.

Для створення додатку контекстного пошуку, спочатку потрібно розібрати, як працює пошук подібності між векторами, який здійснюється на основі косинусу подібності.

Косинус подібності – коефіцієнт подібності двох не нульових векторів у предгільбертовому просторі, який обчислюється як косинус кута між ними. Косинус 0° дорівнює 1, а для всіх інших значень кута в інтервалі $(0,\pi]$ буде менше за 1. Отож, це оцінка напрямку, а не величини: два вектори з однаковим напрямком мають косинус подібності 1, а два вектора, які утворюють кут 90° один відносно одного,

мають подібність 0, а два діаметрально направлені вектори мають подібність -1, незалежно від їх довжини [6].

Формула косинусної подібності між двома ненульовими векторами:

$$cos(\theta) = \frac{AB}{|A||B|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

Кожне слово можна представити у вигляді вектора з певними значеннями. Наприклад, представимо зображені на рис. 5. слова у вигляді ненульових векторів у просторі: Кіт [(48,12)] та Пес [(50,10)].

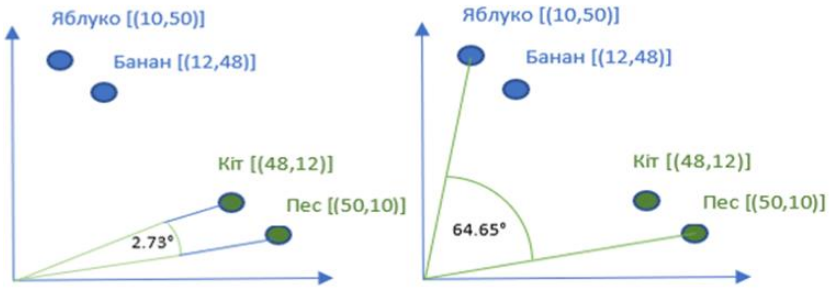

Рис. 5. Приклад знаходження подібності векторів

Використовуючи формулу подібності, обчислимо значення векторів:

$$AB = 50 \times 48 + 10 \times 12 = 2400 + 120 = 2520$$

$$|B| = \sqrt{48^2 + 12^2} = \sqrt{2304 + 144} = \sqrt{2488} \approx 49.5$$

$$|A| = \sqrt{50^2 + 10^2} = \sqrt{2500 + 100} = \sqrt{2600} \approx 51$$

$$cosine\ similarity = \frac{2520}{51 \cdot 49.5} \approx 0.998$$

cos(2,73)= 0.998 – Кіт і пес;

cos(64,65)= 0.4284 – Яблуко і пес.

Згідно до отриманих коефіцієнтів між словами Кіт та Пес подібність буде більшою, ніж між словами Яблуко та Пес.

Контекстний пошук

У контексті штучного інтелекту контекстний пошук стосується здатності системи розуміти та обробляти запити користувачів на основі намірів і змісту запиту, а не просто покладатися на ключові слова. Контекстний пошук прагне зрозуміти нюанси та зв'язки слів у запиті, щоб отримати більш відповідні результати [7].

Для генеративного штучного інтелекту семантичний пошук є критично важливим, оскільки мова йде не лише про отримання інформації, але й про створення вмісту, який відповідає запитам користувача. Наприклад, для підтримання розмови з користувачем, штучному інтелекту потрібно буде розуміти тему, яку обговорюють та будувати діалог на основі написаного користувачем.

Контекстний пошук оснований на векторному пошуку, який дозволяє ранжувати вміст на основі важливості контексту та релевантності наміру. Векторний пошук кодує деталі інформації, доступної для пошуку, у поля пов'язаних термінів або елементів, а потім порівнює їх, щоб визначити, які з них найбільш схожі. Схема роботи контекстного пошуку представлена на (рис. 6).

Коли виконується запит, пошукова система перетворює запит на ембедінги, які є числовими представленнями даних і зберігає їх у вигляді векторів.

