

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Система технічної підтримки клієнтів на основі
штучного інтелекту та Natural Language Processing»

на здобуття освітнього ступеня магістра
зі спеціальності 122 Комп'ютерні науки
(код, найменування спеціальності)
освітньо-професійної програми Комп'ютерні науки
(назва)

*Кваліфікаційна робота містить результати власних досліджень.
Використання ідей, результатів і текстів інших авторів мають посилання
на відповідне джерело*

(підпис)

Рябков Данило
(Ім'я, ПРІЗВИЩЕ здобувача)

Виконав:
здобувач вищої освіти
група КНДМ-61

Данило РЯБКОВ

Керівник:
науковий ступінь,
вчене звання

Олександр ЗВЕНИГОРОДСЬКИЙ
д.т.н., професор

Рецензент:
науковий ступінь,
вчене звання

(Ім'я, ПРІЗВИЩЕ)

Київ 2024

**ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ**

Навчально-науковий інститут інформаційних технологій

Кафедра Комп'ютерних наук

Ступінь вищої освіти Магістр

Спеціальність 122 Комп'ютерні науки

Освітньо-професійна програма Комп'ютерні науки

ЗАТВЕРДЖУЮ

Завідувач кафедру Комп'ютерних наук

_____ Віктор ВИШНІВСЬКИЙ
« _____ » _____ 2024 р.

**ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ**

_____ Рябкову Данилу Олександровичу

(прізвище, ім'я, по батькові здобувача)

1. Тема кваліфікаційної роботи: «Система технічної підтримки клієнтів на основі штучного інтелекту та Natural Language Processing»

керівник кваліфікаційної роботи Олександр ЗВЕНІГОРОДСЬКИЙ д.т.н.,
професор,

(Ім'я, ПРИЗВИЩЕ науковий ступінь, вчене звання)

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «15» жовтня 2024р. №320

2. Строк подання кваліфікаційної роботи «15» грудня 2024р.

3. Вихідні дані до кваліфікаційної роботи: науково-технічна література, методи статистичного аналізу та візуалізації даних, методи розробки програмного забезпечення.

4. Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

4.1. Дослідження ситем та можливостей NLP.

4.2. Проектування прототипу системи та підбір технологій для її реалізації.

4.3. Розробка прототипу системи візуалізації.

5. Перелік графічного матеріалу: *презентація*

5.1 Основні методи візуалізації даних за допомогою діаграм та таблиць.

5.2 Підходи до проектування прототипу системи.

5.3 Приклади програмного коду.

5.4 Головні компоненти створеної системи.

6. Дата видачі завдання «16» вересня 2024 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Аналіз наявної науково-технічної літератури	16.09-27.09.24	виконано
2	Дослідження ситем та можливостей NLP	30.09-11.10.24	виконано
3	Проектування та підбір технологій реалізації	14.10-25.10.24	виконано
4	Розробка прототипу	28.10-08.11.24	виконано
6	Головні компоненти створеної системи	25.11-29.11.24	виконано
7	Оформлення роботи: вступ, висновки, реферат	02.12-06.12.24	виконано
8	Розробка демонстраційних матеріалів	09.12-13.12.24	виконано

Здобувач вищої освіти

(підпис)

Данило РЯБКОВ

(Ім'я, ПРІЗВИЩЕ)

Керівник

кваліфікаційної роботи

(підпис)

Олександр ЗВЕНІГОРОДСЬКИЙ

(Ім'я, ПРІЗВИЩЕ)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 82 стор., 34 рис., 22 джерела.

Мета роботи – розробка Proof of Concept (PoC) системи автоматизації обробки клієнтських запитів із використанням генеративного штучного інтелекту (NLP, Gen AI), спрямованої на підвищення ефективності, швидкості та якості обслуговування клієнтів.

Об'єкт дослідження – процеси автоматизації обробки клієнтських запитів у сучасних організаціях.

Предмет дослідження – Proof of Concept (PoC) система автоматизації, заснована на технологіях NLP, Generative AI та RAG.

Короткий зміст роботи: Кваліфікаційна робота присвячена створенню концептуального прототипу системи автоматизації обробки клієнтських запитів із використанням генеративного штучного інтелекту.

Робота складається з наступних основних частин:

- Аналіз розвитку технологій Generative AI та NLP.
- Дослідження методів інтеграції технологій Gen. AI в організаційні процеси.
- Розробка та оцінка PoC системи автоматизації клієнтських запитів.

Під час виконання роботи було досліджено основні концепції та методології у сфері Generative AI та NLP. Розглянуто ключові можливості та обмеження. Розроблено PoC-систему автоматизації запитів із використанням RAG для підвищення релевантності та точності відповідей. Проведено тестування функціональності, оцінено ефективність та обґрунтовано потенційний вплив системи на організаційні процеси.

КЛЮЧОВІ СЛОВА: GENERATIVE AI, АВТОМАТИЗАЦІЯ ОБРОБКИ ЗАПИТІВ, PROOF OF CONCEPT, RETRIEVAL-AUGMENTED GENERATION, КЛІЄНТСЬКИЙ СЕРВІС, ВЕЛИКІ МОВНІ МОДЕЛІ.

ABSTRACT

The textual part of the qualification thesis for obtaining the master's degree: 82 pages, 34 figures, 22 references.

The purpose of the work is to develop a Proof of Concept (PoC) system for automating customer query processing using Generative Artificial Intelligence (NLP, Gen AI), aimed at improving efficiency, speed, and quality of customer service.

The object of the study is the processes of automating customer query processing in modern organizations.

The subject of the study is a Proof of Concept (PoC) automation system based on NLP, Generative AI, and Retrieval-Augmented Generation (RAG) technologies.

Summary of the work: The qualification thesis is devoted to the creation of a conceptual prototype of a system for automating customer query processing using Generative Artificial Intelligence.

The work consists of the following main parts:

- Analysis of the development of Generative AI and NLP technologies.
- Research on methods for integrating Generative AI technologies into organizational processes.
- Development and evaluation of a PoC system for automating customer queries.

During the work, the main concepts and methodologies in the field of Generative AI and NLP were studied. Key capabilities and limitations were considered. A PoC query automation system was developed using RAG to enhance the relevance and accuracy of responses. Functionality testing was conducted, effectiveness was evaluated, and the potential impact of the system on organizational processes was substantiated.

KEYWORDS: GENERATIVE AI, QUERY PROCESSING AUTOMATION, PROOF OF CONCEPT, RETRIEVAL-AUGMENTED GENERATION, CUSTOMER SERVICE, LARGE LANGUAGE MODELS.

ЗМІСТ

ВСТУП.....	10
1 АНАЛІЗ РОЗВИТКУ ТЕХНОЛОГІЙ GEN AI та NLP	11
1.1 Передумови.....	11
1.2 Мета дослідження та обсяг	12
1.3 Методи дослідження	13
2 Інформаційні технології як драйвер створення цінності	15
2.1 Процес створення цінності.....	15
2.2 Моделі створення цінності за допомогою ІТ	16
2.3 Управління знаннями.....	20
3 Генеративний штучний інтелект	23
3.1 Штучний інтелект	23
3.2 Генеративний штучний інтелект (Генеративний ШІ)	24
3.3 Великі мовні моделі (Large Language Models, LLM).....	25
3.4 Векторні подання (embeddings)	28
3.5 Генерація з доповненням отримання	29
3.6 Застосування генеративного ШІ у клієнтському сервісі	32
3.7 Впровадження генеративного ШІ в організаційному контексті	33
4 Дослідження дизайну	35
4.1 Методологія Action Design Research (ADR).....	36
4.2 Вибір Action Design Research (ADR) для дослідження	37
5 Розробка та оцінка артефакту	40
5.1 Proof of Concept (PoC).....	40
5.2 Ресурси	42
5.3 Дані	43
5.4 Процес дизайну	45
5.5 Архітектура.....	47
5.6 Приклади запитів користувачів	48
5.7 Інженерія підказок	49
5.8 Оцінка.....	51
6 Обговорення.....	57

6.1 Продуктивність RoC-асистента.....	57
6.2 Вплив на організаційні процеси	59
6.3 Генеративний ШІ як драйвер цінності.....	62
6.4 Стратегічний потенціал RoC для обробки запитів.....	63
6.5 Інтеграція, управління змінами та перспективи	65
6.6 Обмеження та напрямки для вдосконалення	65
ВИСНОВКИ.....	67
Перелік посилань.....	68
ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ	72

ВСТУП

У сучасному світі компанії стикаються з постійно зростаючим потоком звернень клієнтів до служби технічної підтримки. Традиційні методи обробки запитів часто виявляються неефективними, призводять до тривалого часу очікування та незадоволеності користувачів. Впровадження машинного навчання в існуючі бізнес процеси підтримки користувачів призведе до уніфікації загально поширених проблем користувачів, які виникають, автоматизує і пришвидшить отримання відповідей на них. Ця магістерська робота присвячена дослідженню сучасних методологій впровадження застосунків генеративного штучного інтелекту (Generative Artificial Intelligence, Gen AI). Основна мета дослідження — розробка рішення для оптимізації внутрішніх процесів обробки запитів клієнтів.

