

Міністерство освіти і науки України  
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук  
(повна назва)

Кафедра Інформаційних управляючих систем  
(повна назва)

**КВАЛІФІКАЦІЙНА РОБОТА**  
**Пояснювальна записка**

рівень вищої освіти другий (магістерський)

Дослідження локальних мовних моделей для побудови чат-ботів в системах електронної комерції  
(тема)

Виконав:

студент 2 курсу, групи ІУСТМ-22-1

Безсонов Олександр Сергійович  
(прізвище, ім'я, по батькові)

Спеціальність 122 Комп'ютерні науки  
(код і повна назва спеціальності)

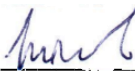
Тип програми освітньо-професійна  
(освітньо-професійна або освітньо-наукова)

Освітня програма Інформаційні  
управляючі системи та технології  
(повна назва освітньої програми)

Керівник Сергій Чалий  
(Власне ім'я ПРІЗВИЩЕ)

Допускається до захисту

Зав. кафедри

  
(підпис)

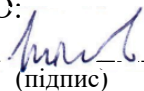
Костянтин ПЕТРОВ  
(Власне ім'я ПРІЗВИЩЕ)

2024 р.

## Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук  
(повна назва)Кафедра Інформаційних управляючих систем  
(повна назва)Рівень вищої освіти другий (магістерський)Спеціальність 122 Комп'ютерні науки  
(код і повна назва)Тип програми освітньо-професійна  
(освітньо-професійна або освітньо-наукова)Освітня програма Інформаційні управляючі системи та технології  
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри   
(підпис)«20» 11 2023 р.**ЗАВДАННЯ**  
НА КВАЛІФІКАЦІЙНУ РОБОТУстудентові Безсонову Олександр Сергійовичу  
(прізвище, ім'я, по батькові)

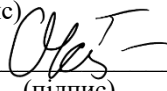
1. Тема роботи Дослідження локальних мовних моделей для побудови чат ботів в системах електронної комерції  
затверджена наказом університету від 16 11 2023 р. № 1359Ст
2. Термін подання студентом роботи до екзаменаційної комісії 20 01 2024 р.
3. Вихідні дані до роботи наукові дослідження та статті з методів побудови чат-ботів, методи адаптації локальних мовних моделей, методи оцінки локальних мовних моделей
4. Перелік питань, що потрібно опрацювати в роботі аналіз методів адаптації систем чат-ботів для систем електронної комерції, аналіз існуючих методів включення специфічних знань в систему чат-бота, удосконалення методу побудови контекстних даних для чат-бота, розробка інформаційної технології для використання локальних мовних моделей, опис інформаційної технології, імплементація інформаційної технології, програмна реалізація технології, експериментальна перевірка розробленої технології з удосконаленим методом

## КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання	20.11.2023	Виконано
2	Виконання аналізу методів побудови і методів оцінки пояснень	20.11.2023-28.11.2023	Виконано
3	Дослідження критеріїв оцінки пояснень	30.11.2023-05.12.2023	Виконано
4	Удосконалення можливісного методу оцінки пояснень	06.12.2023-11.12.2023	Виконано
5	Опис та імплементація інформаційної технології впровадження удосконаленого методу оцінки пояснень	12.12.2023-15.12.2023	Виконано
6	Програмна реалізація удосконаленого методу оцінки пояснень	16.12.2023-24.12.2023	Виконано
7	Експериментальна перевірка удосконаленого методу оцінки пояснень	25.12.2023-26.12.2023	Виконано
8	Оформлення пояснювальної записки до кваліфікаційної роботи	28.12.2023-06.01.2024	Виконано
9	Захист роботи	22.01.2024	Виконано

Дата видачі завдання 20 листопада 2023 р.

Студент   
(підпис)

Керівник роботи  — проф. каф. ІУС Чалий С. Ф.  
(підпис) (посада, прізвище, ініціали)

## РЕФЕРАТ

Робота містить: 85 сторінок, 27 рисунків, 2 таблиць, 1 додаток, 10 джерел.

ВЕЛИКІ МОВНІ МОДЕЛІ, ІНФОРМАЦІЙНА ІНТЕЛЕКТУАЛЬНА СИСТЕМА, ЛОКАЛЬНІ МОВНІ МОДЕЛІ, МАШИННЕ НАВЧАННЯ, ЧАТ-БОТИ.

У даній роботі виконано аналіз методів адаптації локальної мовних моделей для систем електронної комерції. Було досліджено властивості цих систем, їх здатності до генерації відповідей на запитання користувача. Досліджено існуючі методи побудови чат-ботів, що включає в себе оцінку їх ефективності. Розглянуто методи створення чат-ботів за архітектурою RAG (пошук та збільшення) та методи оцінки згенерованих відповідей. Окрема увага приділена методам включення специфічних знань в системи чат-ботів.

В роботі також представлено удосконалення методу побудови контекстних даних для чат-бота, що полягає в узагальненні вхідних документів для збільшення ефективності семантичного пошуку з подальшим використанням оригінальних фрагментів даних для запобігання втрат інформації.

Завершальна частина роботи присвячена програмній реалізації удосконаленого методу побудови контекстних даних для чат-бота, включаючи експериментальну перевірку його ефективності. Результати дослідження підтверджують значення удосконаленого методу побудови контекстних даних для чат-бота у поліпшенні релевантності згенерованих відповідей та користувацького досвіду в інформаційних системах електронної комерції.

## ABSTRACT

The work contains: 85 pages, 27 figures, 2 tables, 1 appendix, 10 sources.

LARGE LANGUAGE MODELS, INFORMATION INTELLIGENT SYSTEM, LOCAL LANGUAGE MODELS, MACHINE LEARNING, CHAT-BOTS.

This paper analyzes the methods of adaptation of local language models for the e-commerce system. The properties of these systems, their ability to generate answers to the user's questions were investigated. The existing methods of building chatbots have been studied, which includes an assessment of their effectiveness. Methods of creating chatbots for the RAG architecture (search and extension) and methods of evaluating the generated responses are considered. Particular attention is paid to the methods of incorporating special knowledge into chatbot systems.

The work also presents the improvement of the method of constructing contextual data for a chatbot, which arises from the summarization of input documents to increase the efficiency of semantic search with the subsequent use of original data fragments to prevent information loss.

The final part of the work is devoted to the software implementation of the improved method of building contextual data for the chatbot, including experimental verification of its effectiveness. The results of the study confirm the importance of an improved method of constructing contextual data for chatbots in improving the relevance of generated responses and user experience in e-commerce information systems.

## ЗМІСТ

Вступ.....	8
1. Аналіз предметної області.....	10
1.1 Аналіз структури та властивості систем електронної комерції.....	10
1.2 Аналіз задач чат ботів в системах електронної комерції.....	11
1.3 Дослідження локальних мовних моделей .....	17
1.4 Постановка задачі дослідження.....	23
2. Побудова контекстних даних для чат-бота на основі локальних мовних моделей.....	25
2.1 Розробка промптів мовної моделі для узагальнення .....	25
2.2 Удосконалення методу побудови контекстних даних для чат-бота.....	29
3. Технологія використання локальних мовних моделей для чат-ботів в системі електронної комерції.....	38
3.1 Підхід до вирішення проблеми обмежених даних локальної моделі.....	38
3.2 Модель зберігача (Retriever model) для етапу формування вбудовувань інформаційної технології .....	39
3.2 Семантичний пошук .....	40
4. Реалізація та експериментальна перевірка використання локальної мовної моделі для чат-бота .....	42
4.1 Вибір платформи для інтерфейсу користувача.....	42
4.2 Опис компонентів чат-бота.....	42
4.3 Імплементация RAG-чатботу.....	46
4.4 Оцінка відповідей чат-бота .....	53
Висновки .....	57
Перелік джерел посилання .....	58
Додаток А Графічний матеріал.....	60

## СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАЧКИ

ІС – інформаційна система

МН – машинне навчання

ШІ – штучний інтелект

ID – identity document (унікальна ознака об'єкта)

LLM – large language model (велика мовна модель)

RAG – retrieval and augmentation (пошук і збільшення)

## ВСТУП

У сучасній ері електронної комерції, де взаємодія з клієнтами набуває новітні форми та переходить у віртуальний простір, чат-боти виявляються ключовим інструментом для забезпечення ефективного та природного спілкування з покупцями. Враховуючи зростання інтересу споживачів до онлайн-покупок та інтерактивних послуг, постійне вдосконалення інтерфейсів та інтерактивних аспектів електронної комерції стає невід'ємною частиною конкурентного середовища.

Магістерська робота покликана вивчити та вдосконалити локальні мовні моделі для оптимізації функціонування чат-ботів у системах електронної комерції. Специфіка мовлення та взаємодії у цьому контексті вимагає глибокого розуміння потреб споживачів та вивчення їхніх уподобань для забезпечення ефективної та інтуїтивно зрозумілої взаємодії. Розглядатимуться аспекти автоматизованого відповідання на запитання, персоналізованого обслуговування та вдосконалення алгоритмів обробки мовленнєвих даних для максимізації користувацького досвіду у сфері електронної комерції.

Мета роботи полягає в розробці підходу до створення та управління чат-ботами, які відповідають вимогам сучасного електронного бізнесу. Дослідження локальних мовних моделей та їх підвищення шляхом оптимізації семантичного аналізу тексту спрямовані на забезпечення найвищого ступеня релевантності та точності відповідей чат-ботів. Розробка інформаційної технології для впровадження вдосконаленого методу та програмна реалізація експерименту із запитаннями в реальних умовах дозволять не лише перевірити практичну значущість розробленого підходу, а й визначити його ефективність та надійність у сучасних умовах електронної комерції.

У роботі описано удосконалений метод побудови контекстних даних для чат-бота, який дозволяє позбавитися шуму в даних для більш ефективного

вибору контекстних даних. Метод сприяє генерації більш релевантних відповідей чат-бота.

## 1 АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ

### 1.1 Аналіз структури та властивості систем електронної комерції

Електронна комерція, або e-commerce, визначається як процес купівлі та продажу товарів чи послуг через Інтернет. Це електронна форма торгівлі, де бізнес-операції здійснюються за допомогою електронних засобів зв'язку, таких як комп'ютери, смартфони чи інші мобільні пристрої [1].

Основні характеристики електронної комерції включають:

– онлайн-платформи: товари та послуги продаються через веб-сайти або мобільні додатки, які надають доступ клієнтам з будь-якого місця, де є Інтернет;

– електронні платежі: оплата здійснюється електронними засобами, такими як кредитні картки, електронні гаманці чи інші онлайн-платіжні системи;

– взаємодія з клієнтами: комунікація між покупцями та продавцями часто здійснюється через електронні канали, такі як чат-боти, електронна пошта, або системи онлайн-зворотного зв'язку;

– онлайн-маркетинг: Застосування цифрових маркетингових стратегій для просування товарів та залучення уваги покупців;

– автоматизація бізнес-процесів: Використання технологій для оптимізації управління запасами, обробки замовлень, та інших бізнес-процесів;

– Електронна комерція дозволяє підприємствам розширити свій ринок, забезпечити зручність для покупців та оптимізувати бізнес-процеси через використання технологій та Інтернет-ресурсів.

Переваги використання електронної комерції [1]:

– зниження вартості – відомо, що у деяких інтернет-магазинах ціни можуть бути вищими, ніж у традиційних магазинах. Онлайн-продавцям не потрібна фізична торгова площа та велика кількість персоналу, яка доглядає

за електронним комерційним веб-сайтом. Це призводить до зниження їхніх загальних витрат порівняно з тими у класичних продавців, і саме це є основною причиною, що дозволяє їм пропонувати споживачам більш доступні ціни на товари та послуги, а також забезпечувати вигідні умови доставки [2];

- швидше і легше купувати – за декілька хвилин клієнт може замовити потрібний товар чи послуги та отримати їх на свою домашню адресу;

- можливість роботи 24/7 – перевага електронної комерції для всіх клієнтів, порівняно із традиційними магазинами, які мають фіксований робочий графік [2];

- доступність інформації – при пошуку на сайтах електронної комерції можливо знайти різну інформацію про товари та послуги продавців: усі продукти, які вони пропонують, усі розміри та кольори продуктів, які вони пропонують, години роботи їхніх звичайних магазинів;

- різноманіття товарів і послуг на електронних комерційних сайтах зазвичай значно перевищує те, що можна знайти в класичних магазинах, охоплюючи широкий спектр розмірів, кольорів та інших характеристик.;

- створення бази клієнтів – при реєстрації в інтернет-магазині відвідувачам зазвичай пропонують ввести дані про себе. Використовуючи зібрані дані, продавець може ефективно пристосовувати свою пропозицію під потреби потенційного покупця та утримувати зв'язок з ним [2].

Діяльність підприємства регламентовано наступними нормативними документами [3]:

- закон України "Про електронну комерцію" №675";
- закон України "Про захист прав споживачів";
- закон України "Про інформацію";
- закон України "Про платіжні послуги";
- закон України "Про рекламу";
- закон України "Про захист інформації в інформаційно-телекомунікаційних системах";

– правила роздрібної торгівлі продовольчими товарами, затвержені наказом Мінекономіки України.

Головні загальні риси електронної комерції включають:

- операції купівлі-продажу здійснюються в електронному вигляді;
- інтернет є основним каналом для здійснення транзакцій (включаючи оплату) і спілкування;
- доставка товарів або послуг може, але не обов'язково, відбуватися онлайн.

Можна знайти й інші більш-менш подібні визначення електронної комерції у джерельній літературі, що впливає з потреби адаптувати концепції до сьогодення тенденції глобалізації та бажання споживачів – головна причина, однак, є невинний і швидкий розвиток нових технологій. Зазначені подібності не виключають відмінностей і неточностей у визначеннях електронної комерції, і багато розбіжностей є результатом плутанини або поєднання понять електронної комерції та електронного бізнесу та інші терміни, близькі за значенням.

У зв'язку з профілем їх діяльності, каналами зв'язку, продажі та завершення транзакцій, компанії електронної комерції створюють, приймають, і використовувати величезну кількість інформації. Порівняння інформації та потоки даних представлені на рисунку 1.1.

Використання інтелектуальних чат-ботів та традиційної комунікації з клієнтами в електронній комерції представляє суттєві відмінності.

Інтелектуальні чат-боти вирізняються автоматизованою та миттєвою відповіддю на запитання клієнтів. Вони дозволяють забезпечити цілодобову підтримку, оптимізують процес оформлення замовлень, а також здатні аналізувати дані для надання персоналізованих рекомендацій. Використання чат-ботів спрощує взаємодію та забезпечує ефективність обслуговування клієнтів [4].