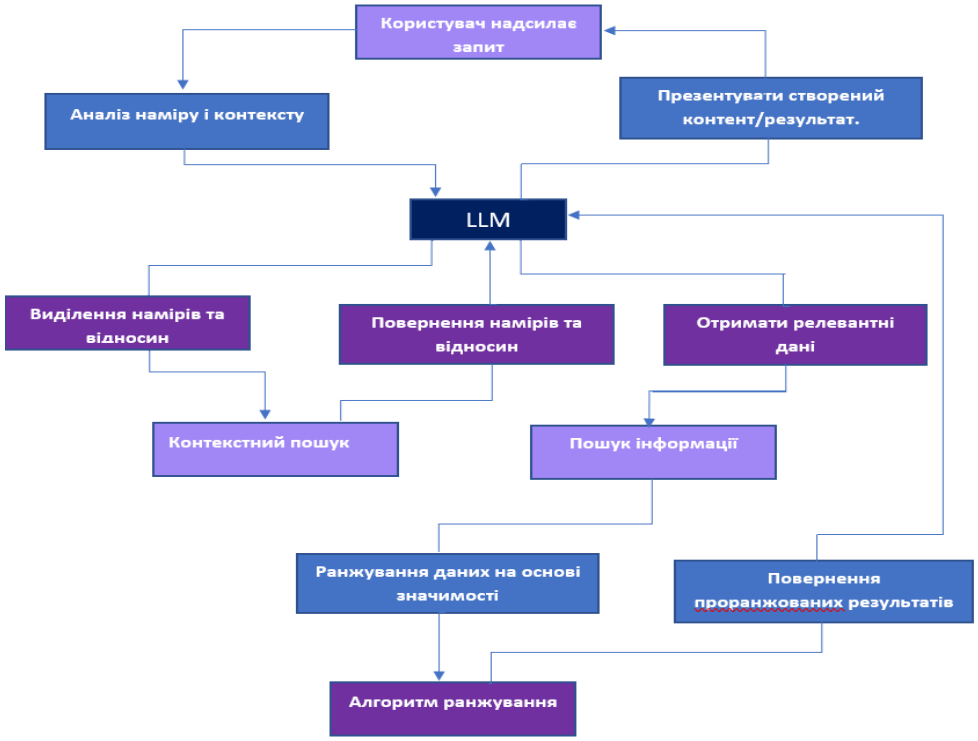

Рис. 6. Схема роботи контекстного пошуку

Ембеддінг

Для роботи додатку використовується *Retrieval Augmented Generation* (*RAG*). *RAG* – це метод роботи з великими мовними моделями, який дає змогу навчати їх речам, яким вони навчені не були, наприклад приватним документам чи нещодавно опублікованій інформації. Тобто, можливість додавати до контексту запиту до мовної моделі додаткову інформацію, на основі якої мовна модель може надати користувачу більш повну і точну відповідь. В випадку нашого додатку – це клієнтська база даних [8].

Спочатку користувач ініціює процес, задавши пошуковий запит. Далі *LLM* аналізує запит, щоб зрозуміти наміри користувача та контекст запиту. Контекстний пошук обробляє запит, щоб визначити зв'язки між термінами та загальний зміст запиту, які потім надсилаються назад до *LLM*. Після цього *LLM* отримує дані, які мають відношення до запиту. Далі алгоритм ранжування оцінює отримані дані з векторної бази даних, розміщує їх за порядком важливості у відповідності запиту. Відсортовані результати надсилаються назад до *LLM*, де генеруються результати пошуку для користувача (рис. 7).

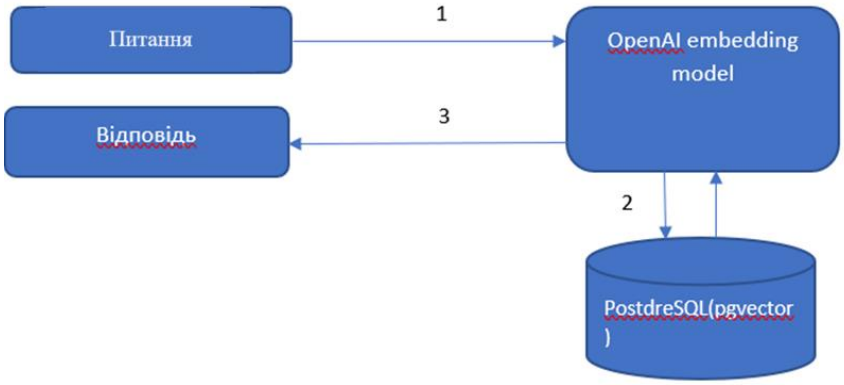

Рис. 7. Концепт роботи ембеддінгу

Ембеддінг – техніка за допомогою якої слова можуть бути представлені у вигляді векторів. В результаті слова з схожими значеннями, будуть мати схожі вектори. Тобто завдяки цьому можна будувати семантичні зв'язки між словами та фразам, що в свою чергу дозволяє машинам розуміти та оброблювати людську мову та давати людиноподібні відповіді. Для забезпечення точної та актуальної інформації створені вектори зберігають у векторній базі даних. *OpenAI* надає модель для ембеддінгу *Ada V2* [9].