Застосунок на базі генеративного штучного інтелекту буде представлений як доказ концепції (Proof of Concept, PoC), який дозволить оцінити доцільність, потенціал та функціональність нового підходу до обробки запитів. Головна задача дослідження — вивчити сучасні методології та ключові можливості організації для розробки та впровадження рішень Gen AI та NLP у підтримці клієнтів.

1 АНАЛІЗ РОЗВИТКУ ТЕХНОЛОГІЙ GEN AI ТА NLP

1.1 Передумови

Інтеграція машинного навчання в існуючі бізнес-процеси підтримки клієнтів може значно підвищити їхню ефективність. Це дозволить уніфікувати найпоширеніші проблеми, автоматизувати вирішення повторюваних питань і прискорити процес надання відповідей. Технології машинного навчання забезпечують можливість виявлення шаблонів у запитах клієнтів, що дає змогу швидко знаходити відповідні рішення.

Ця магістерська робота спрямована на дослідження сучасних методологій впровадження застосунків генеративного штучного інтелекту (Generative Artificial Intelligence, Gen AI). Основною метою є створення інноваційного рішення для оптимізації внутрішніх процесів обробки клієнтських запитів.

Запропонований застосунок на базі генеративного штучного інтелекту буде представлено як доказ концепції (Proof of Concept, PoC). Такий підхід дозволить оцінити доцільність використання нових технологій, потенціал автоматизації процесів і функціональність запропонованого рішення.

Розробка цього рішення передбачає дослідження сучасних методологій і визначення ключових можливостей організації для впровадження рішень на базі Gen AI та NLP.

Однією з найбільш революційних гілок штучного інтелекту, що з'явилися останніми роками, є генеративний штучний інтелект (Generative AI). Ця технологія спеціалізується на створенні нового, унікального контенту на основі навчальних даних, які можуть включати текст, зображення, музику або навіть відео. Генеративний ШІ відкрив величезний спектр нових можливостей, від творчих додатків до автоматизованих систем обслуговування клієнтів.

Значущим етапом у розвитку і популяризації Gen AI стало випуск ChatGPT компанією OpenAI у листопаді 2022 року. Ця подія стала каталізатором глобальної гонки за бізнес-можливостями, у якій брали участь стартапи, розробники додатків і найбільші корпорації світу.

До 2024 року інтеграція генеративного штучного інтелекту у бізнес-операції стала звичним явищем. Проте впровадження таких технологій супроводжується низкою викликів. Зокрема, це вимагає значних ресурсів, спеціальних навичок і глибокого розуміння технологій. Крім того, академічних досліджень у цій сфері все ще недостатньо, що ускладнює адаптацію інструментів для специфічних бізнес-потреб.

Таким чином, у рамках цього дослідження ставиться мета не лише розробити застосунок, але й продемонструвати його ефективність та практичну цінність через Proof of Concept (PoC), що може послужити основою для подальшого розвитку технологій у цій сфері.

1.2 Мета дослідження та обсяг

Основна мета цього дослідження полягає в розробці інноваційного рішення для оптимізації внутрішніх процесів обробки клієнтських запитів у компаніях. Дослідження зосереджене на вивченні сучасних методологій інтеграції штучного інтелекту, зокрема генеративного ШІ.

Завдяки глибокому аналізу поточного стану технологій і процесів, дослідження прагне досягти ефективного впровадження генеративного штучного інтелекту (Gen AI) у реальні бізнес-операції. Для цього запропоноване рішення функціонуватиме як Proof of Concept (PoC), що продемонструє його можливості для подальшого розвитку. Це підкреслює експериментальний характер роботи, оскільки її результати спрямовані на практичне застосування.

Процес обробки запитів у багатьох компаніях здійснюється вручну командами, що відповідають за виконання цих завдань. Часто цей процес потребує залучення співробітників з інших підрозділів, що збільшує час на підготовку

відповідей. Такий підхід не лише витратний, але й знижує оперативність реагування.

У рамках дослідження ставляться два ключові питання:

- Які основні можливості необхідні для розробки та впровадження генеративних ШІ-застосунків в організаційному середовищі?
- Наскільки ефективним може бути PoC-асистент у вирішенні запитів клієнтів, пов'язаних із питаннями сталого розвитку, на основі доступних знань?

Це дослідження зосереджується на внутрішніх процесах компанії, зокрема на обробці клієнтських запитів, що стосуються питань сталого розвитку. У ньому також розглядаються ключові аспекти генеративного ШІ, великі мовні моделі (LLMs) та концепції генерації з доповненням пошуком (Retrieval Augmented Generation, RAG).

1.3 Методи дослідження

Успішне виконання дослідження вимагає ретельного вибору відповідних методів, які найкраще відповідають цілям і завданням роботи. Вибір методів повинен базуватися на логічному обґрунтуванні та забезпечувати максимальну ефективність отриманих результатів.

Основною метою цього дослідження є створення артефакту, що демонструє практичну цінність розробки. З огляду на це, було обрано конструктивний підхід до дослідження, відомий як **Action Design Science Research (ADR)**. Цей метод є ідеальним для організаційного контексту, оскільки він дозволяє досліджувати складні системи та водночас розробляти інноваційні рішення. Детальний опис етапів ADR представлено в наступних розділах роботи.

Для створення артефакту досліднику необхідно зібрати релевантну та достовірну інформацію. З цією метою у рамках роботи використовується неконтрольований огляд літератури. Огляд літератури спрямований на:

- визначення ключових питань і тенденцій у досліджуваній сфері;
- підтвердження або спростування існуючих теорій та концепцій;

- ідентифікацію прогалин або суперечностей у попередніх дослідженнях.

Теоретична база дослідження здебільшого будується на академічній літературі, щоб забезпечити необхідний рівень достовірності та науковості. Цей підхід дозволяє сформулювати цілісний погляд на проблему, розробити артефакт і оцінити його відповідність контексту та ефективність в організаційному середовищі.

Паралельно з теоретичним аналізом у дослідженні передбачено практичну реалізацію рішення. Цей підхід включає створення PoC, який дозволяє не лише продемонструвати можливості застосування, але й оцінити його функціональність та потенціал для подальшого впровадження.

Методологія ADR у поєднанні з глибоким аналізом літератури та експериментальним підходом робить це дослідження комплексним і спрямованим на вирішення конкретних бізнес-проблем.

2 ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ ЯК ДРАЙВЕР СТВОРЕННЯ ЦІННОСТІ

Цей розділ оглядає відповідну академічну літературу про інформаційні технології (ІТ), їхню роль у процесі створення цінності та управлінні знаннями в організаціях. Розділ поділено на дві частини. У першій частині ІТ розглядаються як складова організації та аналізуються як драйвер створення цінності в організаційному контексті на основі двох інтегративних моделей. Друга частина зосереджується на аналізі практик управління знаннями та впливу генеративного штучного інтелекту (Gen AI) на ці процеси, що розглядається в наступному розділі.

2.1 Процес створення цінності

Здатність адаптуватися до ринкових можливостей є ключовим фактором успіху для бізнесу в сучасному середовищі. Глобалізація, високий рівень конкуренції, зміна потреб клієнтів і швидкі технологічні зміни створюють умови, за яких підтримання конкурентних переваг стає складним завданням [3, с. 1]. Технології є практичною реалізацією знань та інформації для створення пристроїв і систем, що вдосконалюють процеси. Впродовж століть технології були основним двигуном інновацій та розвитку умов життя. У бізнесі саме організації стали найактивнішими прихильниками технологій, адже вони дозволяють підвищувати ефективність і продуктивність.

Сьогодні компанії не можуть існувати та успішно конкурувати без ІТ. Вони стали базовим елементом бізнесу, забезпечуючи автоматизовані процеси та аналітику, необхідну для управління [19, с. 3]. ІТ дозволяють перетворити організації з реактивних на проактивні, змінюючи всі аспекти їх функціонування. Однак самі по собі інвестиції в ІТ не гарантують створення цінності чи зростання прибутковості. Щоб отримати цінність від ІТ, організація повинна володіти ІТ-спроможністю — здатністю ефективно використовувати ІТ-ресурси разом з іншими ресурсами [23, с. 94-95].