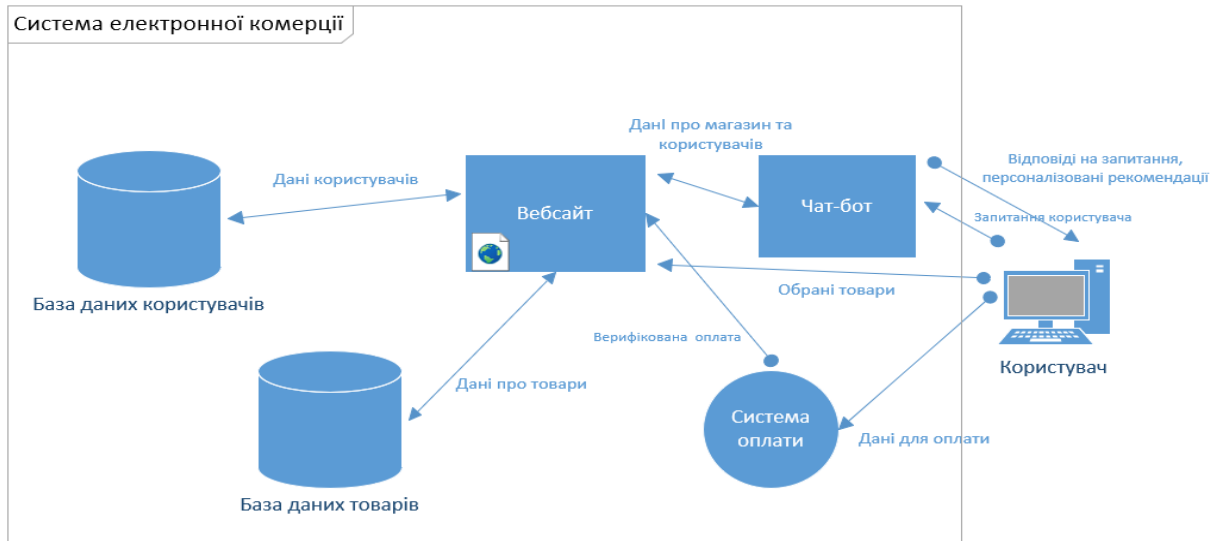


Рисунок 1.1– Взаємодія компонентів системи електронної комерції

## 1.2 Аналіз задач чат ботів в системах електронної комерції

У той час як традиційна комунікація передбачає прямий контакт з людьми, що може призводити до більш гнучкої та адаптивної відповіді на унікальні ситуації. Однак вона обмежена робочими годинами та може вимагати значних людських ресурсів.

В обох випадках важливо збалансувати автоматизацію та людський елемент для досягнення найкращих результатів в обслуговуванні клієнтів та оптимізації бізнес-процесів.

Чат-боти можуть вирішувати різноманітні завдання в електронній комерції, спрямовані на поліпшення користувацького досвіду, автоматизацію процесів та збільшення продажів. Нижче наведено перелік основних задач чат-ботів в системах електронної комерції (рисунок 1.2).

#### підтримка комунікації з користувачами:

- чат-боти можуть відповідати на запитання клієнтів щодо продуктів, послуг, цін, графіку роботи тощо;
- вони можуть надавати консультації та рекомендації користувачам.

замовлення та оплата – чат-боти допомагають клієнтам зробити замовлення та оплатити їх, що зменшує кількість кроків для здійснення покупки;

- вони можуть надавати інформацію про різні способи оплати та доставки.

#### відстеження замовлень:

- чат-боти можуть надсилати користувачам статуси їхніх замовлень та інформацію про доставку;
- вони дозволяють клієнтам отримувати актуальну інформацію про свої покупки.

#### рекомендації та персоналізація:

- чат-боти можуть аналізувати попередні покупки та інші дані користувачів для надання рекомендацій щодо подібних товарів або акцій;
- вони допомагають створити персоналізований досвід для кожного користувача.

#### вирішення проблем користувачів та повернення товарів:

- чат-боти можуть надавати інструкції та допомогу у вирішенні проблем з доставкою, товарами або поверненням товарів;
- вони можуть запускати процеси повернення та обміну.

#### маркетинг та реклама:

- чат-боти можуть повідомляти користувачів про нові товари, акції та знижки;
- вони можуть проводити опитування та збирати відгуки від клієнтів.

#### аналіз даних та звітність:

Рисунок 1.2 – Можливі варіанти використання чат-бота

В основі алгоритмів роботи чат-ботів може бути:

- набір правил;
- машинне навчання.

Жорсткий набір правил - саме такий швидкий і простий спосіб автоматизації спілкування з клієнтами використовує більшість компаній.

Інструменти цього типу витягають із реплік ключових слів користувача та на їх основі формують відповідь із заданих заздалегідь. Якщо в тексті не буде ключових слів, алгоритм не зможе відповісти і переключити розмову зі співробітником. Тому в такі програми часто встановлюють спеціальні кнопки з варіантами запитів: так користувач буде гарантовано прийнятий.

Машинне навчання - це більш продвинуті інструменти, зі штучним інтелектом і складним алгоритмом роботи. Вони уміють інтерпретувати репліки навіть без ключових слів. Відповідь готується за допомогою генеративних мереж, що навчаються або евристичних методів. Завдяки НЛП-технологіям боти підтримують природний діалог. Можливі діалоги з користувачем наведено на рисунку 1.3.

#### **Деталі доставки:**

- *Клієнт:* "Які варіанти доставки ви пропонуєте для мого регіону?"
- *Чатбот:* "В нашому магазині доступні різні варіанти доставки: стандартна, експрес, та кур'єрська. Де саме ви знаходитесь, щоб я міг надати більше інформації?"

#### **Питання про терміни доставки:**

- *Клієнт:* "Скільки часу триває стандартна доставка?"
- *Чатбот:* "Зазвичай, стандартна доставка займає від 3 до 5 робочих днів. Однак, точні терміни можуть залежати від вашого місця розташування. Ви бажаєте знати терміни для конкретного регіону?"

#### **Питання про повернення товару:**

- *Клієнт:* "Як я можу повернути товар, який мені не підійшов?"
- *Чатбот:* "Шкода, що ви не задоволені покупкою. Для повернення товару будь ласка, зверніться до нашого сервісу підтримки. Якщо у вас є конкретні питання про процес повернення, я готовий вам допомогти."

#### **Питання про оплату:**

- *Клієнт:* "Чи можливо оплатити замовлення при отриманні?"
- *Чатбот:* "На жаль, ми не надаємо оплату при отриманні. Проте, ми приймаємо різні види онлайн-оплати, такі як кредитні картки та електронні платіжні системи. Якщо у вас є конкретні питання щодо оплати, дайте мені знати."

Рисунок 1.3 – Варіанти питань до чат-бота і відповідей на них

Мета інструмента — з'ясувати проблеми клієнта, а потім допомогти їх вирішити. Послідовність роботи чат-бота:

- після звернення користувач задає питання – формулює сам або вибирає з готових;
- програма поділить репліку користувача на окремі слова і шукає серед них ключові;
- якщо це передбачено сценарієм, чат-бот за допомогою нейросети класифікує питання, визначає намір користувача або його наміри;
- програма вилучає з фраз сутності, необхідні для генерації відповіді, наприклад, дати, сім'ю, місто. При необхідності задає уточнюючі питання;
- далі інструмент шукає відповідь на питання користувача в базі знань або, генерує його за допомогою навчальної нейросемережі;
- при необхідності програма через API звертається до підсистем, розміщених на сторонніх серверах, і просить виконати визначені вичислення;
- користувачу показується відповідь за допомогою веб інтерфейсу (рисунок 1.4).

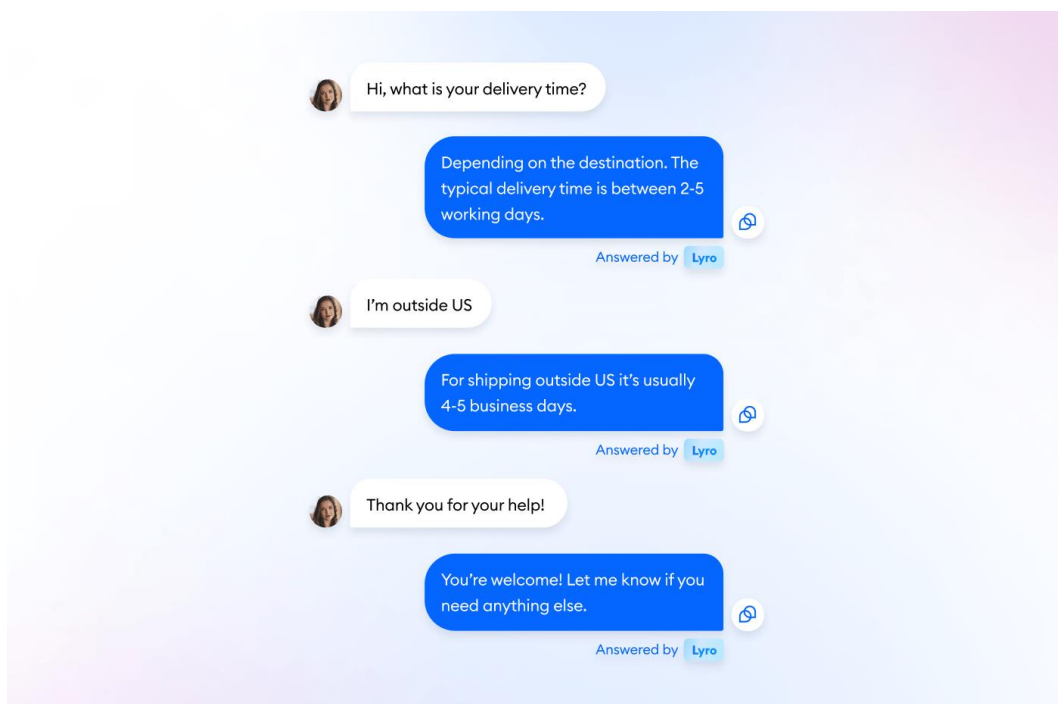


Рисунок 1.4 – Приклад діалогу з користувачем у веб-інтерфейсі

### 1.3 Дослідження локальних мовних моделей

LLM – це великі мовні моделі на основі нейронних мереж, навчені на величезній кількості даних, щоб імітувати людську поведінку, виконуючи різноманітні подальші завдання, як-от відповіді на запитання, переклад і резюмування. LLM підірвали світ із запровадженням таких інструментів, як ChatGPT і GitHub Copilot [7].

Локальні мовні моделі – це LLM з відкритим вихідним кодом відрізняються від своїх закритих аналогів вихідним кодом (а іноді й іншими компонентами). У закритих LLM вихідний код, який пояснює, як структурована модель і як працюють алгоритми навчання, не публікується.

Локальні LLM з відкритим вихідним кодом допомагають галузі в цілому: оскільки так багато людей роблять внески, їх можна розробляти швидше, ніж закриті моделі. Вони також можуть бути більш ефективними для крайніх випадків або конкретних програм (наприклад, підтримка місцевої мови).

Таблиця 1.1 – Порівняння локальних LLM

Модель	Розробник	Рік Виходу	Кількість Параметрів	Архітектура	Задачі, для яких призначена	Комерційне використання
OpenLLaMA	Berkeley AI Research	2023	3 млрд, 7 млрд, 13 млрд	Transformer	Загальні задачі обробки природної мови	Так
BERT	Google	2018	340 млн	Transformer	Пошукові системи, витягування інформації	Ні

Кінець таблиці 1.1

Модель	Розробник	Рік Виходу	Кількість Параметрів	Архітектура	Задачі, для яких призначена	Комерційне використання
Falcon	Інститут технологічних інновацій Абу-Дабі	2023	7 млрд, 40 млрд	Transformer	Falcon може генерувати людський текст, перекладати мовами та відповідати на запитання	Так
MPT	MosaicML	2023	7 млрд, 30 млрд	Transformer	Для віртуальних помічників, чат-ботів та інших інтерактивних інструментів	Так
FastChat-T5	FastChat	2023	3 млрд	Transformer	Підходить для додатків, що потребують розуміння мови, як-от віртуальні помічники, системи підтримки клієнтів та інтерактивні платформи.	Так

GPT-2	OpenAI	2019	1.5 млрд	Transformer	Загальні задачі	Ні
-------	--------	------	----------	-------------	-----------------	----

Але закриті моделі, які часто створюють великі компанії, також мають переваги. По-перше, вони вбудовані в системи з фільтрами для упередженої інформації, ненормативної лексики та іншого сумнівного вмісту. Вони також часто мають заходи безпеки. Крім того, вони не потребують тонкого налаштування, спеціального набору навичок, які вимагають відданих людей і команд.

Багато компаній експериментують із ChatGPT та іншими великими моделями мови чи зображень. Загалом вони вважали їх вражаючими з точки зору їхньої здатності висловлювати складні ідеї чіткою мовою. Однак більшість користувачів усвідомлюють, що ці системи насамперед навчаються на основі інформації в Інтернеті та не можуть відповідати на підказки чи запитання щодо конфіденційного вмісту чи знань. [7]

Використовувати знання компанії про пристойність є критично важливим для її здатності конкурувати та впроваджувати інновації, особливо в сучасному нестабільному середовищі. Організаційні інновації підживлюються ефективним і гнучким створенням, управлінням, застосуванням, рекомбінацією та розгортанням ресурсів знань і ноу-хау. Однак знання в організаціях зазвичай генеруються та фіксуються з різних джерел і форм, включаючи індивідуальні думки, процеси, політики, звіти, операційні транзакції, дошки обговорень, а також онлайн-чати та зустрічі. Таким чином, комплексні знання компанії часто не враховуються, і їх важко організувати та розгорнути там, де це необхідно.

Технологія включення знань певної галузі організації в LLM швидко розвивається. На даний момент існує три основні підходи до включення власного вмісту в генеративну модель: навчання LLM з нуля, точне налаштування існуючого LLM та оперативна настройка існуючого LLM.

Перший із підходів полягає у створенні та навчанні власної предметно-спеціальної моделі з нуля. Це не поширений підхід, оскільки для навчання

великої мовної моделі потрібна величезна кількість високоякісних даних, а більшість компаній просто не мають цього. Це також вимагає доступу до значної обчислювальної потужності та добре навченого таланту в галузі обробки даних.

Однією з компаній, яка застосувала цей підхід, є Bloomberg, яка нещодавно оголосила, що створила BloombergGPT для фінансового контенту та природного мовного інтерфейсу з терміналом даних. Bloomberg має фінансові дані, новини та документи за понад 40 років, які він об'єднав із великим обсягом тексту з фінансових декларацій та даних Інтернету. Загалом дослідники даних Bloomberg використали 700 мільярдів токенів, або близько 350 мільярдів слів, 50 мільярдів параметрів і 1,3 мільйона годин графічного процесора. Небагато компаній мають такі ресурси.

Другий підхід полягає в «тонкому налаштуванні» навчання існуючого LLM для додавання конкретного доменного вмісту до системи, яка вже навчена загальним знанням і взаємодії на основі мови. Цей підхід передбачає коригування деяких параметрів базової моделі та зазвичай потребує значно менше даних — зазвичай лише сотні чи тисячі документів, а не мільйони чи мільярди — і менше часу на обчислення, ніж створення нової моделі з нуля.