PostgreSQL – це реляційна база даних з відкритим кодом, є однією з найвідоміших реляційних баз даних. *PostgreSQL* підтримує як реляційні, так і нереляційні запити. Для перетворення *PostgreSQL* в векторну базу даних, використовують розширення *pgvector* , яке допомагає працювати з векторами.

Принцип роботи додатку контекстного пошуку

Принцип роботи додатку контекстного пошуку – генерація векторів для бази знань за допомогою *OpenAI API* та пошук найбільш схожих відповідей за допомогою схожості по косинусу використовуючи *PostgreSQL*, та розширення *pgvector*. Створений проект має власну векторну базу даних, використовує *Open AI* через *REST Api* та надає власний *REST API* для роботи сторонніх чат-ботів (рис. 8).

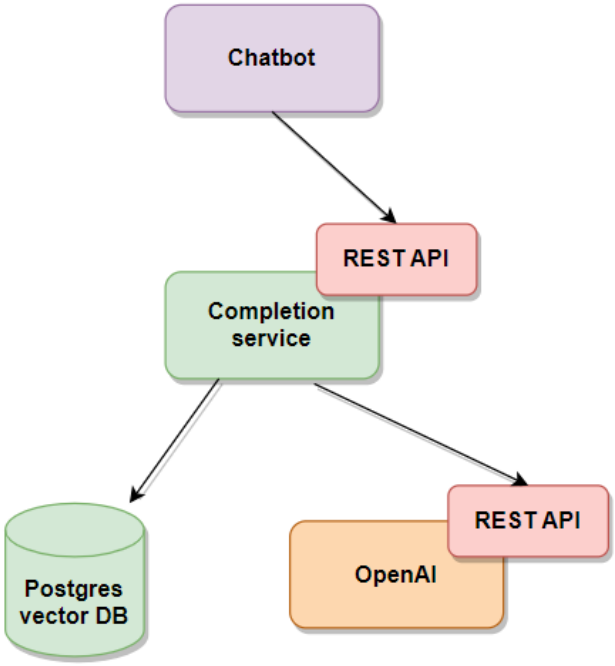

Рис. 8.Архітектура додатку контекстного пошуку

Через виклик *REST API* відправляються вхідні дані з яких, використовуючи надану *OpenAI* модель для ембеддінгу – *text-embedding-ada*-002, генеруються вектори. Створені вектори додаються до векторної бази даних *Postgres* та далі можуть використовуватись для подальшої обробки (рис. 9).

Отримання відповіді на запит відбувається наступним чином: після того як користувач надсилає повідомлення за допомогою чат-боту, на основі цього повідомлення генерується вектор за допомогою *embeddings API* від *OpenAI*. Отриманий вектор запиту користувача порівнюється з наявними в базі знань питаннями, та за допомогою косинусу подібності знаходяться *N* запитань з найбільш схожими векторами. Далі цей запит, разом з контекстом побудованим на основі знайдених схожих запитань надсилається до *OpenAI completion API* для отримання людино-подібної відповіді і вже отриманий результат надсилається користувачу (рис. 10).