Стратегії компаній повинні інтегрувати ІТ у загальну корпоративну стратегію, щоб використовувати ІТ як інструмент досягнення стратегічних цілей. ІТ-стратегія має бути взаємопов'язаною з корпоративною стратегією і сприяти створенню нових конкурентних переваг [19, с. 4-6].

ІТ-інфраструктура є основою для всіх ІТ-ресурсів і послуг компанії. Це не лише апаратне забезпечення, але й операційні системи, бази даних, інструменти адміністрування, дата-центри та сервери. Гнучкість цієї інфраструктури має вирішальне значення для здатності організації адаптуватися до змін [11, с. 1].

Інновації також є критично важливими для конкурентоспроможності. Хоча самі ІТ не створюють інновацій, вони забезпечують умови для розвитку інноваційних продуктів і послуг. Нещодавні технологічні прориви, такі як хмарні обчислення, великі дані, Інтернет речей та штучний інтелект, надають величезний потенціал [27].

2.2 Моделі створення цінності за допомогою ІТ

Мелвіла та Краймер [26] запропонували модель, яка інтегрує елементи, що впливають на роль ІТ у підвищенні організаційної ефективності. У центрі моделі — використання внутрішніх можливостей організації для створення бізнес-цінності у зовнішньому середовищі. До ключових компонентів моделі належать:

1) Ресурси:

- Технологічні ресурси (TIR): інфраструктура та бізнес-додатки.
- Людські ресурси (HIR): технічні та управлінські знання.
- Організаційні ресурси: політики, культура, різноманітність.

2) Процеси та результати:

- Підвищення ефективності бізнес-процесів.
- Організаційна продуктивність: зростання доходів, зниження витрат.

3) Зовнішні фактори:

- Конкурентне середовище: динаміка ринку, швидкість змін.
- Макроекономічні умови: регуляції, культурні аспекти.

Гровер, В. та ін. [13, с. 398] пропонують концептуальну двокомпонентну модель, яка включає розвиток спроможностей і їх реалізацію, для пояснення, як організації

можуть використовувати великі дані (BDA) з метою покращення процесів прийняття рішень, ефективності та інновацій. Як показано на рисунку 2.1, розвиток спроможностей починається зі створення інфраструктури для підтримки BDA. Ключовими елементами цієї інфраструктури є активи великих даних, портфель аналітичних інструментів і людські ресурси. Саме людські ресурси є найбільш важливим компонентом у цій системі, оскільки для реалізації стратегій BDA потрібні досвід і експертні знання. Коли інфраструктура готова, організація може створити спроможності для інтеграції, управління, аналізу та обміну даними в різних формах, що дозволяє створювати цінність.

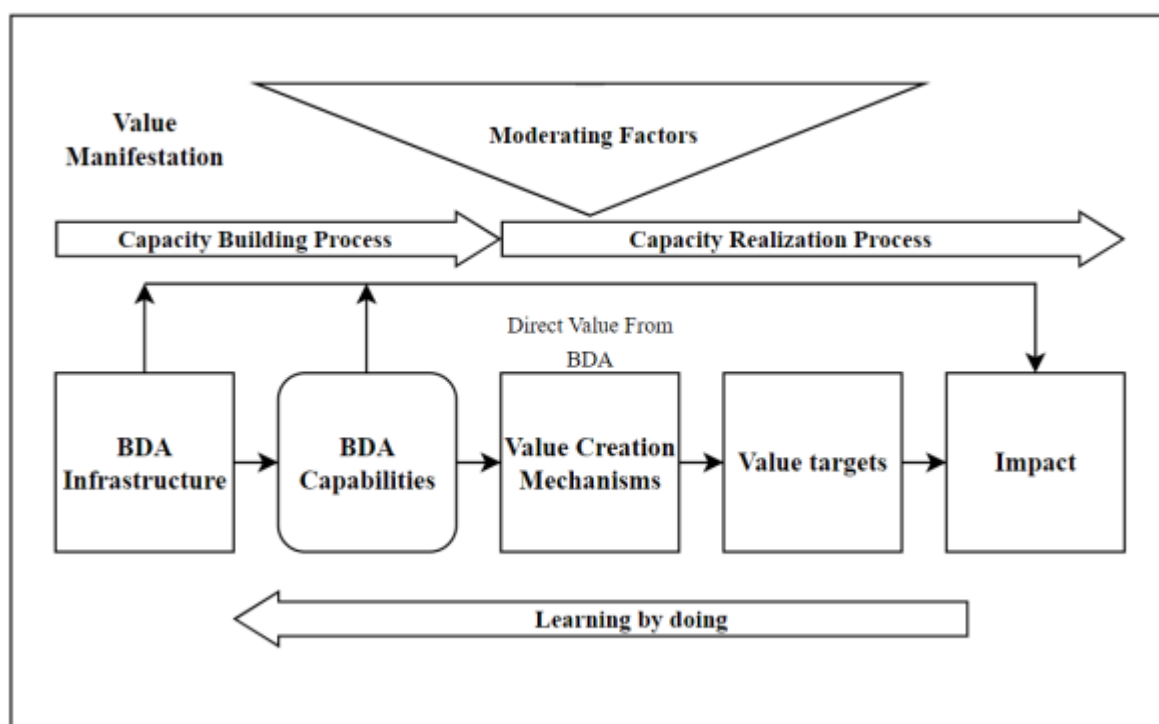


Рисунок 2.1 – Створення цінності за допомогою аналізу великих даних

Після процесу розбудови спроможностей Гровер та ін. переходить до процесу реалізації спроможностей до процесу реалізації потенціалу. Вони пропонують шість різних механізмів, які можуть призвести до створення цінності, проілюстровані в таблиці 2.1.

Таблиця 2.1 – Механізми створення цінності, запропоновані Гровер та ін. [13, с. 401-403]

Механізм	Значення
Прозорість та доступ	Зробити інформацію легкодоступною та доступною для всіх відповідним зацікавленим сторонам в організації, гарантуючи, що особи, які приймають рішення на всіх рівнях особи, які приймають рішення на всіх рівнях, мають необхідну інформацію, щоб розуміти та ефективно реагувати на умови ведення бізнесу.
Відкриття та експерименти практики	Дозволяє організаціям відкривати нові ідеї та закономірності з величезних масивів даних, підтримуючи експериментальну перевірку гіпотез гіпотез в реальних умовах, що дозволяє впроваджувати інновації та зниження ризиків
Прогнозування та оптимізація	Використовує предиктивну аналітику для прогнозування майбутніх тенденцій, поведінки та ринкових змін, що дозволяє бізнесу готуватися та адаптуватися проактивно. Алгоритми оптимізації забезпечують ефективний розподіл ресурсів для максимізації результатів на основі прогнозних інсайтів.
Кастомізація та таргетинг	Сприяє створенню високо персоналізованого клієнтського досвіду шляхом пристосовуючи продукти, послуги та взаємодію до індивідуальних вподобаннями та поведінкою, а також точно орієнтуючись на конкретні клієнтські сегментів клієнтів з точністю, підвищуючи ефективність маркетингу та підвищуючи задоволеність клієнтів.
Навчання та краудсорсинг	Підтримує середовище безперервного навчання, дозволяючи аналіз нових потоків даних для постійного розуміння. Краудсорсинг використовує колективні знання та та досвід великої групи, часто з-поза меж організації, для вирішення проблем і генерування інноваційних організації, для вирішення проблем і генерування інноваційних ідей.
Безперервний моніторинг та проактивна адаптація	Включає моніторинг систем і процесів у режимі реального часу для виявлення та реагування на проблеми до їх ескалації, що дозволяє бізнесу передбачати проблеми та коригувати стратегії динамічно коригувати стратегії, підтримуючи ефективність та запобігаючи потенційним збоям.

Після опису механізмів створення цінності, Гровер та ін. [13, с. 402] у своїй моделі розглядають цілі створення цінності. Дослідники виділяють чотири основні цілі. Першою є підвищення ефективності організації. Завдяки використанню великих даних (BDA) організація може покращити процеси прийняття рішень. Це вимагає прозорості даних у межах організації, що дозволяє будувати моделі та отримувати інсайти.

Друга ціль — вдосконалення бізнес-процесів. Завдяки BDA можна визначати сильні та слабкі сторони бізнес-процесів на основі емпіричного аналізу великих обсягів даних.

Третя — отримання цінності від інновацій у продуктах і послугах. Аналіз даних, зокрема кліків користувачів чи їхніх покупок, може сприяти розробці промоційних стратегій або індивідуалізованих вебінтерфейсів. Збір та аналіз інсайтів щодо досвіду клієнтів допомагає виявити часті проблеми з продуктами чи послугами, а також визначити, що саме клієнти бажають.