Google використав точне налаштування навчання на своїй моделі Med-PaLM2 (друга версія) для медичних знань. Дослідницький проект розпочався із загального PaLM2 LLM від Google і перепідготував його на ретельно відібраних медичних знаннях із різноманітних публічних наборів медичних даних. Модель змогла відповісти на 85% запитань іспиту на медичну ліцензію США — майже на 20% краще, ніж перша версія системи. Незважаючи на цей швидкий прогрес, під час тестування за такими критеріями, як наукова достовірність, точність, медичний консенсус, аргументація, упередженість і шкода, а також оцінювання експертами з багатьох країн, команда розробників визнала, що система все ще потребує суттєвого вдосконалення перед тим, як її прийняти для клінічних досліджень. практика.

Підхід до тонкого налаштування має деякі обмеження. Незважаючи на те, що навчання вимагає набагато менше обчислювальної потужності та часу, ніж навчання з нуля, навчання все одно може бути дорогим, що не було проблемою для Google, але було б для багатьох інших компаній. Це вимагає значної експертизи в галузі даних; наукова стаття для проекту Google, мала 31 співавтора. Деякі дослідники даних стверджують, що він найкраще підходить не для додавання нового вмісту, а для додавання нових форматів і стилів вмісту (таких як чат або написання, як у Вільяма Шекспіра). Крім того, деякі постачальники LLM (такі як OpenAI) не дозволяють тонке налаштування своїх останніх LLM, наприклад GPT-4.

Третім та найпоширенішим підходом до налаштування вмісту LLM для компаній-постачальників нехмарних технологій є його налаштування за допомогою підказок. За допомогою цього підходу вихідна модель залишається замороженою та змінюється за допомогою підказок у контекстному вікні, які містять предметно-спеціальні знання. Після швидкого налаштування модель може відповісти на запитання, пов'язані з цими знаннями. Цей підхід є найбільш обчислювально ефективним із трьох, і він не вимагає величезного обсягу даних для навчання на новому домені вмісту.

Morgan Stanley, використовував оперативне налаштування для навчання моделі GPT-4 OpenAI, використовуючи ретельно підібраний набір із 100 000 документів із важливими знаннями про інвестування, загальний бізнес та інвестиційний процес. Мета полягала в тому, щоб надати фінансовим консультантам компанії точні та легкодоступні знання щодо ключових питань, з якими вони стикаються під час консультування клієнтів. Система швидкого навчання працює в приватній хмарі, доступ до якої мають лише співробітники Morgan Stanley.

Хоча це і найпростіший із трьох підходів для організації, він не позбавлений технічних проблем. У разі використання неструктурованих даних, як-от текст, як вхідних даних для LLM, дані, ймовірно, будуть занадто великими з надто великою кількістю важливих атрибутів, щоб вводити їх

безпосередньо в контекстне вікно для LLM. Альтернативою є створення векторних вбудовувань — масивів числових значень, створених із тексту за допомогою іншої попередньо навченої моделі машинного навчання. Така архітектура називається RAG — Retrieval Augmented Generation. Векторні вкладення є більш компактним представленням цих даних, яке зберігає контекстуальні зв'язки в тексті. Коли користувач вводить підказку в систему, алгоритм подібності визначає, які вектори слід надіслати в модель GPT [8]

RAG і тонке налаштування є методами адаптації попередньо навчених мовних моделей до конкретних завдань або доменів. Порівняння цих двох підходів наведено в таблиці 1.2

Таблиця 1.2 – Порівняння підходів включення знань в LLM

Характеристика	Retrieval Augmented Generation (RAG)	Fine-tuning
Визначення	Поєднує широкомасштабний пошук знань із генерацією послідовності. Отримує відповідні документи та генерує відповідь, використовуючи їх.	Удосконалює попередньо навчену модель для конкретних завдань, використовуючи менший набір даних. Регулює вагу моделі, щоб спеціалізувати її для певного завдання.
Переваги	Може використовувати величезні зовнішні знання.	Може досягти високих результатів у конкретних завданнях. Ефективно, особливо коли для виконання завдання доступні обмежені дані.

Кінець таблиці 1.2

Характеристика	Retrieval Augmented Generation (RAG)	Fine-tuning
Можливі складності	<p>Потрібен ефективний механізм пошуку.</p> <p>Можливість отримати нерелевантні документи.</p> <p>Обчислювально більш інтенсивний.</p>	<p>Ризик переобладнання, якщо доступно недостатньо даних або якщо навчання є занадто агресивним.</p> <p>Знання обмежені до останнього відрізка навчання</p>
Сфера застосування	<p>Відповіді на питання відкритого домену (QA).</p> <p>Динамічні відповіді в чат-ботах.</p> <p>Ситуації, коли часто з'являються нові дані.</p>	<p>Спеціальні програми, такі як аналіз настроїв,</p> <p>Спеціальні домени з унікальними наборами даних.</p>
Приклади	<p>Модель OpenAI RAG для контролю якості.</p>	<p>Точне налаштування моделей GPT для конкретних доменів або завдань.</p>

## 1.4 Постановка задачі дослідження

Об'єктом дослідження є процес адаптації локальних систем чат-ботів для систем електронної комерції.

Предметом дослідження виступають методи побудови контекстних даних для чат-ботів.

Метою даної роботи є проведення дослідження та аналіз локальних мовних моделей з метою їхнього використання для побудови чат-ботів в системах електронної комерції. Робота спрямована на вдосконалення ефективності та ефективності чат-ботів, призначених для взаємодії з користувачами в контексті електронної комерції.

Науковою новизна роботи полягає в удосконаленні методу побудови контекстних даних для чат-ботів. Це передбачає узагальнення вхідних даних для підвищення ймовірності знаходження релевантних фрагментів за допомогою семантичного пошуку. Запропоноване удосконалення підвищує якість згенерованих відповідей, надаючи користувачу достовірні дані у зрозумілому вигляді.

Задачі роботи:

- аналіз методів адаптації систем чат-ботів для систем електронної комерції;
- аналіз існуючих методів включення специфічних знань в систему чат-бота;
- удосконалення методу побудови контекстних даних для чат-бота;
- розробка інформаційної технології для використання локальних мовних моделей, що надає можливість ефективно використати удосконалений метод побудови контекстних даних для чат-бота;
- програмна реалізація результатів теоретичних досліджень для практичного використання в системах електронної комерції;
- оцінка ефективності чат-боту для системи електронної комерції.

## 2. ПОБУДОВА КОНТЕКСТНИХ ДАНИХ ДЛЯ ЧАТ-БОТА НА ОСНОВІ ЛОКАЛЬНИХ МОВНИХ МОДЕЛЕЙ

2.1 Підхід до вирішення проблеми обмежених знань локальної мовної моделі

LLM мають обмеження — вони попередньо навчаються загальнодоступній інформації з Інтернету та не мають доступу до приватних даних, таких, як наприклад регламент роботи інтернет магазину в галузі електронної комерції. Саме тут вступає в дію архітектура Retrieval Augmented Generation (RAG), яка використовує векторну базу даних для доповнення LLM відповідними фрагментами приватних даних. Діаграма послідовності для системи RAG наведена на рисунку 2.1.

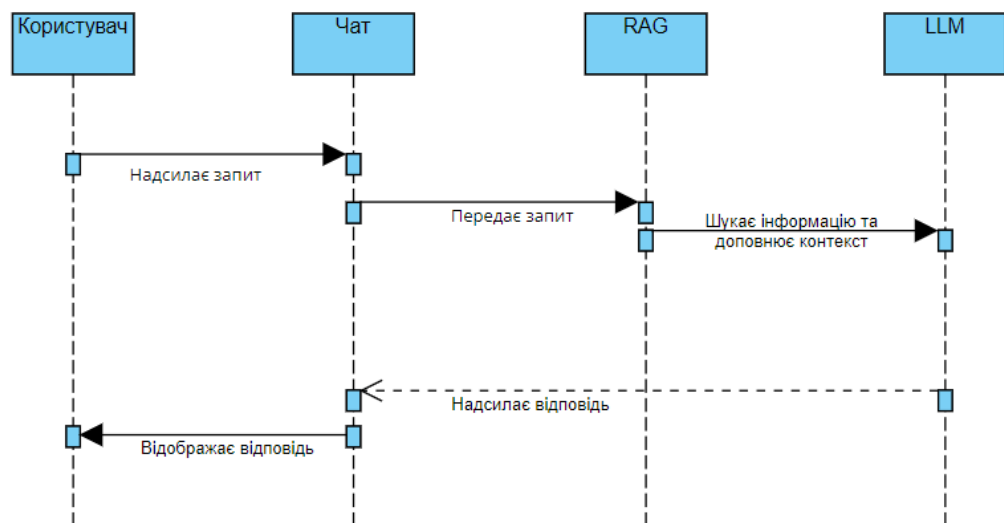


Рисунок 2.1 – Діаграма послідовності для RAG

RAG усуває прогалину в знаннях, динамічно отримуючи релевантну інформацію із зовнішніх джерел, гарантуючи, що згенеровані відповіді є не лише фактичними, але й актуальними. Векторні бази даних відіграють невід’ємну роль у механізмі RAG, забезпечуючи ефективний і семантичний пошук інформації. Ці бази даних зберігають інформацію як вектори, що

дозволяє RAG швидко й точно отримувати найбільш релевантні документи або точки даних на основі семантичної подібності вхідного запиту, підвищуючи точність і релевантність згенерованих LLM відповідей [5].

У загальному вигляді роботу взаємодію компонентів RAG-системи можна описати так:

- користувач вводить запит (запитання), що являє собою вхідні дані;
- зберігач кодує запит —  $q(x)$ . Це передбачає перетворення запиту на числовий вектор за допомогою моделі вбудовування (Embedding model). Цей вектор є згорнутим числовим представленням запиту;
- на основі закодованого запиту зберігач ( $P_{\pi}$ ) здійснює пошук в базі даних за вектором запиту, щоб знайти тісно пов'язані розділи документа, що досягається шляхом обчислення подібності на основі «відстані» між вектором запитання та векторами документів у базі даних. Тобто зберігач буде шукати топ- $K$  найближчих документів за допомогою пошуку максимального внутрішнього добутку (MIPS);
- отримані документи та запит будуть передані генератору ( $P_{\theta}$ ) для створення виводу. Генератор отримує запит разом із найбільш релевантними розділами з документів. Використовуючи цю інформацію, він генерує послідовну та відповідну контексту відповідь. Ефективна оперативна інженерія має важливе значення для спрямування ємовної моделі до більш точних результатів;
- система надає згенеровану відповідь користувачу.

Формула моделі [6]:

$$\begin{aligned}
 (p_{\text{RAG-Sequence}}(y | x)) &\approx \sum_{z \in \text{top-}k(p(\cdot | x))} p_{\pi}(z | x) p_{\theta}(y | x, z) \\
 &= \sum_{z \in \text{top-}k(p(\cdot | x))} p_{\pi}(z | x) \prod_i^N p_{\theta}(y_i | x, z, y_{1:i-1})
 \end{aligned}$$

Детальна схема роботи з прикладами даних наведена на рисунку 2.2.

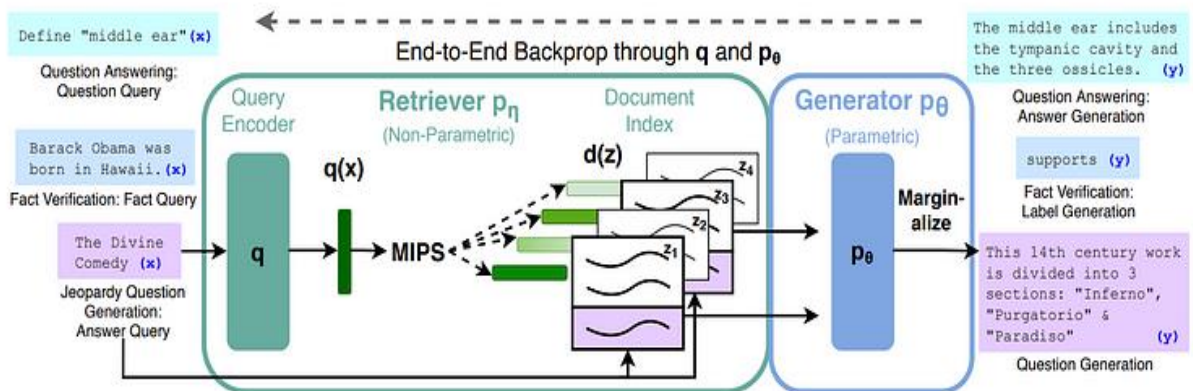


Рисунок 2.2 – Формування відповіді за допомогою RAG-системи

Такий підхід значно підвищує економічну ефективність операцій LLM. Основною причиною є легкість оновлення індексів пошуку на відміну від більш ресурсомісткого процесу постійного тонкого налаштування попередньо навчених моделей.

Крім того, впровадження RAG спрощує доступ до поточних даних. Це призводить до економії часу та коштів, роблячи роботу з актуальною інформацією більш практичною та ефективною в LLM.

Ще одною перевагою є те, що Retrieval Augmented Generation дозволяє уникати помилок або «галюцинацій», оскільки відповіді моделей базуються на контексті, отриманому із зовнішніх джерел даних.

Галюцинації мовних моделей відносяться до явищ, коли штучні інтелектуальні системи, такі як мовні моделі, створюють контент, який здається правдоподібним, але на самому місці не відповідає реальним даним або може бути неправильним, неточним або навіть фантастичним.

Ці галюцинації можуть виникнути із-за обмежених навчальних даних, неповного представлення про світ, а також із-за випадкових шумів або артефактів у процесі навчання.

В системі Retrieval Augmented Generation разом із контекстом модель завжди отримує промпт з чіткою підказкою, що також запобігає виникненню «галюцинацій».

Спеціальний промпт питання-відповідь для RAG працює наступним чином: підказка вказує LLM використати фрагмент інформації («контекст»), щоб відповісти на запитання, з додатковими вказівками, щоб запобігти вигадування некоректної відповіді генеративною моделлю. Запитання міститься внизу, а потім інструкція, яка просить LLM надати коротку відповідь.

Приклад схеми промпту із запитанням та контекстом наведено на рисунку 2.2.

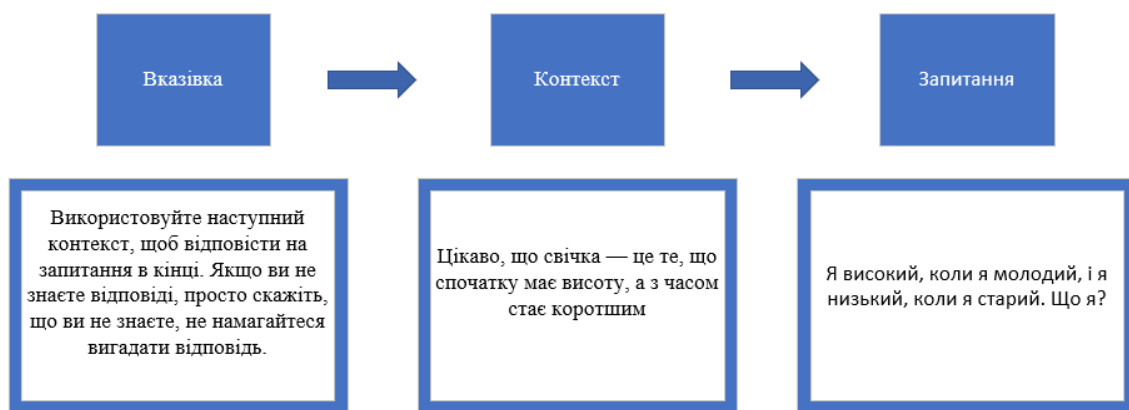


Рисунок 2.2 – Приклад вхідних даних для великої мовної моделі

Враховуючи цей контекст та інструкції, LLM може легко знайти відповідь на загадку та видає відповідь: "ти свічка". Слід зазначити, що таке завдання вимагає великого розуміння контексту і може бути вирішене лише за допомогою технології штучних нейронних мереж.