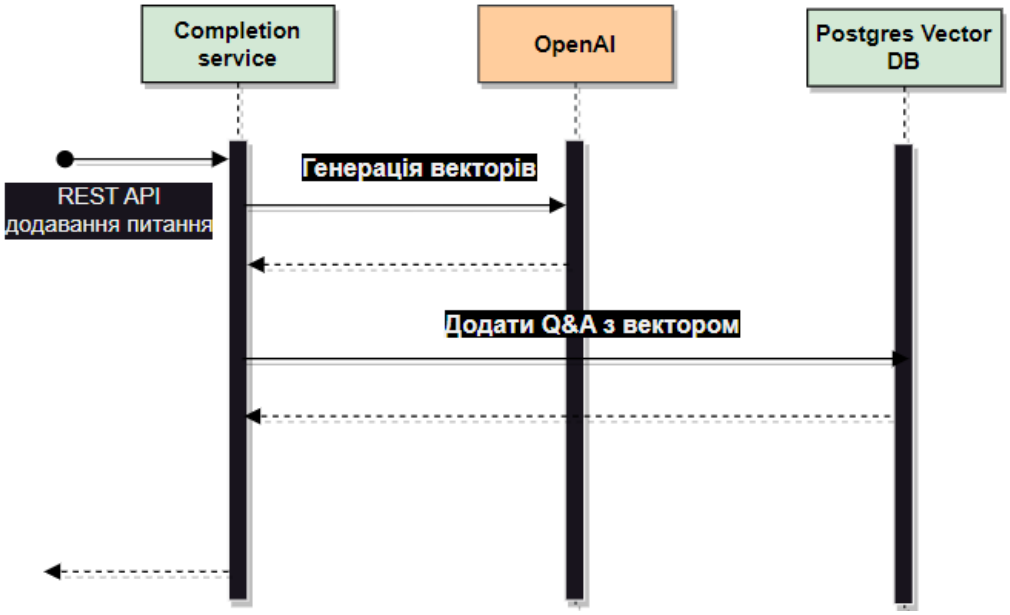

Рис. 9. Діаграма послідовності додавання питання до бази даних

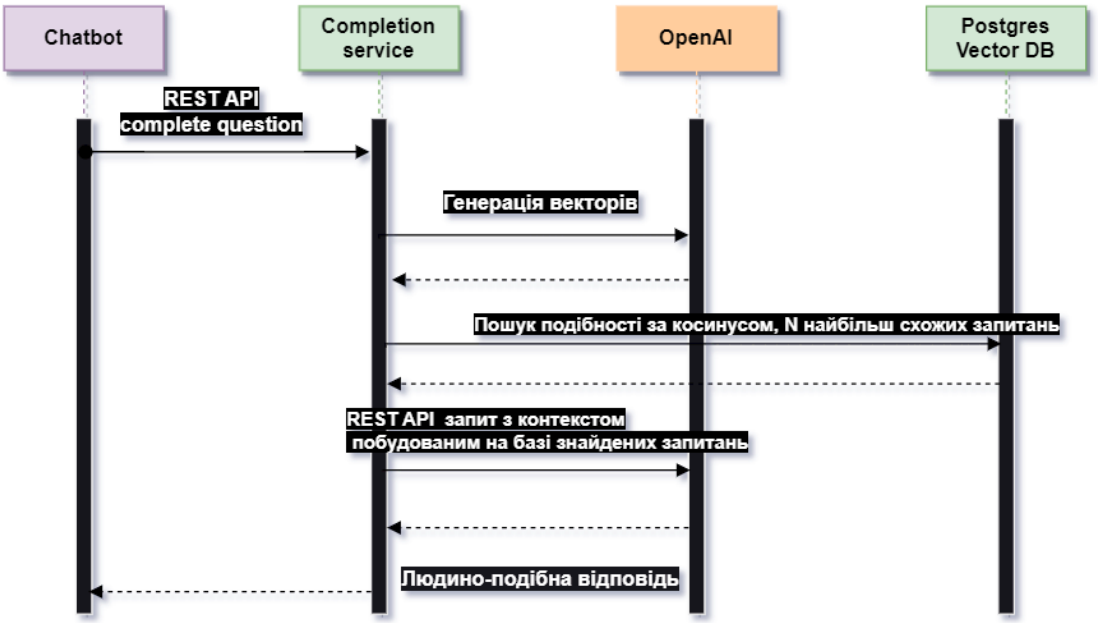

Рис. 10. Діаграма послідовності отримання відповіді

Висновки

У цій статті досліджено та проаналізовано значущий потенціал синергії косинус-подібності та технологій штучного інтелекту у контексті розробки контекстного пошуку на базі штучного інтелекту, пристосованого до мікро-сервісної архітектури. Використання косинусної подібності для оцінки семантичної близькості між текстовими даними є ефективним методом, особливо у поєднанні з генеративним штучним інтелектом.

Зокрема використання штучного інтелекту для перетворення даних в векторні представлення дозволяє отримувати зручне представлення для подальшого використання у вимірюванні схожості. Також поєднання штучного інтелекту та косинус-подібності дозволяє не просто проводити пошук, а й враховувати контекст і зв'язки між словами у запитах, що може поліпшити якість результатів пошуку для складних запитів або запитів з невизначеною семантикою.

Створений додаток свідчить про значний потенціал такого підходу до контекстного пошуку, що може сприяти поліпшенню швидкості та точності пошуку інформації для користувачів у великих обсягах даних. Додатки такого типу стрімко набувають популярності та мають широкий потенціал застосування в різноманітних галузях. Їх впровадження та розгортання дає змогу не тільки покращити процес взаємодії для користувачів, а й цілком змінити підхід до інтернет пошуку.