Четверта ціль — покращення клієнтського досвіду та розширення ринкових можливостей через виявлення конкурентної розвідки або залучення нових клієнтів за допомогою аналітики соціальних мереж. Наприклад, конкурентів можна ідентифікувати через огляди та коментарі.

На вершині моделі Гровер та ін. [13] розташовані модераційні фактори. До них належать стратегія, лідерство, довіра, технологічний та галузевий контекст, підтримка управління, конкурентна динаміка та культура, заснована на даних. Ці фактори сприяють перетворенню інвестицій у BDA на цінність.

Стратегія забезпечує відповідність ініціатив BDA цілям організації, а сильне лідерство сприяє їх впровадженню. Технологічний та галузевий контекст впливають на застосовність та ефективність BDA, що вимагає адаптації до конкретних умов. Підтримка управління гарантує, що практики управління даними відповідають стандартам цілісності та безпеки. Конкурентна динаміка, яку також виділяли в моделі Мелвіля та ін. [26], визначає, наскільки активно організація інвестує в BDA для збереження або покращення своєї ринкової позиції. Культура,

заснована на даних, є основою всіх зусиль, інтегруючи дані у процеси організації та прийняття рішень, що максимізує потенціал створення цінності з BDA.

Довіра в організації стимулює використання даних для прийняття рішень. Якщо довіри немає, BDA-рішення не будуть застосовуватися [13, с. 403].

Концепція довіри також врахована у сучасних дослідженнях штучного інтелекту. Крепс, Джордж і Рао [23] запропонували поняття «парадоксу довіри», який вони визначають як «готовність громадськості підтримувати або використовувати технології, що базуються на ШІ, яким вони не довіряють». У контексті організацій розуміння та врахування цього «парадоксу довіри» є важливим для ефективного впровадження як BDA, так і технологій ШІ.

2.3 Управління знаннями

Управління знаннями (Knowledge Management, KM) тісно пов'язане з інформаційними технологіями (IT), і разом вони відіграють ключову роль у створенні цінності для організації. Awad [2, с. 26–28] зазначає, що розвиток інформаційних технологій спричинив виникнення розривів у доступності та управлінні інформацією й знаннями. Хоча ці розриви поступово заповнюються, залишаються кілька важливих викликів: як використовувати технології, щоб перетворити проблеми на можливості? Як створювати знання з інформації? Як генерувати цінність зі знань? Ефективне впровадження практик KM може допомогти відповісти на ці питання.

KM має багато визначень, але основним його завданням є перебудова технічних можливостей організації для створення знань, що сприяють її розвитку. KM спрямоване на те, щоб знання, зібрані в бізнес-процесах, продуктах і послугах, використовувалися прозоро та сприяли зростанню організації. Наприклад, Шопфлін та Уолш [33, с. 3] визначають KM як набір методичних стратегій, за допомогою яких організація оптимізує використання існуючих знань і експертизи для досягнення успіху інформація та знання: піраміда знань

Терміни "дані", "інформація" та "знання" часто вживаються як синоніми, хоча між ними існують відмінності. Для пояснення цих понять Девенпорт і Прусак [8] запропонували піраміду знань (рис. 2.2).

- Дані — це основа піраміди, зібрана та збережена за допомогою різних інструментів і технологій. Вони не мають великої цінності, якщо не організовані чи оброблені. Їхня цінність визначається тим, де і як вони можуть бути використані.
- Інформація — це дані, які були оброблені, збагачені й отримали значення та мету.
- Знання — це результат синтезу інформації з різних джерел, що дозволяє виявляти закономірності, тренди та зв'язки. Воно включає розуміння, інсайти та досвід, що перетворюють інформацію на практичні рекомендації для прийняття рішень.

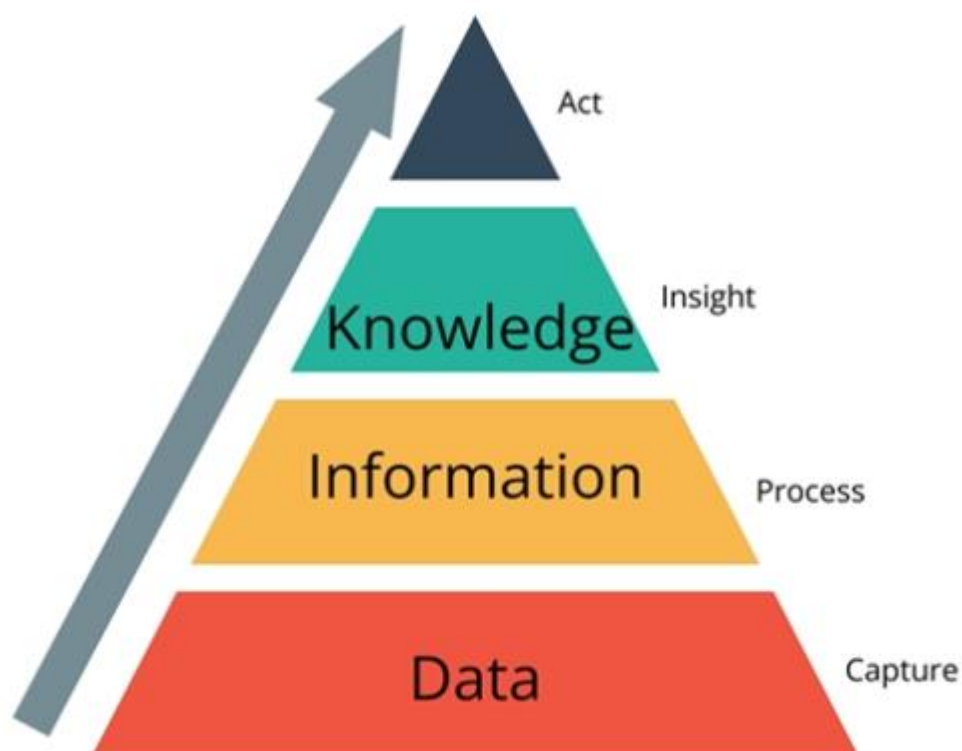


Рисунок 2.2 – Дані, інформація та знання: піраміда знань

Знання в організації поділяються на два основних типи: явні і неявні [5, с. 27].

- Неявні знання — це знання, що виникають із дій, досвіду та участі людей у конкретних контекстах. Вони включають когнітивні елементи (ментальні моделі, переконання) і технічні (практичні навички, "ноу-хау"). До 95% знань в організації вважаються неявними, і саме вони лежать в основі створення явних знань.
- Явні знання — це формалізовані, систематизовані знання, які можна легко передати чи задокументувати, наприклад у звітах чи документах.

У піраміді знань неявні знання розташовані на верхівці, адже вони менш помітні й важче формалізуються. Явні знання також належать до вершини піраміди, оскільки вони є результатом обробки даних і інформації.

Штучний інтелект (ШІ), зокрема генеративний ШІ, має значний вплив на практики КМ. Джаррахі та ін. [17] дослідили потенціал ШІ у підтримці КМ, включаючи створення, зберігання, пошук, обмін і застосування знань.

ШІ суттєво покращує процеси зберігання та пошуку явних знань. Наприклад, системи, засновані на AI, можуть класифікувати, організовувати та знаходити знання швидше, ніж традиційні методи. Одним із таких рішень є Retrieval-Augmented Generation (RAG). ШІ сприяє застосуванню знань через чат-ботів і цифрових асистентів, які знижують бар'єри доступу до знань і забезпечують їх рівномірний розподіл у межах організації. Застосування ШІ у КМ дозволяє створювати більш ефективні й доступні рішення для організаційного зростання.

3 ГЕНЕРАТИВНИЙ ШТУЧНИЙ ІНТЕЛЕКТ

Третя частина дослідження спрямована на вивчення релевантної літератури та інших джерел для створення ґрунтового розуміння концепцій ШІ та генеративного ШІ. У цьому розділі також розглядаються основні елементи, що забезпечують роботу додатків генеративного ШІ. Передостанній розділ цієї частини аналізує застосування генеративного ШІ у сфері обслуговування клієнтів. Останній розділ досліджує впровадження генеративного ШІ у контексті організацій і намагається знайти зв'язок між другим і третім розділами цього дослідження. Мета цього розділу — не надавати вичерпного пояснення функціональних механізмів технологій, що вивчаються.

3.1 Штучний інтелект

Штучний інтелект (ШІ) охоплює майже всі аспекти інженерії та виходить за їх межі. Існує безліч академічних визначень ШІ, однак визначення, наведене Хенлейн і Каплан, часто цитують у наукових роботах. Вони визначають ШІ як «здатність системи коректно інтерпретувати зовнішні дані, навчатися на їх основі та використовувати ці знання для досягнення конкретних цілей і виконання завдань через гнучке пристосування»[38, с. 15-25].