У контексті генерації тексту основною розглядуваною справою є вибір мовної моделі (ММ). Існує безліч варіантів, таких як від OpenAI, Cohere, Facebook та численні відкриті альтернативи. Крім того, кілька фреймворків ММ, таких як LangChain, LlamaIndex та модуль генерації Weaviate, дозволяють легко інтегрувати різноманітні моделі. Вибір конкретної моделі може залежати від таких чинників, як конфіденційність даних, вартість, наявні ресурси і більше.

Одним зі змінних параметрів, який характерний для ММ, є температура. Параметр температури контролює рівень випадковості в генерованому виводі. Температура 0 призводить до більш передбачуваної відповіді з меншою варіацією. З іншого боку, температура 1 надає моделі можливість вводити випадковість та творчість у свої відповіді. Отже, якщо генеративна модель запускається кілька разів і має температуру 1, відповіді можуть відрізнятись після кожного повторного запуску.

Ще одним важливим параметром є ліміт токенів, з якими здатна працювати модель. Цей ліміт визначає скільки токенів може існувати у контекстному вікні, яке являє собою вхідні дані для мовної моделі. У контексті великих мовних моделей (LLM) термін «токен» відноситься до фрагмента тексту, який модель читає або генерує. Токен зазвичай не є словом; це може бути менша одиниця, як символ або частина слова, або більша, як ціла фраза. Ліміт токенів задається на етапі розробки архітектури мовної моделі та не може бути змінений

## 2.2 Удосконалення методу побудови контекстних даних для чат-бота

Етап пошуку відповідних фрагментів вхідного документа є ключовим для RAG систем, оскільки найбільше впливає на повноту та правильність генерованої відповіді. Також, етап пошуку тісно пов'язаний з етапом «індексації», який являє собою розбиття вхідних даних, переведення їх у векторне представлення та зберігання.

Етап пошуку можливо удосконалити шляхом використання підходу підсумовування (узагальнення) вхідного тексту з подальшим пошуком саме по підсумованим даним та зіставлення знайдених фрагментів з оригінальними даними. Ідея полягає в тому, що використання узагальненого вмісту робить

семантичний пошук більш ефективним, тоді як використання більшої частини оригінального вмісту корисніше під час відповіді на запитання [7].

Загальна схема удосконаленого методу наведена на рисунку 2.3.

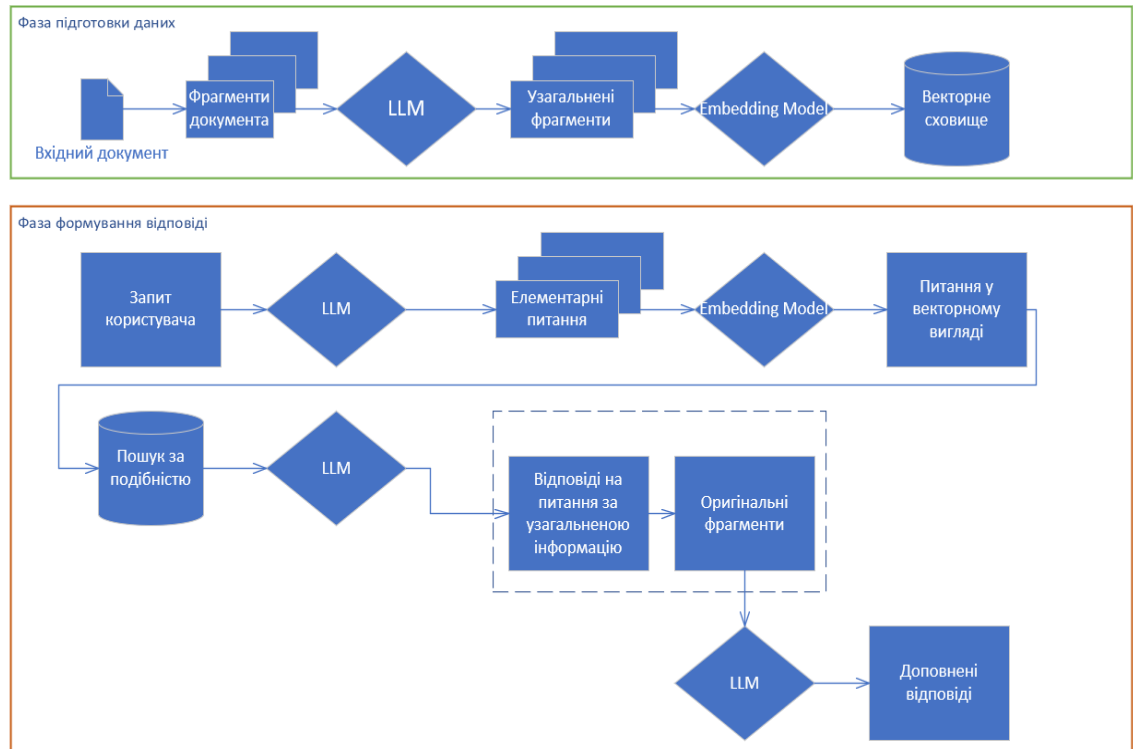


Рисунок 2.3 – Загальна схема удосконаленого методу

Під час запити процес виглядає дуже схожим на базовий варіант RAG, з тією відмінністю, що тепер пошук відбувається по відношенню до підсумованих даних. Виконуючи векторний семантичний пошук по підсумованих даних, підвищується ймовірність того, що запитання користувача відповідатиме відповідній частині вмісту. Це пов'язано з тим, що більші фрагменти збільшують ймовірність того, що інформація буде вирвана з контексту, а процес узагальнення зменшує будь-який відволікаючий шум, який може бути присутнім у вихідному контексті. Підсумований текст зберігає основний зміст документа.

Проте, під час підсумовування відбуваються неминучі втрати дрібних деталей. Для запобігання цих втрат необхідно передавати LLM оригінальний фрагмент, який містить у собі більше інформації.



Рисунок 2.4 – Схема фази підготовки даних з прикладами даних

Обраний підхід починається на етапі індексування вхідних даних. Схему наведено на рисунку 2.4.

Для створення підсумків великих фрагментів використовується велика мовна модель, оскільки є найбільш ефективною для задач обробки природної мови.

Спочатку оригінальний текстовий документ, буде сегментовано з використанням більшого розміру частини. Визначення оптимального розміру блоку не має конкретної формули, що вимагає експериментів, щоб визначити найбільш прийнятний розмір для конкретного випадку використання. У цьому випадку було вибрано розмір блоку, утричі більший, ніж підхід Basic RAG, спочатку встановлений на 400, що призвело до загального розміру 1300. Додатковий контекст забезпечується за допомогою введення невеликого доповнення для цього більшого розміру блоку. Крім того, для покращення узгодженості сегментованого тексту реалізовано ширше перекриття з 80 символів.

Пізніше мовна модель (LLM) використовується для створення підсумків для кожного з цих розширених блоків. Процес передбачає надання промπτ для LLM, де повідомляється про намір створити короткий виклад даного тексту. Примітно, що в промπτі немає ніяких хитросплетінь; бажана дія для LLM просто сформульована, що є коротким підсумовуванням наданого тексту. Вигляд промπτу наведено на рисунку 2.5.

```
"""Summarize the block of text below.
```

```
Text:
```

```
-----
```

```
{text}
```

```
-----
```

```
Your summary: """
```

Рисунок 2.5 – Промпт для створення підсумків тексту

Після здійснення вище зазначеної команди модель поверне підсумований фрагмент вхідного документу. Результат виконання наведено на рисунку 2.6.

```
-- Summarizing chunk: -----
```

```
Walmart Policies and Guidelines
```

```
Find a list of Walmart's most frequently requested public policies and guidelines, including our store return policy, coupon policy and more.
```

```
(1098) total length
```

```
--- Summary: -----
```

```
Walmart and Sam's Club are committed to providing safe, affordable, and sustainable food while ensuring the humane treatment of animals, specifically focusing on their.... (454) total length
```

Рисунок 2.6 – Результат виконання промπτу для підсумовування

Здійснення узагальнення фрагменту вхідного тексту скоротило його обсяг більше ніж вдвічі, з 1098 токенів до 454 та зберегло основний зміст, що вже свідчить про ефективність даного підходу.

Замість того, щоб зберігати підсумки як звичайні текстові метадані у векторному сховищі, пропонується зберігати лише ідентифікатор, який вказує

на відповідний великий контекстний елемент у структурі даних за межами векторного сховища. Для того щоб мати можливість знайти оригінальний вміст великого фрагмента необхідно зберегти індекс відповідного вихідного документа як «source\_id» у векторному сховищі. Детальну схему роботи з прикладами наведено на рисунку 2.7.

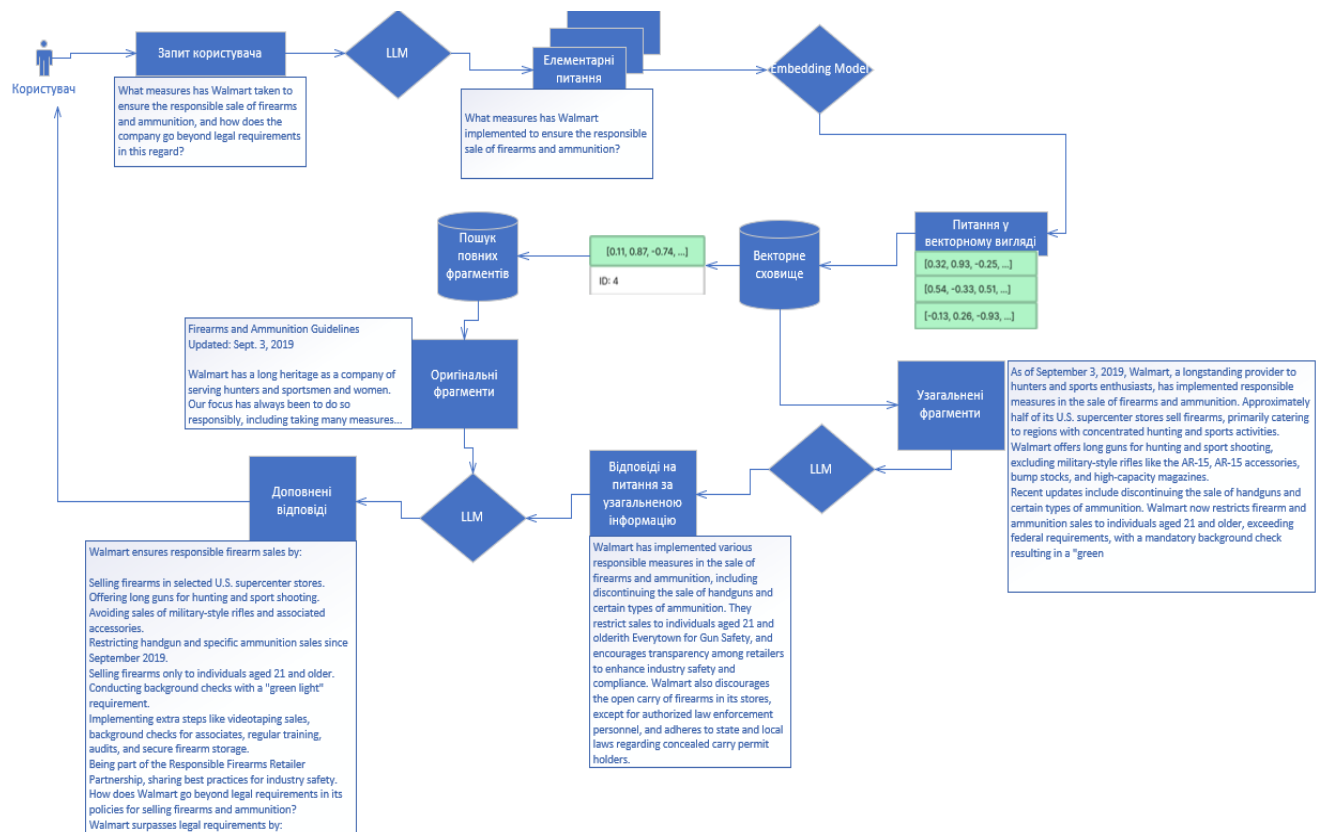


Рисунок 2.7 – Схема фази формування відповіді

Під час запиту виконується семантичний пошук запитання користувача по узагальненим фрагментам, а потім використовується збережений ідентифікатор фрагмента відповідного результату пошуку для зіставлення з великим фрагментами, щоб отримати оригінальний великий фрагмент контексту для використання в промпті для LLM.

Генерація відповіді проходить в два етапи:

- генерація прототипу відповіді на основі узагальненої інформації;

– деталізація прототипу відповіді за допомогою надання мовної моделі повної інформації з оригінального фрагменту.

Роботу даного методу перевірено шляхом введення запитання, яке потребує специфічних знань, що знаходяться у векторному сховищі та надійшли з вхідного документа. Запитання, знайдений контекст, прототипи відповіді та доповнену відповідь наведено на рисунку 2.8.