Подальші наукові дослідження та вдосконалення алгоритму можуть ще більше розкрити можливості поєднання косинус-подібності та штучного інтелекту в цьому напрямку.

Література

1. The value of getting personalization right–or wrong–is multiplying. URL: https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/the-value-of-getting-personalization-right-or-wrong-is-multiplying (date of access: 16.01.2024).

2. Artificial-intelligence-marke. URL: https://www.nextmsc.com/report/artificial-intelligence-market (date of access: 16.01.2024).

3. Chatbot Statistics: What Businesses Need to Know About Digital Assistants. URL: https://masterofcode.com/blog/ai-statistics (date of access: 16.01.2024).

4. How ChatGPT and Generative AI Will Alter the Future of Work. URL: https://www.aberdeen.com/blog-posts/how-chatgpt-and-generative-ai-will-alter-the-future-of-work/ (date of access: 16.01.2024).

5. The 2023 State of Digital Customer Experience Report. URL: https://www.verint.com/wp-content/uploads/2023-state-of-digital-cx-report.pdf (date of access: 16.01.2024).

6. Dataplot Reference Manual. Volume 2: LET Subcommands and Library.-National Institute of Standards and Technologie. URL: https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm (date of access: 17.01.2024).

7. What is semantic search? URL: https://www.elastic.co/what-is/semantic-search (date of access: 17.01.2024).

8. PostgreSQL as a Vector Database: Create, Store, and Query OpenAI Embeddings With pgvector. URL: https://www.timescale.com/blog/postgresql-as-a-vector-database-create-store-and-query-openai-embeddings-with-pgvector/ (date of access: 17.01.2024).

9. Introduction to Text Embeddings with the OpenAI API. URL: https://www.datacamp.com/tutorial/introduction-to-text-embeddings-with-the-open-ai-api (date of access: 07.02.2024).

Холявкіна Т.В., Чуба І.В., Шолупата А.Ю.

СИНЕРГІЯ КОСИНУС-ПОДІБНОСТІ ТА ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ

Насамперед, синергія косинус-подібності та генеративного штучного інтелекту в контексті розробки додатку контекстного пошуку є перспективним напрямком, особливо у зв'язку з тим, що більшість пошукових систем використовують пошук на основі ключових слів. Новий підхід, що пропонує штучний інтелект базується на нейронному і семантичному пошуку і він покладається не на окремі ключові слова, а на їх контекст та зв'язки між ними, що відкриває широкі можливості для покращення точності та релевантності результатів пошуку.

Для того, щоб задовольнити потреби користувачів створюють чат-ботів. Чат-боти використовуються в різних сферах, включаючи обслуговування клієнтів, консультування, підтримку користувачів та інші завдання, що вимагають взаємодії з людьми через текстовий чи голосовий інтерфейс. Їх популярність постійно зростає відповідно

до попиту від користувачів, більшість з яких регулярно використовує чат-ботів в повсякденному житті.

Стаття охоплює концепції штучного інтелекту, чат-ботів на основі штучного інтелекту, контекстного пошуку та косинусу подібності. Предметом розгляду є додаток контекстного пошуку на базі штучного інтелекту. Надається пояснення підходу та конкретних деталей реалізації його розробки.

***Ключові слова**: штучний інтелект; чат-бот; контекстний пошук; косинус подібності; вектор; ембеддінг; LLM; Java; PostgreSQL; pgvector; OpenAI API.*

Kholyavkina T.V., Chuba I.V., Sholupata A.Y.

SYNERGY OF COSINE SIMILARITY AND GENERATIVE ARTIFICIAL INTELLIGENCE

First and foremost, the synergy between cosine similarity and generative artificial intelligence in the context of developing a contextual search application is a promising direction, especially given that most search systems rely on keyword-based search. The new approach proposed by artificial intelligence is based on neural and semantic search, focusing not on individual keywords but on their context and relationships, which opens up broad possibilities for improving the accuracy and relevance of search results.

To meet users' needs, chatbots are being developed. Chatbots are utilized in various fields, including customer service, consultation, user support, and other tasks that require interaction with people through text or voice interfaces. Their popularity is constantly increasing in response to demand from users, the majority of whom regularly use chatbots in their everyday lives. The article covers concepts of artificial intelligence, AI-based chatbots, contextual search, and cosine similarity. The focus is on the application of contextual search based on artificial intelligence. An explanation of the approach and specific details of the implementation of its development is provided.

***Keywords**: artificial intelligence; chatbot; contextual search; cosine similarity; vector; embedding; LLM; Java; PostgreSQL; pgvector; OpenAI API.*