ШІ вперше з'явився як академічна дисципліна в 1950-х роках, але протягом півстоліття не привертав значної уваги. Сьогодні ШІ є однією з найцікавіших тем у наукових дослідженнях, і компанії з усього світу змагаються за те, хто зможе використати його найкраще, з урахуванням нормативних обмежень, таких як Європейський закон про ШІ [10].

ШІ виступає як узагальнюючий термін, що охоплює різноманітні алгоритми, здатні виконувати завдання, які зазвичай потребують інтелекту, схожого на людський. Це завдання, такі як прийняття рішень, розуміння природної мови, навчання на основі досвіду й розпізнавання шаблонів [6]. Одним із підходів до реалізації ШІ є машинне навчання (ML). ML є підгалуззю ШІ, яка базується на

пошуку закономірностей у даних і навчанні на їх основі [29]. ML це набір методів, що дозволяють автоматично виявляти закономірності в даних, а потім використовувати ці закономірності для прогнозування майбутніх даних або прийняття рішень в умовах невизначеності [30]. Замість того щоб моделювати поведінку системи на основі фізичних законів, ML використовує дані для навчання кореляціям. ML ідеально підходить для виявлення аномалій, оскільки здатне вчитися на досвіді. Якість і доступність даних є критично важливими для успішного використання ML [24]. ML далі розвивається в глибоке навчання (DL), де багатоваршова нейронна мережа тренується на великому обсязі історичних даних для виконання конкретних завдань. У DL мережа поступово витягує інформацію з даних, навчаючи комп'ютер імітувати людську поведінку [35, с. 7]. ML і DL формують основу ШІ, надаючи інструменти для створення систем, які можуть навчатися на даних, адаптуватися до нових сценаріїв і виконувати завдання, подібні до людських, із зростаючою точністю.

Ціль ШІ — не лише імітація людського інтелекту, але й створення програмного забезпечення, здатного ефективно вирішувати реальні проблеми [25]. Як і раніше, розвиток ШІ залишається динамічним і змінним, обіцяючи революційні зміни у взаємодії з технологіями й розумінні інтелекту. Впровадження ШІ сьогодні є важливим для підтримки конкурентної переваги у різних галузях, таких як охорона здоров'я, виробництво та споживчі товари.

3.2 Генеративний штучний інтелект (Генеративний ШІ)

Останніми роками з'явилися численні нові технології та концепції, пов'язані зі штучним інтелектом. Генеративні моделі ШІ привернули значну увагу як науковців, так і широкої громадськості через їхню здатність вирішувати складні завдання на основі мови та створювати креативний, схожий на людський контент. Генеративний ШІ може створювати зображення, відео, музику та консолідувати дані з різних джерел для аналізу [9, с. 4].

Історія генеративного ШІ сягає 1950-х років, коли були розроблені моделі прихованих марковських процесів (НММ) і гаусових сумішей (ГММ). НММ використовувалися для генерації послідовних даних, таких як прогнози погоди, а ГММ — для моделювання ймовірнісних розподілів [5, с. 4].

Найпоширеніші генеративні моделі базуються на так званих базових моделях (foundation models), що є великими штучними нейронними мережами. Ці моделі використовують глибокі нейронні мережі, які значно перевершують традиційні за складністю. Глибокі нейронні мережі складаються з багатошарової структури взаємопов'язаних нейронів, здатних навчатися на великих обсягах даних [22].

Базові моделі виступають як «попередньо треновані» моделі, що слугують основою для створення специфічних додатків або виконання завдань. Ці моделі доопрацьовуються для конкретних цілей чи сфер, використовуючи знання, отримані під час попереднього навчання, і спеціалізуються на основі нових даних. Вони також здатні до перенесення навчання, тобто знання, здобуті для однієї задачі, застосовуються для покращення виконання іншої задачі.

3.3 Великі мовні моделі (Large Language Models, LLM)

Однією з найпопулярніших тем у сфері генеративного ШІ є ChatGPT від OpenAI, який працює на основі GPT-4 — високопродуктивної великої мовної моделі (LLM). З моменту свого запуску ChatGPT привернув значну увагу завдяки своїм вражаючим можливостям. Популярність цієї моделі відображається в трендах пошуку, де шкала значень демонструє інтерес до запиту (рис. 3.1).

ChatGPT належить до класу мовних моделей, які називаються Generative Pre-trained Transformers (GPT) і входять до категорії LLM. Ці моделі базуються на технологіях глибокого навчання (DL) і проходять тренування на великих обсягах даних. ChatGPT спеціально оптимізований для ведення діалогу, що дозволяє створювати відповіді, схожі на ті, які формулює людина, завдяки широкій базі знань та інформації.

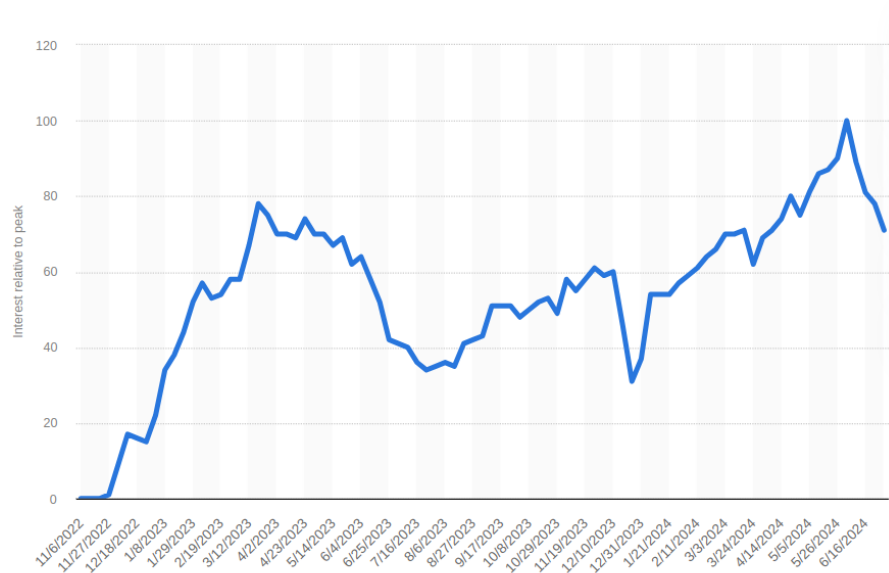


Рисунок 3.1 – Пошукова зацікавленість «ChatGPT» в Google по всьому світу

Основою LLM є обробка природної мови (NLP). Мета NLP — подолати розрив між людською мовою і розумінням комп'ютерів, що дозволяє машинам аналізувати дані й формувати відповіді природною мовою. Поєднання великих наборів даних, статистичних методів і глибокого навчання сприяло широкому впровадженню NLP на початку XXI століття. Поєднання слів зображено на рисунку 3.2 Це дало змогу розробити практичні застосунки, такі як віртуальні асистенти, чат-боти для підтримки клієнтів, інструменти аналізу настроїв і сервіси машинного перекладу.

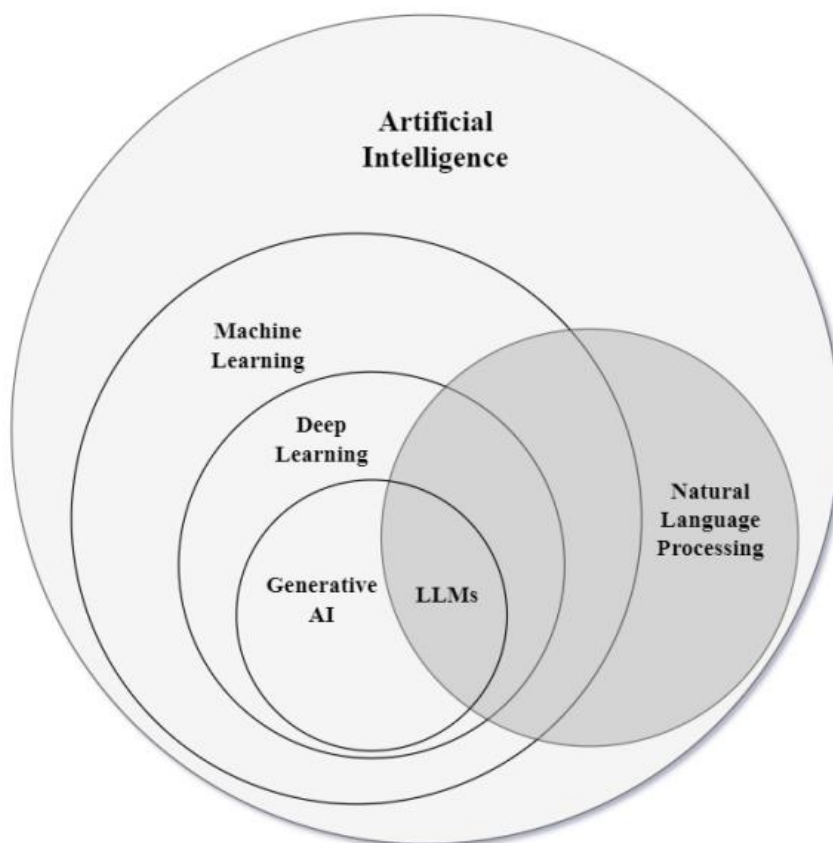


Рисунок 3.2 – Поєднання ІІІ слоїв

В обробці природної мови (NLP) існує безліч практичних сценаріїв використання. Одним із них є класифікація тексту, де модель присвоює тексту певну мітку або категорію. Наприклад, це може бути класифікація електронних листів як спам чи ні, тематичне категоризування або аналіз настроїв. Ще одним популярним застосуванням є машинний переклад, що автоматично перекладає текст із однієї мови на іншу. Генерація тексту дозволяє створювати тексти, які нагадують людські, виконувати їх підсумовування та автоматично створювати контент, наприклад, для чат-ботів.