Питання:	Повна інформація	Узагальнена інформація	Прототип відповіді	Результуюча доповнена відповідь
<p>What measures has Walmart taken to ensure the responsible sale of firearms and ammunition, and how does the company go beyond legal requirements in this regard?</p>	<p><b>Firearms and Ammunition Guidelines</b> Updated: Sept. 3, 2019</p> <p>Walmart has a long heritage as a company of serving hunters and sportsmen and women. Our focus has always been to do so responsibly, including taking many measures that go beyond what's required by law.</p> <p><b>Where We Sell Firearms and Ammunition</b> Walmart sells firearms in approximately half of our U.S. supercenter stores. While we are not the largest volume firearms seller in the United States, we do serve many areas of the country where there is a concentration of hunters and sportsmen/women.</p> <p><b>Items We Sell</b> Walmart sells long guns for hunting and sport shooting, including shotguns, single-shot hunting rifles and light sporting rifles, as well as BB and pellet guns.</p>	<p>As of September 3, 2019, Walmart, a longstanding provider to hunters and sports enthusiasts, has implemented responsible measures in the sale of firearms and ammunition. Approximately half of its U.S. supercenter stores sell firearms, primarily catering to regions with concentrated hunting and sports activities. Walmart offers long guns for hunting and sport shooting, excluding military-style rifles like the AR-15, AR-15 accessories, bump stocks, and high-capacity magazines. Recent updates include discontinuing the sale of handguns and certain types of ammunition. Walmart now restricts firearm and ammunition sales to individuals aged 21 and older, exceeding federal requirements, with a mandatory background check resulting in a "green light."</p> <p>To ensure responsible sales, Walmart undertakes additional measures beyond legal obligations, such as videotaping sales, restricting authorized associates for firearm sales, regular training, inventory audits, and secure firearm storage. Non-compliance with policies can result in disciplinary actions, including termination. Walmart actively participates in the Responsible Firearms Retailer Partnership, collaborating with Everytown for Gun Safety and sharing best practices to</p>	<p>Walmart has implemented various responsible measures in the sale of firearms and ammunition, including discontinuing the sale of handguns and certain types of ammunition. They restrict sales to individuals aged 21 and older, conduct mandatory background checks, and go beyond legal requirements by videotaping sales, limiting authorized associates for firearm sales, providing regular training, conducting inventory audits, and ensuring secure firearm storage. The company actively participates in the Responsible Firearms Retailer Partnership, collaborates with Everytown for Gun Safety, and encourages transparency among retailers to enhance industry safety and compliance. Walmart also discourages the open carry of firearms in its stores, except for authorized law enforcement personnel, and adheres to state and local laws regarding concealed carry permit holders.</p>	<p>Walmart ensures responsible firearm sales by:</p> <ul style="list-style-type: none"> <li>Selling firearms in selected U.S. supercenter stores.</li> <li>Offering long guns for hunting and sport shooting.</li> <li>Avoiding sales of military-style rifles and associated accessories.</li> <li>Restricting handgun and specific ammunition sales since September 2019.</li> <li>Selling firearms only to individuals aged 21 and older.</li> <li>Conducting background checks with a "green light" requirement.</li> <li>Implementing extra steps like videotaping sales, background checks for associates, regular training, audits, and secure firearm storage.</li> <li>Being part of the Responsible Firearms Retailer Partnership, sharing best practices for industry safety.</li> <li>How does Walmart go beyond legal requirements in its policies for selling firearms and ammunition?</li> <li>Walmart surpasses legal requirements by: <ul style="list-style-type: none"> <li>Selling firearms in selected supercenter stores, exceeding volume considerations.</li> <li>Avoiding sales of military-style rifles and discontinuing certain handgun and ammunition types.</li> <li>Selling firearms only to those aged 21 and older, surpassing federal age requirements.</li> <li>Requiring a "green light" in background checks, exceeding the federal three-day waiting period.</li> <li>Taking extra steps like videotaping sales, background checks for selling associates, regular training, audits, and secure firearm storage.</li> </ul> </li> </ul>

Рисунок 2.8 – Перевірка роботи методу

Важливо те, що це дозволяє LLM мати багатий набір інформації, яка, швидше за все, містить відповідь на запитання користувача, і LLM може відповідати з таким самим рівнем глибини та нюансів, як представлено в оригінальному документі.

Векторні бази даних, аналогічно іншим типам баз даних, стають більш витратними по мірі збільшення обсягу сховища інформації. Тому оптимізація обсягу даних в базі є ключовим аспектом для зниження витрат. Зменшення обсягу даних у векторній базі даних на 3-кратну кількість, яке супроводжується підвищенням якості відповідей додатків являє собою подвійну вигоду та суттєву економію коштів у перспективі.

Необхідно враховувати, що при застосуванні даної стратегії узагальнення збільшаться витрати на індексацію через використання LLM для формування підсумованого тексту. Можливо, це буде відноситися до єдиночасних витрат, але при статичних даних важливо врахувати їх вплив на загальні експлуатаційні витрати при запуску додатків.

### 2.3 Оцінка подібності для семантичного пошуку контекстних даних для чат-боту

Під час керування веб-сайтом або сайтом електронної комерції з величезним набором вмісту чи продуктів стандартний пошук за ключовими словами може бути неефективним.

Вони покладаються на відповідність конкретних слів у запиті, що часто призводить до результатів, які не мають контексту чи наміру пошуку. Семантичний пошук удосконалює це, охоплюючи значення запиту, вибираючи вміст, релевантний за значенням, а не лише за збігом слів.

Візьмемо такий сценарій ви вводите в рядок пошуку «Посади початкового рівня в секторі відновлюваної енергетики». Базовий пошук за ключовими словами може відображати сторінки, що містять «початковий рівень», «позиції», «відновлювані джерела енергії», «енергетика» та «сектор». Але це не означає, що ви знайдете списки вакансій. Натомість ви можете побачити освітні статті чи новини галузі.

Однак за допомогою семантичного пошуку система розуміє, що ви шукаєте вакансії у сфері відновлюваної енергетики початкового рівня.

Потім він представляє конкретні списки вакансій, наприклад «Молодший монтажник сонячних панелей» або «Стажер-технік вітрових турбін», безпосередньо відповідаючи на ваш пошуковий намір. Семантичний пошук з'єднує слова та значення, надаючи вам важливі результати.

Один з механізмів який забезпечує роботу семантичного пошуку - косинус подібності. Косинусна подібність – це показник, який визначає, наскільки схожі два документи, незалежно від їх розміру. Цей метод обчислює косинус кута між двома ненульовими векторами в багатовимірному просторі — ці вектори представляють текстовий вміст.

Це інструмент, який доводить свою високу ефективність у семантичному пошуку, який зосереджується на пошуку матеріалу, який поділяє значення із запитом, а не лише ідентичних ключових слів.

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

Візуальне представлення косинусної подібності наведено на рисунку 2.9. Зображення є тривимірним сюжетом із трьома осями, кожна з яких представляє окрему тему: McDonald's, Popeyes, Mixed documents і Mona Lisa. У цьому просторі зображено чотири документи: один пов'язаний із McDonald's, Моною Лізою, один пов'язаний із Popeyes і один, який поєднує обидві теми. Відстань між цими точками обчислюється за допомогою косинусної подібності, яка вимірює, наскільки документи пов'язані з кожною темою на основі кута між ними. Чим ближче документи один до одного, тим більше вони схожі. Документ про Мону Лізу побудований далеко від двох інших, показуючи, що це зовсім інша тема.

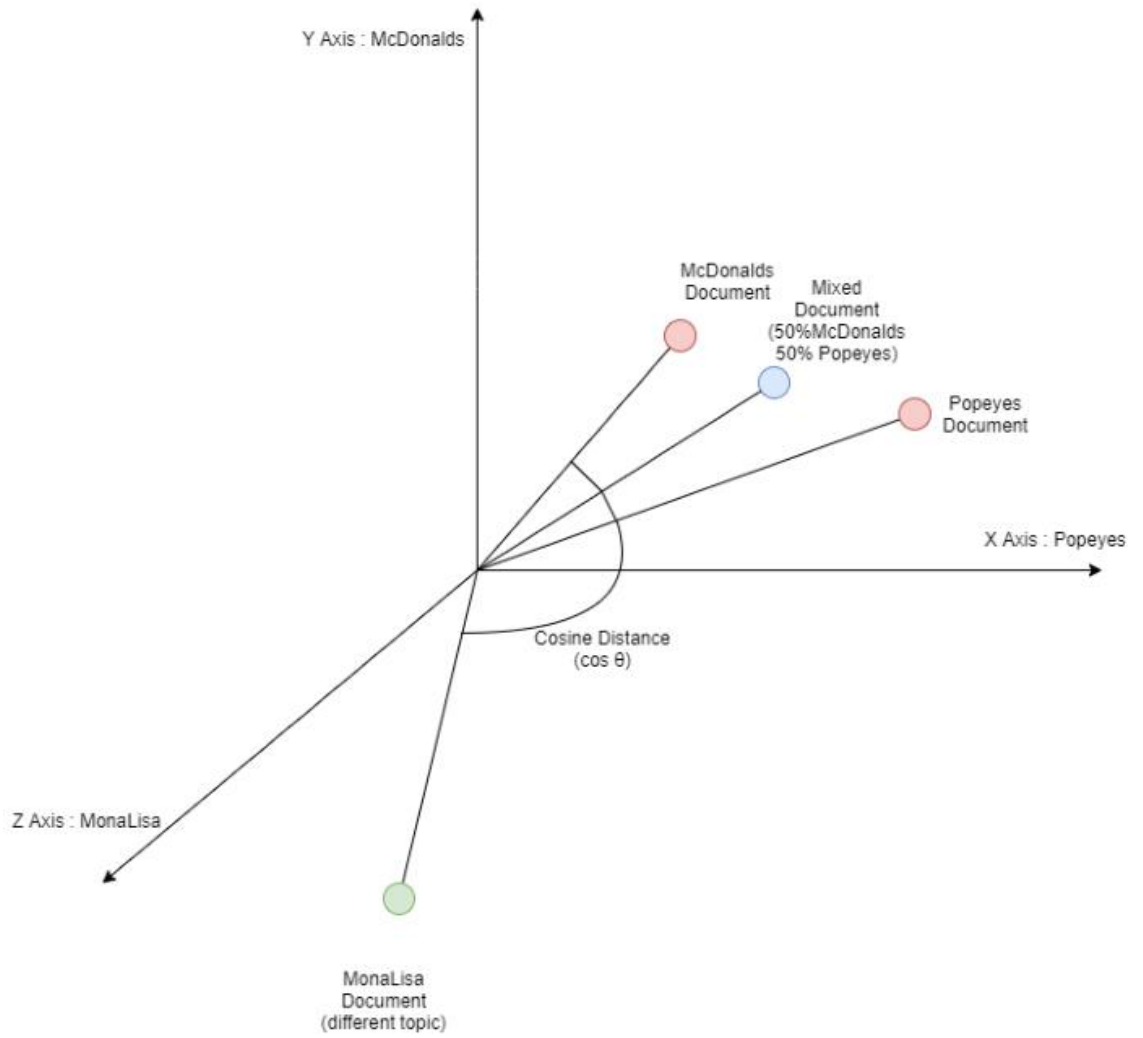


Рисунок 2.9 – Візуальне представлення косинусної подібності

## 3 ТЕХНОЛОГІЯ ВИКОРИСТАННЯ ЛОКАЛЬНИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ЧАТ-БОТІВ В СИСТЕМІ ЕЛЕКТРОННОЇ КОМЕРЦІЇ

### 3.1 Огляд технології

Перш ніж створювати систему RAG, важливо ознайомитися з інструментами, які роблять цей процес можливим. Кожен інструмент відіграє певну роль, забезпечуючи ефективну та результативну роботу системи RAG.

- мовна модель (LLM) – в основі системи лежить LLM, основна модель штучного інтелекту, відповідальна за генерацію людських текстових відповідей;

- vector store – це спеціальна система зберігання, у якій зберігаються вбудовані файли та відповідні їм текстові дані, що забезпечує швидкий та ефективний пошук;

- vector store retriever – являє собою пошукову систему. Vector Store Retriever отримує відповідні документи шляхом порівняння схожості векторів, гарантуючи, що найбільш актуальна інформація завжди під рукою;

- embedder – перш ніж зберігати або отримувати дані, нам потрібно перетворити текстову інформацію у формат, зрозумілий системі. Embedder бере на себе цю роль, перетворюючи текст у векторне представлення;

- підказка – кожна взаємодія починається із запиту або заяви користувача. Підказка фіксує цей початковий вхід, готуючи основу для процесів пошуку та генерації;

- завантажувач документів. Оскільки необхідно обробляти величезні обсяги даних, завантажувач документів є важливим. Він імпортує та читає документи, готуючи їх до фрагментації та вбудовування;

- document chunker – щоб зробити дані більш керованими та ефективними для пошуку, Document Chunker розбиває документи на менші частини, які легше засвоювати;

- запит користувача – останнє, але не менш важливе, інструмент

введення користувача фіксує запит або заяву, надану кінцевим користувачем, ініціюючи весь процес RAG.

На рисунку 3.1 наведено діаграму для реалізації технології удосконаленого методу, а саме для фази підготовки даних, тобто створення векторної бази довідкових матеріалів.

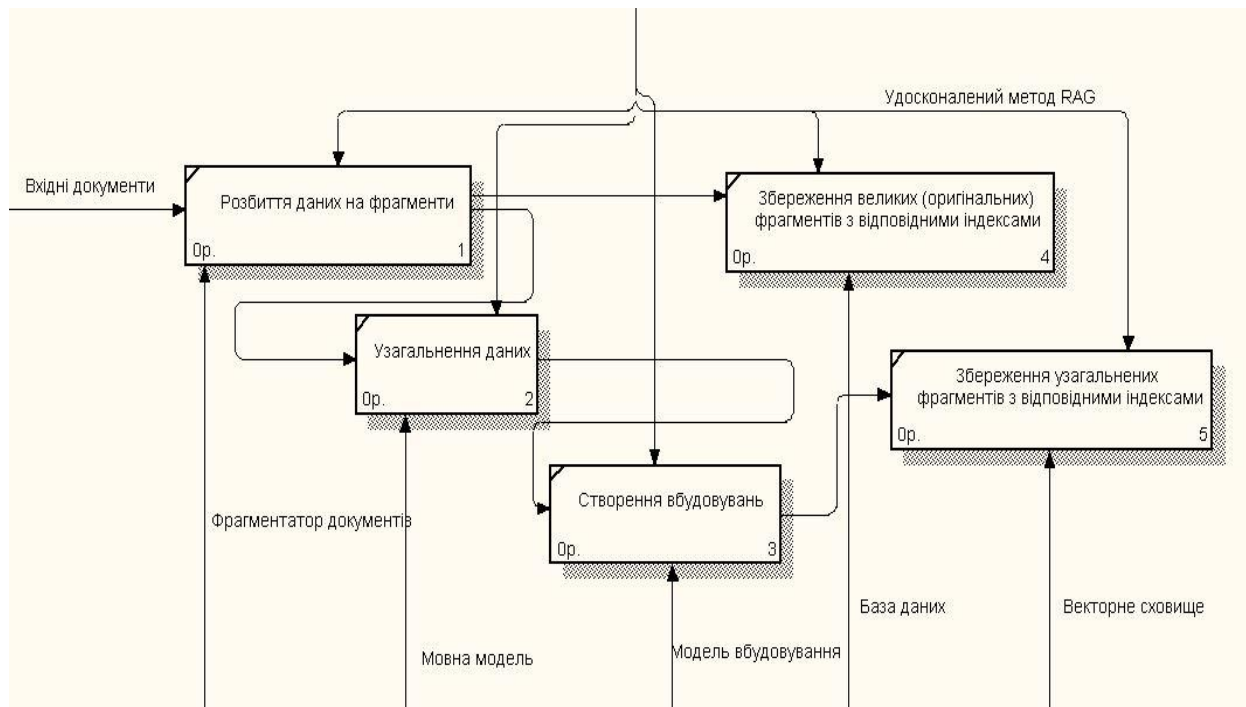


Рисунок 3.1 – Діаграма для фази підготовки даних про предметну область

На рисунку 3.2 наведено діаграму реалізації фази формування відповіді чат-бота, що включає в себе:

- розбиття запиту користувача: являє собою пошукову систему – за допомогою мовної моделі запит розбивається на окремі елементарні питання;
- вбудовування елементарних питань – перетворення кожного питання у векторний вигляд за допомогою моделі вбудовування;
- пошук узагальненої інформації за подібністю у векторному сховищі – пошук узагальнених фрагментів тексту за подібністю до кожного питання;
- генерація відповідей на основі узагальненої інформації – генерація прототипів відповідей на основі узагальненої інформації за допомогою LLM;

– знаходження оригінальних фрагментів за за індексом – пошук великих, не узагальнених фрагментів за спільним індексом з узагальненими фрагментами у базі даних;

– розширення прототипів відповідей за допомогою оригінальних фрагментів – формування результуючої відповіді за допомогою контексту з оригінальних даних та прототипів відповідей.