NLP також забезпечує такі можливості, як розпізнавання мовлення, відповіді на запитання, моделювання мов і реалізацію різних генеративних додатків. Для досягнення цього NLP використовує низку ключових концепцій.

Токенізація розділяє текст на дрібніші одиниці, які називаються токенами, і є важливим етапом попередньої обробки тексту в більшості завдань NLP. Крім того, важливу роль відіграє парсинг, який аналізує взаємозв'язки між словами й фразами для створення контенту за допомогою векторного поєднання.

3.4 Векторні подання (embeddings)

Векторні подання являють собою метод представлення слів у вигляді точок у багатовимірному просторі, де кожна точка відображає певне слово чи фразу. У цьому просторі семантично схожі слова розташовані ближче одне до одного, а несхожі – далі. Такий підхід дозволяє фіксувати семантичні зв'язки між словами, надаючи можливість алгоритмам NLP розуміти синтаксичну і семантичну подібність, а також контекст використання слів у тексті. Ці векторні подання є фундаментальними для роботи великих мовних моделей (LLM), оскільки забезпечують можливість розуміння, аналізу й моделювання мовних нюансів. Завдяки їм LLM здатні враховувати контекст, розпізнавати значення в різних мовних ситуаціях і створювати тексти, що імітують людську мову.

Векторизація слів слугує ключовою технологією, яка робить можливим автоматизовану обробку, аналіз і генерацію текстів. Вона лежить в основі таких завдань, як семантичний пошук, текстова класифікація, машинний переклад, аналіз настроїв, а також створення контенту, подібного до людського.

Ця складна технологічна база дозволяє застосовувати генеративний штучний інтелект у вирішенні реальних завдань у різних галузях. Об'єднання методів машинного навчання, глибокого навчання та векторизації дало змогу значно покращити якість взаємодії між людиною й машиною. Це особливо помітно в таких сферах, як обслуговування клієнтів, переклад текстів, аналіз даних, створення оригінального контенту та багато іншого. Завдяки розвитку векторних подань та їх інтеграції в сучасні NLP-системи, багато галузей отримали доступ до інструментів, які раніше здавалися неможливими або занадто складними для реалізації. Наприклад, у сфері обслуговування клієнтів векторизація текстів дозволяє

автоматизувати процес обробки запитів, швидко і точно розпізнаючи їхній зміст і контекст. Це сприяє не лише економії часу, а й покращенню досвіду клієнтів, оскільки відповіді стають більш релевантними та точними. У перекладі текстів векторні подання забезпечують глибше розуміння мовних структур, завдяки чому машинні перекладачі досягають більшої точності. Наприклад, вони можуть враховувати культурні та семантичні відмінності між мовами, що робить переклади більш природними та зрозумілими.

Аналіз тексту, наприклад, в аналізі настроїв, використовує векторні подання для розпізнавання тональності тексту – позитивної, негативної або нейтральної. Це корисно для бізнесу, який аналізує відгуки клієнтів, соціальні мережі або дані з опитувань.

Генерація текстів, зокрема за допомогою генеративного ШІ, відкриває нові можливості в створенні контенту. Векторні подання дозволяють моделі створювати тексти, які відповідають стилю, тону та контексту, заданому користувачем. Це має великий потенціал у таких сферах, як написання статей, створення рекламних матеріалів або навіть творчий контент, як-от сценарії чи поезія.

Таким чином, об'єднання векторизації, моделей глибокого навчання та алгоритмів машинного навчання дозволяє NLP системам досягти нового рівня точності та функціональності. Ці досягнення відкривають двері до подальших інновацій у взаємодії між людьми та технологіями, розширюючи горизонти їх практичного використання.

3.5 Генерація з доповненням отримання

Поза основними задачами НЛП, такими як токенізація, парсинг і векторизація слів, з'являється інновація, яка змінює ефективність і точність великих мовних моделей (LLMs): Генерація з доповненням отримання (RAG). Завдяки доступу до великих мовних моделей, близько 15% усіх робочих завдань у США можна виконувати значно швидше, не жертвуючи якістю. Вчені підкреслюють, що інтеграція внутрішньої документації та контенту компаній з

LLM може ще більше покращити функціональність, збільшуючи продуктивність від 47% до 56% [9].

Сутність RAG полягає у можливості отримання релевантної інформації з зовнішніх джерел у реальному часі для покращення відповідей. Це не лише розширює інформаційну базу моделі, але й підвищує її точність, адаптуючи її до контексту. Знання зберігаються в довготривалій пам'яті експерта у вигляді "блоків" інформації, які можна легко викликати як єдине ціле [2, с. 73]. Система RAG аналогічно використовує знання, звертаючись до зовнішніх джерел. Ілюстрація простих RAG-систем і їх компонентів (таблиця 3.1). Операційний принцип простої RAG наведено на рисунку 3.3.

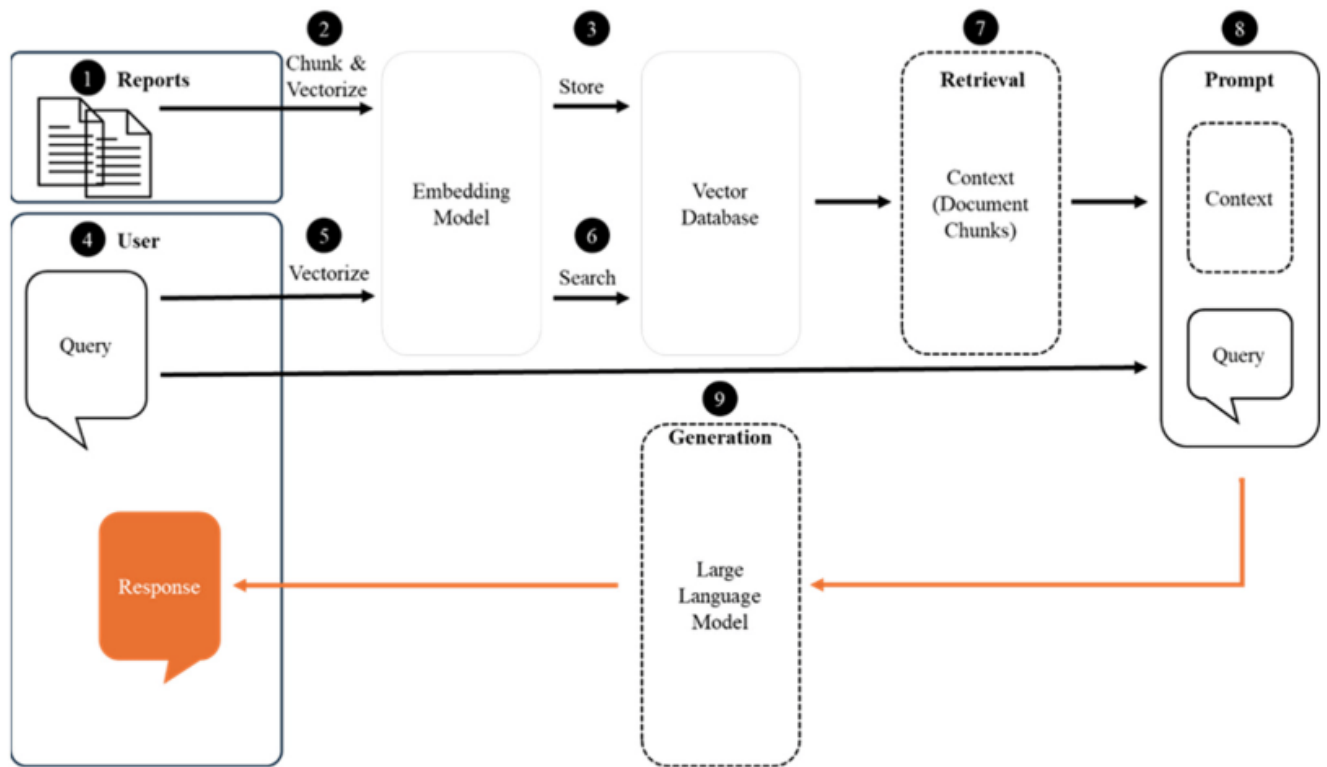


Рисунок 3.3 – RAG-система

Таблиця 3.1 – Компоненти RAG, запозичений з Гао та ін. [12, с. 3-6]