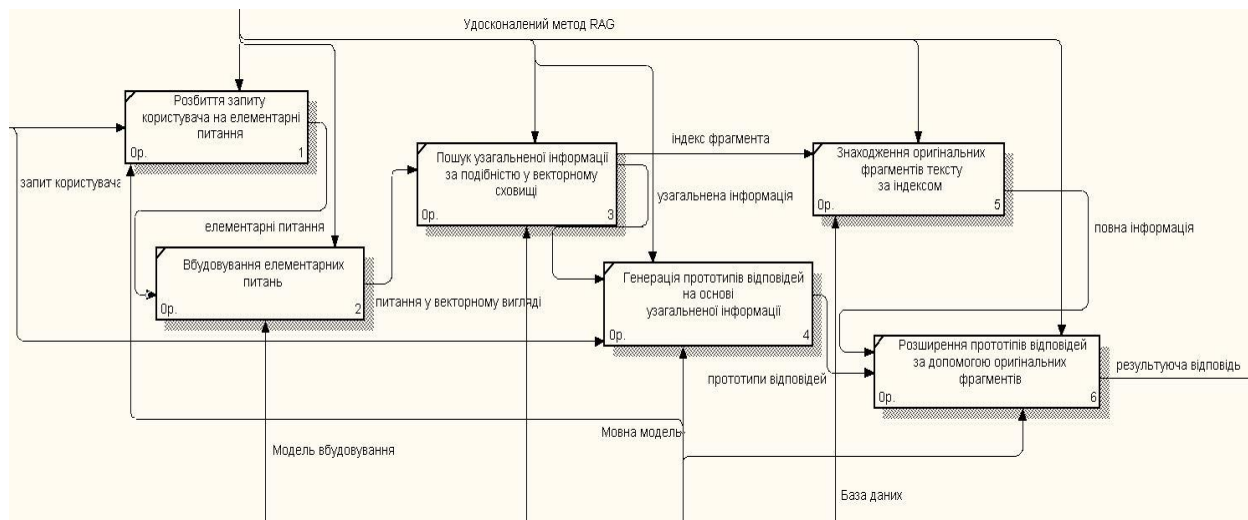


Рисунок 3.2 – Діаграма для фаз формування відповіді чат-бота

### 3.2 Реалізація етапу формування вбудовувань інформаційної технології

Зберігач (Retriever) – це частина RAG призначена для визначення відповідної інформації у величезному наборі даних, що містить у собі модель вбудовування (embedding model). Використовуючи вдосконалені методи, відомі як щільний пошук, зберігач створює числові представлення, які називаються вбудовуваннями, як запитів, так і документів. Він розміщує схожі запити та документи поруч один з одним у великому просторі.

Коли надходить запит, модель використовує семантичні методи пошуку, як-от косинусну подібність, щоб ідентифікувати та надати документи, які найбільше відповідають контексту. Сила цієї моделі полягає в її точності; він

відмінно підходить для швидкого пошуку точної інформації, необхідної з великої маси даних.

Для ілюстрації наведено зображення (рисунок 3.3), яке показує вбудовування слів для таких термінів, як «король», «королева», «чоловік» і «жінка» в тривимірному просторі, демонструючи, як семантично пов'язані терміни групуються разом.

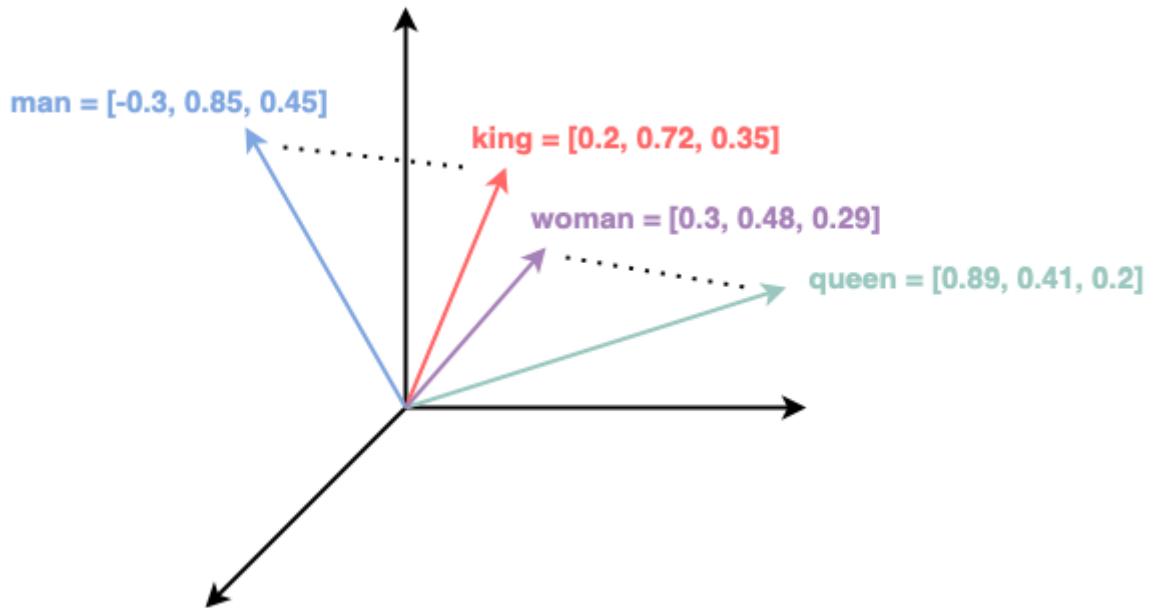


Рисунок 3.3 – Візуалізація векторних даних

Можливо сприймати це як розумну пошукову систему, яка повертає найпопулярніші документи, релевантні для відповіді на певний запит. Потім мовна модель діє як записувач і формує відповідь на основі цих документів.

## 4 РЕАЛІЗАЦІЯ ТА ЕКСПЕРИМЕНТАЛЬНА ПЕРЕВІРКА ВИКОРИСТАННЯ ЛОКАЛЬНОЇ МОВНОЇ МОДЕЛІ ДЛЯ ЧАТ-БОТА

### 4.1 Вибір платформи для побудови інтерфейсу користувача

Один з ключових аспектів реалізації програмного продукту – це користувацький інтерфейс. Поставимо вимоги до інтерфейсу:

- легкість інтеграції з розробленою діалоговою системою, а також можливість легко підтримувати, оновлювати та змінювати логіку роботи без переробки інтерфейсу;

- зручність та зрозумілість інтерфейсу для легкого інтерактивного спілкування з ботом.

Оскільки розроблений продукт є демонстраційним, в якості платформи для інтерфейсу було обрано StreamLit.

Streamlit – це потужна та легка у використанні відкрита бібліотека для мови програмування Python, яка полегшує створення інтерактивних веб-додатків для машинного навчання та науки про дані. За допомогою Streamlit розробники та вчені в галузі даних можуть створювати захопливі, інформативні та візуально привабливі додатки всього кількома рядками коду. Ця бібліотека дозволяє вам швидко та ефективно візуалізувати дані, будувати графіки, інтерактивні таблиці та інші компоненти, не витрачаючи багато часу на складну розробку.

Streamlit дозволяє розробникам зосередитися на створенні важливих функцій та взаємодії з користувачем, замість того, щоб докладати великі зусилля на розгортання складних інтерфейсів.

## 4.2 Опис компонентів чат-бота

Розроблена система складається з наступних компонентів:

- embedding модель (HuggingFaceTransformers) – моделі готують основу для більш обґрунтованої, контекстно-насиченої генерації мови, підвищуючи можливості традиційних мовних моделей;

- велика мовна модель (Mistral 7b) – Ці моделі діють як творчі автори, синтезуючи отриману інформацію в зв'язний і контекстуально відповідний текст;

- векторна база знань Facebook AI Similarity Search (Faiss) – Faiss виступає як бібліотека, розроблена для ефективного дослідження подібності та кластеризації в щільних наборах векторних даних. Оснащений алгоритмами, здатними здійснювати пошук серед векторних наборів різного розміру, включаючи ті, що потенційно перевищують обсяг оперативної пам'яті, він також містить супровідний код для оцінки та налаштування параметрів. Faiss написано на C++ і постачається з комплексними оболонками для Python/numpy з певними високоефективними алгоритмами, оптимізованими для використання GPU. Основний центр розробки для Faiss знаходиться в дослідницькій групі Meta Fundamental AI Research;

- логіка додатку (Langchain) LangChain – це платформа для створення програм на основі великих мовних моделей (LLM). Надає інструменти та абстракції для розширення можливостей налаштування, підвищення точності та актуальності інформації, що генерується. Також у LangChain є компоненти, які дозволяють застосовувати LLM до нових наборів даних без повторного навчання.

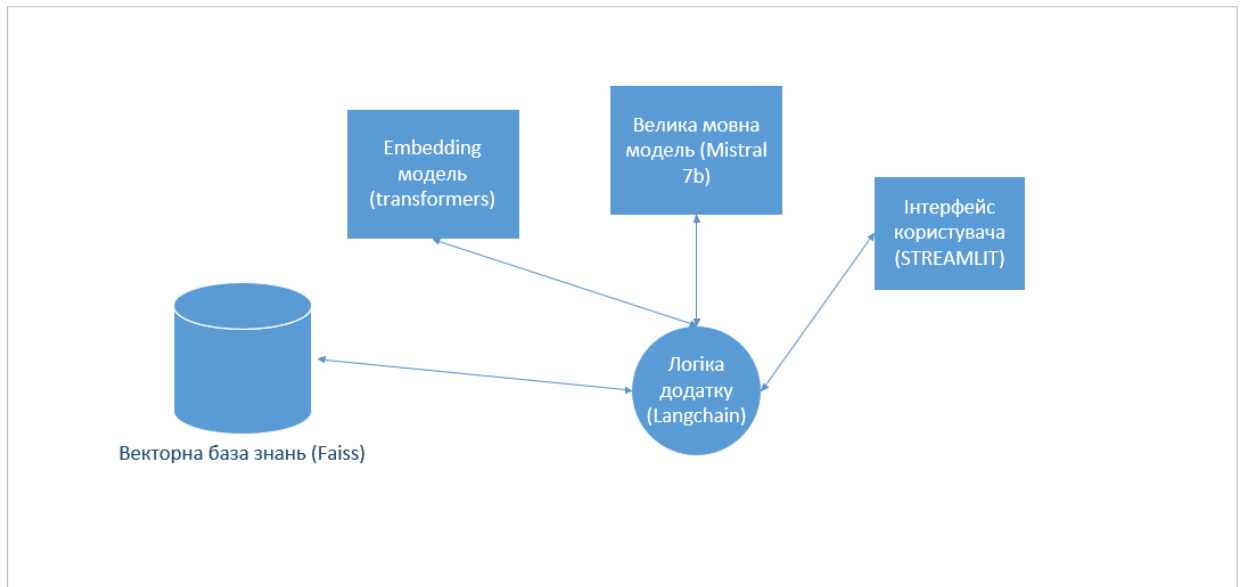


Рисунок 4.1 – Компоненти додатку

В якості моделі для побудови генератора відповідей було обрано «Mistral 7B».

Mistral 7B – це мовна модель із семи мільярдами параметрів, представлена компанією Mistral AI, новачком у галузі штучного інтелекту. Ця модель здобула увагу як одна з найпотужніших моделей із семи мільярдами параметрів.

Mistral 7B представлена як фундаментальна модель, що вказує на її значущість як основного будівельного блоку у обробці природної мови. Вона входить у розвиваючийся ландшафт великих мовних моделей.

Незважаючи на відносно скромний розмір порівняно з деякими більшими моделями, Mistral 7B вражає вражаючою ефективністю у різноманітних завданнях розуміння та генерації природної мови. Вона відмінно справляється із завданнями, такими як текстова резюмзація, класифікація тексту, завершення тексту та автодоповнення коду.

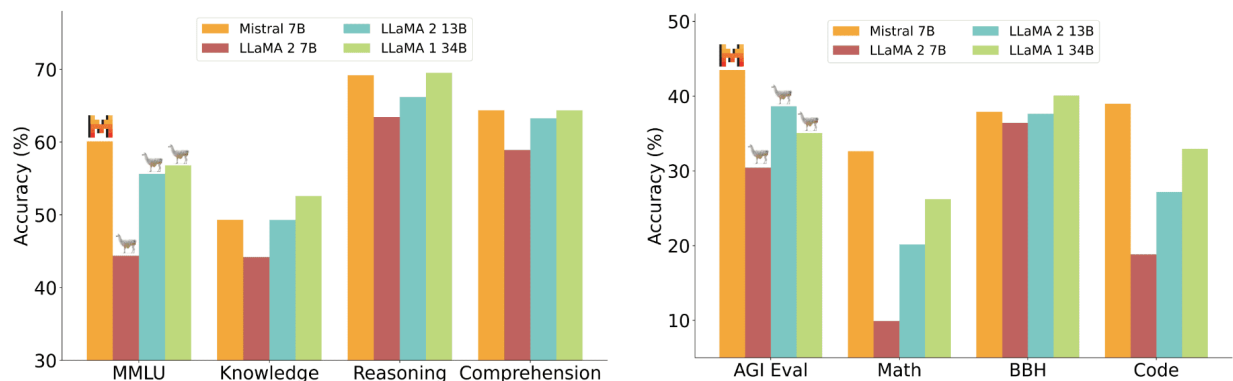
Mistral 7B відрізняється своїм архітектурним підходом. Модель використовує новаторську архітектуру, яка не є повністю визначеною, але вказує на відмінну філософію дизайну порівняно із відомими моделями, такими як GPT-3. Вона використовує Grouped-query attention (GQA) для

швидкого інференсу та Sliding Window Attention (SWA) для обробки довших послідовностей за менший вартісний рахунок. Це призводить до дуже високої продуктивності для такої невеликої моделі.

На даний момент Mistral 7B в основному сконцентрована на англійській мові з можливістю програмування. Однак відомо, що вона має значно велике вікно контексту, що дозволяє їй розуміти та генерувати текст в ширшому контексті порівняно із іншими моделями, такими як GPT-3.

Високі показники ефективності Mistral 7B на різних тестах перевершують результати набагато більших моделей, таких як GPT-3, що свідчить про її ефективність та компетентність у роботі з мовними завданнями.

На рисунку 4.2 наведено порівняльний аналіз Mistral 7B з іншими моделями за різними тестами:



Model	Modality	MMLU	HellaSwag	WinoGrande	PIQA	Arc-e	Arc-c	NQ	TriviaQA	HumanEval	MBPP	MATH	GSM8K
LLaMA 2 7B	Pretrained	44.4%	77.1%	69.5%	77.9%	68.7%	43.2%	24.7%	63.8%	11.6%	26.1%	3.9%	16.0%
LLaMA 2 13B	Pretrained	55.6%	<b>80.7%</b>	72.9%	80.8%	75.2%	48.8%	<b>29.0%</b>	<b>69.6%</b>	18.9%	35.4%	6.0%	34.3%
Code LLaMA 7B	Finetuned	36.9%	62.9%	62.3%	72.8%	59.4%	34.5%	11.0%	34.9%	<b>31.1%</b>	<b>52.5%</b>	5.2%	20.8%
Mistral 7B	Pretrained	<b>60.1%</b>	<b>81.3%</b>	<b>75.3%</b>	<b>83.0%</b>	<b>80.0%</b>	<b>55.5%</b>	<b>28.8%</b>	<b>69.9%</b>	<b>30.5%</b>	47.5%	<b>13.1%</b>	<b>52.1%</b>

Рисунок 4.2 – Результати порівняння Mistral 7b з іншими моделями

Крім того, дана модель випускається під ліцензією Apache 2.0, що дозволяє використовувати її в комерційних цілях без обмежень.

У якості даних для побудови векторної бази знань із якою буде працювати обрана мовна модель було використано FaQ декількох сучасних e-commerce платформ. FAQ - це аббревіатура від англійського вислову "Frequently Asked Questions", що в перекладі на українську означає "Питання, які

найчастіше задаються". Це розділ або сторінка, яка містить відповіді на найчастіші питання з метою надання користувачам швидкої інформації та вирішення їхніх потреб. FAQ часто використовується на веб-сайтах, у документації, програмах, сервісах та інших ресурсах.

### 4.3 Імплементация RAG-чатботу

Логіка додатку розроблена за допомогою мови програмування Python та фреймворку Langchain.

Перший крок являє собою створення функціоналу для завантаження документа. LangChain надає кілька вбудованих завантажувачів документів, які працюють із файлами PDF, файлами JSON або файлами Python у каталозі файлів. Можливо використовувати PyPDFLoader від LangChain, щоб безперешкодно імпортувати PDF. Код для створення завантажувача файлів наведено на рисунку 4.3.

```
1. class FileIngestor:
2.     def __init__(self, uploaded_file):
3.         self.uploaded_file = uploaded_file
4.
5.     def handlefileandigest(self):
6.         with tempfile.NamedTemporaryFile(delete=False) as tmp_file:
7.             tmp_file.write(self.uploaded_file.getvalue())
8.             tmp_file_path = tmp_file.name
9.
10.         loader = PyMuPDFLoader(file_path=tmp_file_path)
11.         data = loader.load()
```

Рисунок 4.3 – Код для створення завантажувача файлів

Другий крок – розділення документи на частини: якщо документ довгий, необхідно розділити текст документа на частини. Існують різні способи

розділення тексту. Пропонується використати метод `CharacterTextSplitter`, щоб розділити на основі символів і виміряти довжину блоку за кількістю символів.

Код для розділення документу наведено на рисунку 4.4.

```
large_chunk_size = 1300
large_chunk_overlap = 80
large_chunk_text_splitter =
RecursiveCharacterTextSplitter(chunk_size=large_chunk_size,
                               chunk_overlap=large_ch
unk_overlap)
large_chunks = large_chunk_text_splitter.split_documents(documents)

print(f"Split document into {len(large_chunks)} chunks of text")
```

Рисунок 4.4 – Код для розділення документу

Третій крок є ключовим для удосконаленого методу і являє собою додавання функціоналу для узагальнення великих фрагментів. Для цього також потрібно використати мовну модель. Код для створення підсумовування даних наведено на рисунку 4.5.

```
from langchain.docstore.document import Document

summary_documents = []
for doc in large_chunks:
    to_summarize = doc.page_content

    print("--- Summarizing chunk: -----")
    print(f"{to_summarize[0:40]}... ({len(to_summarize)}) total length")
    response = openai.Completion.create(
        model="Mistral 7b",
        prompt=prompt_template.format(text=to_summarize),
        temperature=0.0,
        max_tokens=500
    )
    summary = response['choices'][0]['text'].strip()
    summary_documents.append(Document(page_content=summary, metadata=doc.metadata))
```

Рисунок 4.5 – Код для створення підсумків фрагментів

Наступний крок – створення вбудованих текстів: Фрагменти тексту перетворюються на числові вектори за допомогою вбудованих елементів, що дає змогу ефективно працювати з текстовими даними, як-от семантичний пошук. Для цього завдання можливо вибрати постачальника моделі вбудовування, наприклад OpenAI або HuggingFace (рисунок 4.6).

```
import openai
openai.api_key = openai_api_key

def create_embeddings(batch: list[str]):
    model_id = 'text-embedding-ada-002'
    embedding_resp = openai.Embedding.create(input=batch, model=model_id)
    return [emb['embedding'] for emb in embedding_resp['data']]

embeddings = create_embeddings([doc.page_content for doc in small_chunks])
```

Рисунок 4.6 – Код для створення вбудовувань

Розроблена система являє собою веб-додаток розроблений за допомогою сучасної платформи Streamlit, що покращує користувацький досвід.

Додаток складається з трьох сторінок:

- чат;
- налаштування;
- завантаження документів.

На рисунку 4.7 зображено початковий інтерфейс діалогової системи, який являє собою власне сторінку з чатом. Сторінка для спілкування з чат-ботом є інтерфейсом, який дозволяє користувачам взаємодіяти з ботом і отримувати відповіді на свої запитання. Основне вікно займає центральну частину сторінки і призначене для обміну повідомленнями з чат-ботом.

Включає в себе наступні ключові елементи:

- інтерфейс введення тексту. Це поле, в якому користувач може вводити свої запитання або команди та кнопка для відправлення запиту;

– інтерфейс виведення. Тут відображаються відповіді чат-бота, тобто згенерований текст, що надає користувачу необхідну інформацію.

Повідомлення користувача та відповіді чату відображаються у вигляді бесіди. Також є індикатор, який вказує на те, що чат-бот обробляє запит користувача. Поки система обробляє запит замість відповіді виводиться повідомлення «друкую...», що робить використання чатботом ще більш схожим на комунікації в будь-якому сучасному додатку-месенджері.

Функціонал даного інтерфейсу може бути розширений для потреб систем електронної комерції шляхом створення додаткових елементів управління таких як кнопки із заздалегідь визначеними командами або варіантами відповідей, щоб облегчити взаємодію з чат-ботом.

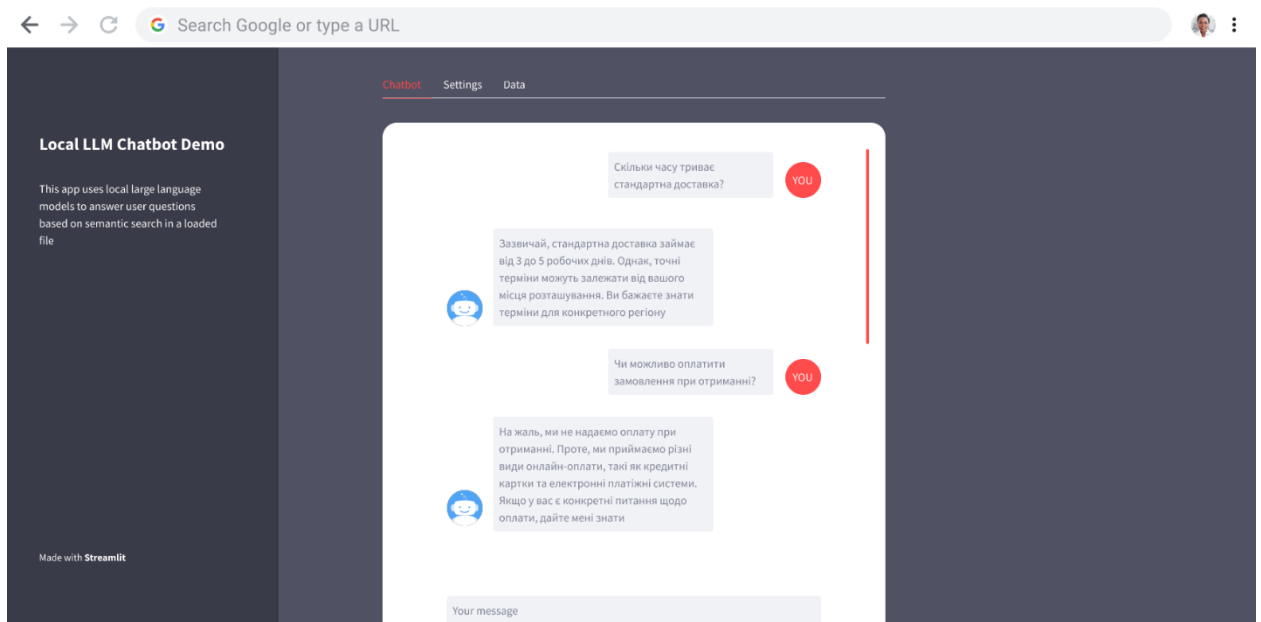


Рисунок 4.7 – Інтерфейс для спілкування з чат-ботом

Наступна сторінка додатку – налаштування. Сторінку зображено на рисунку 4.8.

Налаштування потрібні, як для вибору великої мовної моделі так і для зміни її параметрів, таких як:

- температура;
- top P;

- максимальна довжина;
- штраф за частоту;
- штраф за присутність.

Температура – безпосередньо впливає на детермінізм результатів. Нижчі температури забезпечують детермінований результат, послідовно вибираючи найбільш ймовірний наступний маркер. І навпаки, вищі температури вводять випадковість, збільшуючи ймовірність різноманітних або творчих результатів. По суті, це тягне за собою підвищення ваги альтернативних токенів. Для таких завдань, як відповіді на запитання на основі фактів (QA), доцільно використовувати нижче значення температури, щоб сприяти фактичним і лаконічним відповідям. Навпаки, для таких творчих завдань, як створення віршів, збільшення значення температури може виявитися вигідним.

Тор Р – метод вибірки з температурою, відомий як вибірка ядра, пропонує контроль над детермінованістю моделі. Нижче значення підтримує точні та фактичні відповіді, тоді як вище значення сприяє більш різноманітним результатам. З Тор Р лише токени в межах маси ймовірності `top_r` сприяють відповідям. Низьке значення `top_r` вибирає найбільш впевнені відповіді, тоді як високе значення `top_r` розширює облік моделі, включаючи менш ймовірні слова, що призводить до більш різноманітних результатів. Загальна рекомендація полягає в тому, щоб змінити або температуру, або Тор Р, але не обидва одночасно.

Максимальна довжина – це кількість токенів, згенерованих моделлю, можна регулювати, регулюючи максимальну довжину. Зазначення максимальної довжини допомагає запобігти довгим або нерелевантним відповідям і забезпечує контроль над пов'язаними витратами.

Штраф за частоту – Штраф за частоту застосовує покарання до наступного маркера, пропорційного до того, скільки разів цей маркер уже з'являвся у відповіді та підказці. Чим вищий штраф частоти, тим менша ймовірність появи слова знову. Цей параметр зменшує повторення слів у відповіді моделі, надаючи вищий штраф маркерам, які з'являються більше.

Штраф за присутність – Штраф за присутність також застосовує штраф для повторюваних жетонів, але, на відміну від штрафу за частоту, штраф є однаковим для всіх повторюваних токенів. Токен, який з’являється двічі, і токен, який з’являється 10 разів, караються однаково. Цей параметр не дозволяє моделі надто часто повторювати фрази у своїй відповіді. Якщо ви хочете, щоб модель створювала різноманітний або креативний текст, ви можете використати вищий штраф за присутність. Або, якщо вам потрібно, щоб модель залишалася зосередженою, спробуйте використати менший штраф за присутність.

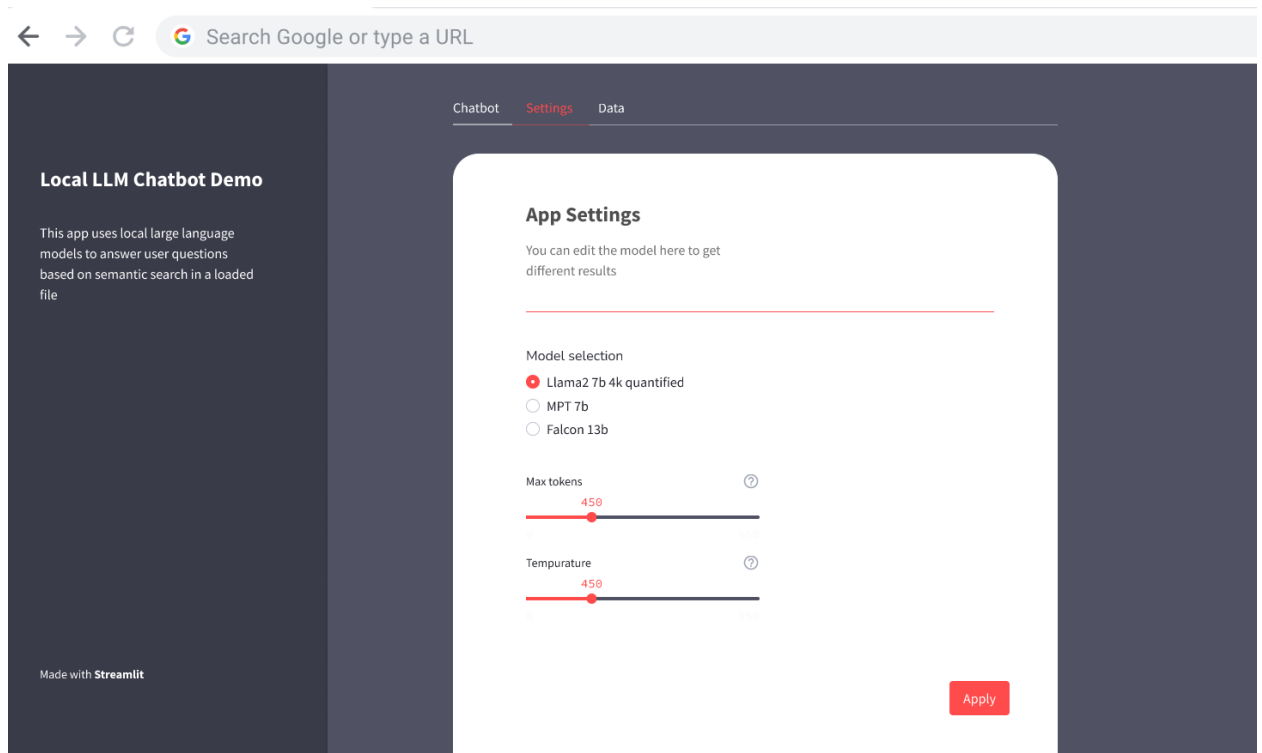


Рисунок 4.8 – Інтерфейс налаштування

Остання сторінка додатку – завантажувач документів. Основна частина сторінки включає область для завантаження файлів з розширенням PDF або TXT.

Кнопка "Вибрати файл" використовується для вибору файлів на диску для завантаження. Альтернативним способом вибору є перетягування файлу з

файлового менеджера, тобто drag and drop. Після успішного завантаження документа буде відображено його назву.