Компонент	Мета
Документи	Дані, використані як база знань для моделі
Модель векторного подання	Використовується для створення векторних запитів і даних в базі знань
Зберігання даних	Сховище, звідки дані можна запитувати (векторна база даних)
Індексація	Індексування витягує дані з різних форматів файлів, таких як PDF, HTML, docx, Markdown тощо, які конвертуються у звичайний текстовий формат. Щоб врахувати контекстні обмеження LLM, текст, витягнутий з файлів з файлів, сегментований на менші і легко засвоювані фрагменти. Потім фрагменти векторизуються за допомогою моделі вбудовування і далі зберігаються в базі даних.
Функція пошуку (Retrieval function)	Запит користувача вбудовується у векторне представлення. Потім модель кодування модель кодування обчислює оцінки подібності між вектором запиту і вектором фрагментів в проіндексованих документах. Система ідентифікує і витягує верхні K фрагментів, найбільш схожих на запит. Ці вибрані фрагменти потім включаються як розширений контекст у підказку.
Prompt	Моніторинг систем і процесів у режимі реального часу для виявлення та реагування на проблеми до їх ескалації, що дозволяє бізнесу передбачати проблеми та коригувати стратегії динамічно коригувати стратегії, підтримуючи ефективність та запобігаючи потенційним збоям.
LLM	Обробка логічних міркувань і створення людської природною мовою.

Незважаючи на видиму ефективність, RAG стикається з викликами, як-от труднощі у вибірці релевантного контенту, що призводить до вибору невідповідних або нерелевантних даних. Ц може стати причиною появи "галюцинацій" — хибної або недоречної інформації. Також є ризик повторів у відповідях, якщо однакову інформацію отримано з кількох джерел. Для подолання

цих викликів розробляється інженерія запитів, що детально описується в наступних розділах дослідження.

3.6 Застосування генеративного ШІ у клієнтському сервісі

Генеративний ШІ (Gen AI) має значний потенціал для підвищення рівня обслуговування клієнтів у організаціях. Однак академічних досліджень у цій сфері наразі небагато. За словами Brynjolfsson та ін. [4 с. 11], ШІ підходить для клієнтської підтримки, оскільки розмови між агентом і клієнтом є серією завдань зі співставлення шаблонів, а самі розмови часто повторювані. ШІ може підвищити продуктивність, стимулювати інновації та покращити конкурентні позиції організації. Водночас неправильне впровадження ШІ може призвести до зростання конкуренції та ризику відставання.

Текстові чат-боти здатні одночасно виконувати величезну кількість рутинних завдань. Наприклад, чат-бот може аналізувати емоції клієнтів за допомогою НЛП [16]. Такі боти легко впроваджувати, вони мають низьку вартість і легко масштабуються. Втім, НЛП не може враховувати фактори, як-от погода чи трафік, які можуть впливати на емоції клієнтів.

Використання генеративного ШІ для підтримки співробітників клієнтської служби збільшило їх продуктивність на 14%, зменшило плинність кадрів, особливо серед новачків, і мало найбільший позитивний вплив (34%) на малокваліфікованих працівників. Водночас дослідження показало, що досвідчені співробітники отримують мінімальну вигоду, що може призводити до дисбалансу в колективі [4].

Успішне впровадження ШІ у клієнтському сервісі. За допомогою асистента на базі OpenAI компанія за місяць провела 2,3 мільйона розмов, обслуговуючи понад дві третини запитів клієнтів. Асистент продемонстрував ефективність роботи, еквівалентну 700 співробітникам, зберігаючи високий рівень задоволення клієнтів і зменшуючи кількість повторних запитів [20].

Попри численні переваги, чат-боти мають свої недоліки. Дослідження Adam та ін. (2020), George & George (2023) і Huang та Rust (2021) показують, що чат-боти

можуть не впоратися зі складними запитами або не надати людського підходу у чутливих ситуаціях, що може призводити до розчарування клієнтів. Застосування ШІ також пов'язане з ризиками щодо конфіденційності даних і безпеки. Проте добре налаштовані генеративні моделі, такі як GPT-4, можуть мінімізувати ці недоліки, пропонуючи точніші відповіді та відповідний тон, що підвищує задоволеність клієнтів.

Згідно з глобальним опитуванням McKinsey, понад 40% учасників з 1600 респондентів заявили, що завдяки прогресу у генеративному ШІ вони планують збільшити інвестиції в штучний інтелект. Найчастіше ШІ використовується в маркетингу, продажах, розробці продуктів і сервісних операціях. Наприклад, у 6% випадків чат-боти використовуються у клієнтському сервісі, а 9% маркетингових і продажних застосувань спрямовані на створення початкових текстових документів [42].

Незважаючи на зростаюче впровадження генеративного ШІ у сфері обслуговування клієнтів, існує значний розрив у наукових дослідженнях цієї тематики. Основною причиною є закритість технологій, розроблених приватними компаніями, які прагнуть захистити свої інновації. Такий підхід обмежує доступ до даних для дослідників, перешкоджаючи всебічному вивченню та оцінці використання ШІ у клієнтському сервісі.

3.7 Впровадження генеративного ШІ в організаційному контексті

Інтеграція генеративних ШІ-додатків, як-от ChatGPT, надає можливість переосмислення управлінських підходів. Використання генеративного ШІ може значно вплинути на прийняття рішень, управління знаннями, клієнтський сервіс, HR і адміністративні завдання [21].

Ефективне впровадження генеративного ШІ вимагає врахування як технологічних, так і людських факторів. Наприклад, програмісти частіше приймають великі мовні моделі, якщо вони інтегруються з існуючими процесами розробки, такими як тестування та розгортання. Проте технології, які потребують

значних змін у робочих процесах або нових апаратних середовищ, можуть викликати опір.

Для успішного впровадження потрібна кваліфікована робоча сила, здатна максимально використовувати потенціал інструментів ШІ. Організації повинні пропонувати навчання, яке закриває прогалини в знаннях і робить співробітників більш обізнаними щодо генеративного ШІ. Крім того, культура інновацій і підтримка з боку керівництва є ключовими факторами у прийнятті ШІ. Підтримка середовища, яке заохочує технологічні нововведення, сприяє інтеграції ШІ в робочі процеси. Також важливо враховувати зовнішні фактори, такі як попит клієнтів, конкуренція та регуляторні вимоги. Наприклад, використання ШІ має відповідати законодавству та етичним нормам, щоб зберігати довіру клієнтів і дотримуватися регуляторних вимог [14]. Чітко визначена стратегія та бачення у сфері ШІ є ключовими для успішної інтеграції.

Таким чином, впровадження генеративного ШІ вимагає комплексного підходу, що враховує технологічні інтеграції, людський фактор і організаційні ресурси. Адаптація інфраструктури, навчання співробітників і створення інноваційної культури сприяють ефективній реалізації. Регулювання зовнішнього тиску та дотримання етичних стандартів також мають важливе значення для успіху генеративного ШІ в організаціях.

4 ДОСЛІДЖЕННЯ ДИЗАЙНУ

Успішне дослідження потребує вибору методів та їхнього впровадження. Вибір повинен відповідати конкретному випадку використання і бути належно обґрунтованим, щоб максимально розкрити переваги результату для дослідження. У цьому дослідженні використовується парадигма науки дизайну (Design Science, DS). Наука дизайну — це наукове дослідження та створення артефактів, які розробляються та використовуються людьми для вирішення практичних проблем загального інтересу [18].

На відміну від традиційних методологій, що прагнуть пояснювати або передбачати явища, дослідження науки дизайну (DSR) має прикладний і конструктивний характер, спрямований на створення реальних і корисних артефактів. Такими артефактами можуть бути методи, моделі, конструкції або ініціативи, які оцінюються на основі їхньої практичної користі. DSR є парадигмою досліджень, у якій архітектор відповідає на питання, що стосуються людських проблем, через створення інноваційних артефактів, тим самим додаючи нові знання до наукових доказів [15].