Окрім кнопки для завантаження даних на сторінці завантажувача також чекбокс, що надає можливість оптимізувати завантажений файл за допомогою використання підходу підсумовування тексту. Тобто, можливо використовувати як базову стратегію RAG, так і удосконалену. Вигляд сторінки завантажувача наведено на рисунку 4.9.

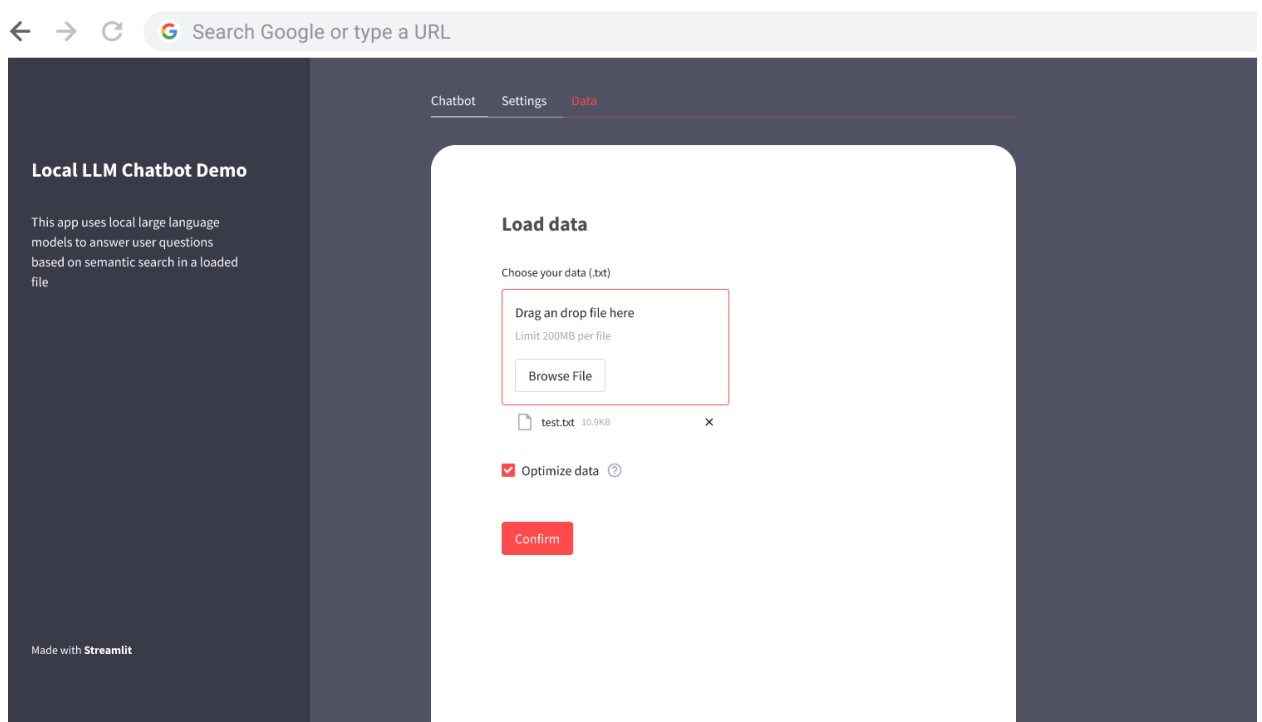


Рисунок 4.9 – Інтерфейс завантажувача

#### 4.4 Оцінка відповідей чат-бота

Оцінка відповідей чатбота була проведена шляхом порівняння їх із відповідями людини за допомогою NLP-моделі. Систематичний підхід до оцінки виводів мовних моделей (LLM), без використання ChatGPT, включає використання інших моделей машинного навчання, які базуються на обробці природної мови (NLP). Конкретні моделі можуть бути застосовані для оцінки виводів за різними метриками, такими як фактична точність, актуальність, упередженість та корисність, не зважаючи на недетермінований характер виводів.

Наприклад, моделі обробки природної мови, які виробляють оцінку унікальності, можуть бути використані для визначення того, наскільки фактично правильною є відповідь за вказаним контекстом. Вищий показник унікальності вказує на більший рівень фактичної точності виводу. Це особливо корисно при оцінці довгих виводів, які можуть бути складними для визначення в термінах фактичної вірності.

Виникає питання, як ці моделі можуть визначити фактичну правильність даного тексту. Для вирішення такої задачі NLP-моделям можна надавати контекстну інформацію, яку вони прийматимуть буквально. Зокрема, така контекстна інформація позначається як контекстуальні істини або посилання. Загальний набір цих посилань часто позначається як набір для оцінки.

Для оцінки розробленої системи було використано бібліотеку Ragas.

Ragas — це структурована структура, призначена для полегшення оцінки конвеєрів Retrieval Augmented Generation (RAG). Термін «RAG» відноситься до категорії додатків мовної моделі (LLM), які покращують свій контекст шляхом включення зовнішніх даних. Хоча вже існують інструменти та інфраструктура для побудови таких конвеєрів, проблема полягає в ефективній оцінці та кількісному вимірі продуктивності цих конвеєрів. Ragas,

або RAG Assessment, розглядає цей конкретний аспект, надаючи комплексне рішення для оцінки RAG.

Ragas пропонує метрики, призначені для окремого оцінювання кожного компонента системи RAG. Серед цих метрик:

- faithfulness;
- answer relevancy;
- context recall;
- context precision;
- context relevancy.

Метрика Faithfulness вимірює фактичну узгодженість згенерованої відповіді з даним контекстом. Він обчислюється на основі відповіді та отриманого контексту. Відповідь масштабується до діапазону (0,1). Чим вище, тим краще.

Метрика Answer relevancy зосереджується на оцінці того, наскільки згенерована відповідь відповідає даному запиту. Нижчий бал присвоюється відповідям, які є неповними або містять зайву інформацію. Цей показник обчислюється за допомогою запитання та відповіді зі значеннями в діапазоні від 0 до 1, де вищі бали вказують на кращу релевантність.

Метрика Context recall вимірює ступінь, до якого отриманий контекст узгоджується з анотованою відповіддю, яка розглядається як основна правда. Він обчислюється на основі основної правди та отриманого контексту, а значення коливаються від 0 до 1, причому вищі значення вказують на кращу продуктивність.

Метрика Context precision – це метрика, яка оцінює, чи всі релевантні елементи контекстів мають вищий рейтинг чи ні. В ідеалі всі відповідні фрагменти повинні з'являтися на перших рядках. Цей показник обчислюється за допомогою запитання та контекстів із значеннями в діапазоні від 0 до 1, де вищі бали вказують на кращу точність.

Метрика Context relevancy вимірює релевантність отриманого контексту, розрахованого на основі як питання, так і контекстів. Значення

знаходяться в діапазоні (0, 1), причому вищі значення вказують на кращу релевантність.

Код для оцінки розробленої системи наведено на рисунку 4.10.

```
from datasets import load_dataset

fiqa_eval = load_dataset("explodinggradients/fiqa", "ragas_eval")
fiqa_eval

DatasetDict({
  baseline: Dataset({
    features: ['question', 'ground_truths', 'answer', 'contexts'],
    num_rows: 30
  })
})
)
)
from ragas.metrics import (
  context_relevancy,
  answer_relevancy,
  faithfulness,
  context_recall,
)

from ragas.metrics.critique import harmfulness
from ragas import evaluate

result = evaluate(
  fiqa_eval["baseline"].select(range(3)),
  metrics=[
    context_relevancy,
    faithfulness,
    answer_relevancy,
    context_recall,
    harmfulness,
  ],
)

result
```

Рисунок 4.10 – Код для оцінювання чатбота

Результати оцінювання відповідей чат-бота наведено на рисунку 4.11.

index	question	ground_truths	answer	contexts	context_relevancy	faithfulness	answer_relevancy	context_recall	harmfulness
0	How to deposit a cheque issued to an associate in my business into my business account?	[Have the check reissued to the proper payee. Just have the associate sign the back and then deposit it. It's called a third party cheque and is perfectly legal. I wouldn't be surprised if it has a longer hold period and, as always, you don't get the money if the cheque doesn't clear. Now, you may have problems if it's a large amount or you're not very well known at the bank. In that case you can have the associate go to the bank and endorse it in front of the teller with some ID. You don't even technically have to be there. Anybody can deposit money to your account if they have the account number. He could also just deposit it in his account and write a cheque to the business.]	The best way to deposit a cheque issued to an associate in your business is to open a business account with the bank. You will need a state-issued "dba" certificate from the county clerk's office as well as an Employer ID Number (EIN) issued by the IRS. Once you have opened the business account, you can have the associate sign the back of the cheque and deposit it into the business account.	[Just have the associate sign the back and then deposit it. It's called a third party cheque and is perfectly legal. I wouldn't be surprised if it has a longer hold period and, as always, you don't get the money if the cheque doesn't clear. Now, you may have problems if it's a large amount or you're not very well known at the bank. In that case you can have the associate go to the bank and endorse it in front of the teller with some ID. You don't even technically have to be there. Anybody can deposit money to your account if they have the account number. He could also just deposit it in his account and write a cheque to the business.] I have checked with Bank of America, and they say the ONLY way to cash (or deposit, or otherwise get access to the funds represented by a check made out to my business) is to open a business account. They tell me this is a Federal regulation, and every bank will say the same thing. To do this, I need a state-issued "dba" certificate (from the county clerk's office) as well as an Employer ID Number (EIN) issued by the IRS. AND their CHEAPEST business banking account costs \$15 / month. I think I can go to the bank that the check is drawn upon, and they will cash it, assuming I have documentation showing that I am the sole proprietor. But I'm not sure... What a racket! When a business asks me to make out a cheque to a person rather than the business name, I take that as a red flag. Frankly it usually means that the person doesn't want the money going through their business account for some reason - probably tax evasion. I'm not saying you are doing that, but it is a frequent issue. If the company makes the cheque out to a person they may run the risk of being party to fraud. Worse still they only have your word for it that you actually own the company, and aren't ripping off your employer by pocketing their payment. Even worse, when the company is audited and finds that cheque, the person who wrote it will have to justify and document why they made it out to you or risk being charged with embezzlement. It's very much in their interests to make the cheque out to the company they did business with. Given that, you should really have an account in the name of your business. It's going to make your life much simpler in the long run.]	0.0883008512583046	0.6666666666666667	0.9774455844595044	0.1111111111111111	0
1	Can I send a money order from USPS as a business?	[Sure you can. You can fill in whatever you want in the From section of a money order, so your business name and address would be fine. The price only includes the money order itself. You can hand deliver it yourself if you want, but if you want to mail it, you'll have to provide an envelope and a stamp. Note that, since you won't have a bank record of this payment, you'll want to make sure you keep other records, such as the stub of the money order. You should probably also ask the contractor to give you a receipt.]	Yes, you can send a money order from USPS as a business. You can fill in whatever you want in the From section of the money order, including your business name and address. The price only includes the money order itself, so you will need to provide an envelope and a stamp if you want to mail it. It is important to keep records of the payment, such as the stub of the money order, and to ask the contractor for a receipt.	[Sure you can. You can fill in whatever you want in the From section of a money order, so your business name and address would be fine. The price only includes the money order itself. You can hand deliver it yourself if you want, but if you want to mail it, you'll have to provide an envelope and a stamp. Note that, since you won't have a bank record of this payment, you'll want to make sure you keep other records, such as the stub of the money order. You should probably also ask the contractor to give you a receipt.]	0.23101050853729246	1.0	0.8830298074690555	0.8	0

Рисунок 4.11 – Результати оцінки чат-бота

Результати оцінювання розробленої системи чат-бота демонструють високу ефективність обраного підходу оскільки згенеровані відповіді були оцінені дуже високо.

## ВИСНОВКИ

В роботі розглянуто проблему побудови відповідей чат-ботів з використанням великих мовних моделей. Запропоновано узагальнювати та в подальшому використовувати також узагальнену інформацію про предметну область.

В роботі виконано аналіз методів адаптації систем чат-ботів для систем електронної комерції, удосконалено метод формування відповіді чат-бота, розроблено інформаційну технологію використання локальних мовних моделей для побудови відповіді чат-бота в системі електронної комерції.

Удосконалено метод формування відповіді чат-бота з використанням мовних моделей шляхом додаткового пошуку прототипів відповідей в узагальненій інформації щодо роботи системи електронної комерції, в якій функціонує чат-бот.

Проведена програмна реалізація вдосконаленого методу побудови контексту для мовної моделі у системах чат-ботів, побудованих за RAG архітектурою, з подальшою перевіркою, включаючи створення запитань, що вимагають специфічних знань з предметної області. Розроблена система здатна генерувати релевантні та зрозумілі відповіді та була високо оцінена за більшістю метрик.

Отримані результати можуть стати важливим внеском у майбутні дослідження та бути використані для створення сучасних систем чат-ботів. Крім того, вони можуть служити основою для вдосконалення існуючих систем з метою покращення їх здатності генерувати релевантні та детальні відповіді.

## ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Переваги електронної комерції для бізнесу та споживачів [Електронний ресурс] – Режим доступу: <https://factory.dev/blog/ecommerce-benefits-businesses-consumers>
2. Ринок e-commerce в Україні в 2020 [Електронний ресурс] – Режим доступу: <https://rau.ua/novyni/e-commerce-v-ukrayini-2020/>
3. Закон України про електронну комерцію [Електронний ресурс] – Режим доступу: <https://zakon.rada.gov.ua/laws/show/675-19#Text>
4. Weiland D. The Role Of Information In E-Commerce. Research journal of the university of Gdańsk. 2017. Vol.68. P. 105-106
5. What Is RAG? (Retrieval Augmented Generation) [Електронний ресурс] – Режим доступу: <https://www.clarifai.com/blog/what-is-rag-retrieval-augmented-generation>
6. Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Küttler H., Lewis M., Yih W., Rocktäschel T., Riedel S., Kiela D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020. Vol.793. P. 1-2
7. Summarization: Improving RAG quality in LLM apps while minimizing vector storage costs споживачів [Електронний ресурс] – Режим доступу: <https://www.ninetack.io/post/improving-rag-quality-by-summarization>
8. Методичні вказівки щодо розробки та оформлення кваліфікаційної роботи (для студентів усіх форм навчання другого (магістерського) рівня програми "Інформаційні управляючі системи та технології") / Упоряд.: Петров К.Е., Левикін В.М., Чалий С.Ф., Євланов М.В., Саєнко В.І., Міхнов Д.К., Міхнова А.В., Чала О.В. Харків: ХНУРЕ, 2021. 30 с.

9. ДСТУ 3008:2015. Інформація та документація. Звіти у сфері науки і техніки. Структура і правила оформлювання. – Чинний від 22.06.2015. – Київ: 105 ДП «УкрНДНЦ», 2016. – 31 с. .

10. ДСТУ 8302:2015. Інформація та документація. Бібліографічні посилання. Загальні положення та правила складання. – Чинний від 04.03.2016. – Київ: ДП «УкрНДНЦ», 2016. – 20 с.