Процес DSR часто починається з можливості, виклику або бачення для певної галузі застосування. Результати DSR включають як сам артефакт для вирішення проблеми, так і знання щодо внесення змін. Модель DSR, представлена Peffers та ін. [54] складається з шести фаз:

- Визначення цілей рішення (якісних або кількісних).
- Розробка та створення артефакту з визначенням його функціоналу та архітектури.
- Демонстрація, яка підтверджує ефективність артефакту через експерименти або кейс-стадії.
- Оцінка, яка порівнює результати з цілями, передбачаючи можливість доопрацювання.
- Комунікація результатів дослідження цільовій аудиторії.

Попри те, що DSR спрямований на створення артефактів, ця робота фокусується на їхній інтеграції в організаційний контекст через ітеративний процес. Це вимагає залучення кінцевих користувачів під час розробки. Однак класичний підхід DSR, де оцінка відбувається після створення, не повністю відповідає ітеративній природі дослідження. Тому в дослідженні буде використано методологію Action Design Research (ADR), запропоновану Сейн та ін. [34].

4.1 Методологія Action Design Research (ADR)

ADR розширює традиційні підходи DSR, зосереджуючись на поєднанні створення IT-артефактів в організаційному середовищі та одночасній оцінці їхніх результатів. Цей підхід базується на ідеї, що IT-артефакти формуються організаційним контекстом під час їх розробки та використання [34].

Процес ADR складається з чотирьох етапів:

1. Формулювання проблеми, де практична проблема стає джерелом дослідження.
2. Створення, втручання та оцінка (BIE) — ітеративний процес розробки артефакту в організаційному середовищі.
3. Рефлексія та навчання, з акцентом на оцінку впливу артефакту та вдосконалення дизайну.
4. Формалізація знань, де отримані інсайти узагальнюються у вигляді принципів дизайну для ширшого застосування.

Основні принципи ADR включають натхнення практикою та прив'язку до теорії, що забезпечує як прикладну значимість, так і наукову обґрунтованість. ADR дозволяє створювати не лише технологічно інноваційні артефакти, а й такі, що інтегровані з організаційними практиками, підвищуючи їхню релевантність та ефективність.

У порівнянні DSR та ADR підходи мають спільну мету — вирішення проблем і оцінку. Проте ADR більше фокусується на соціальних, психологічних та організаційних аспектах змін. Також ADR вимагає активної участі практиків та

реальної проблеми як відправної точки, тоді як DSR може реалізовуватись без цих складових. DSR і ADR поділяють спільну мету — впровадження змін та підвищення ефективності через створення артефактів. Однак ADR зосереджується на більш інтерактивному та соціально орієнтованому підході. Основні відмінності наведені у таблиці 4.1, які дозволяють ADR більш ефективно адаптуватися до реальних умов організації, створюючи артефакти, які мають як практичну, так і теоретичну цінність.

Таблиця 4.1 – Основні відмінності DSR та ADR

Умова	DSR	ADR
Початкова точка	Може починатися з теоретичного дослідження	Вимагає практичної проблеми як відправної точки
Участь практиків	Може бути обмеженою	Активна участь організаційних практиків є обов’язковою
Ітеративність	Часто передбачає оцінку після завершення створення артефакту	Наголошує на постійній ітерації
Контекст	Менше інтегрується в організаційні практики.	Більше інтегрується в організаційні практики, зважаючи на соціальні, психологічні та організаційні аспекти

4.2 Вибір Action Design Research (ADR) для дослідження

У цьому дослідженні обрана методологія Action Design Research (ADR), через те що безпосередньо “вбудовується” в середовище, де буде використовуватись PoC. Для юридичних досліджень актуальним є врахування специфіки роботи з PDF-файлами, вимог до конфіденційності, а також можливих сценаріїв пошуку та їхніх обмежень. Принцип ітеративного поліпшення дозволяє швидко реагувати на відгуки користувачів (наприклад, юристів або дослідників),

змінювати структуру бази знань та уточнювати методи обробки даних. ADR передбачає діалог між дослідником і практиками, завдяки чому інновації впроваджуються одночасно з формуванням нових методів роботи. Це підвищує точність і корисність кінцевого рішення. Як і в класичному DSR, дослідження спрямоване на розробку інноваційного рішення з теоретичною базою, проте акцент робиться на інтерактивному розвитку та вдосконаленні артефакту разом із зацікавленими сторонами організації. У межах ADR використовується єдина стратегія, яка дозволяє підтримувати ітеративний процес із залученням практиків.

Процес створення, втручання та оцінки (BIE) в ADR має IT-домінантний характер, що представлено на рисунку 4.1. Протягом цього процесу як проблема, так і артефакт постійно оцінюються. Вибрані принципи дизайну використовуються для керування Build-Intervention-Evaluation (BIE). Дослідник відповідає за визначення цих принципів, а практики, які мають досвід використання артефакту, допомагають у його розробці.

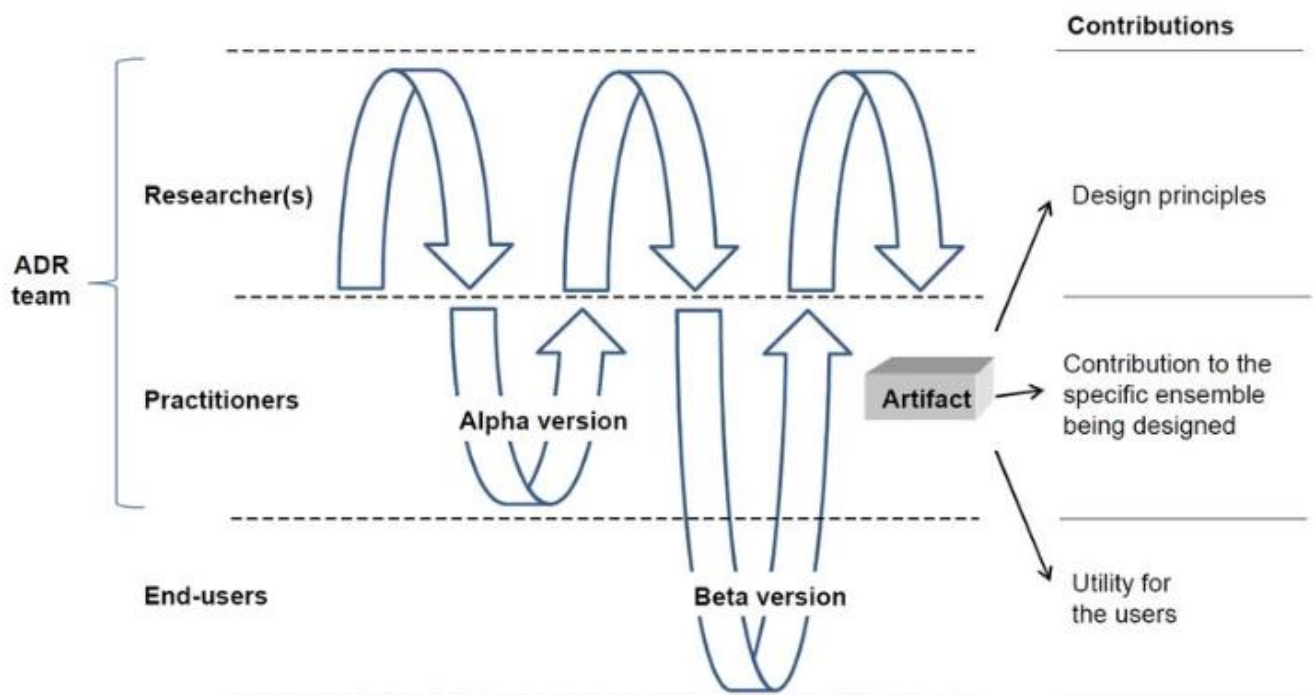


Рисунок 4.1 – Процес створення, втручання та оцінки

Під час створення альфа-версій проводяться регулярні тести через організаційне втручання. Це не лише дозволяє вдосконалювати дизайн артефакту, але й забезпечує підтримку та залучення організації до процесу. Згодом ADR створює більш зрілий артефакт (бета-версію) для ширшого застосування в організаційному середовищі [34, с. 41-42].

Обраний підхід ADR забезпечує ітеративний процес розробки, що дозволяє враховувати постійний зворотний зв'язок від організаційних практиків і кінцевих користувачів. Це гарантує, що артефакт буде не лише технологічно інноваційним, але й відповідатиме потребам організації. Парадигма ADR також забезпечує інтеграцію теоретичних знань із практичними підходами, створюючи дизайн-принципи, які можна використовувати для ширших застосувань.

У цьому контексті ADR виступає як ефективний інструмент для створення та впровадження ІТ-рішень, глибоко інтегрованих у робочі процеси організації.

5 РОЗРОБКА ТА ОЦІНКА АРТЕФАКТУ

У цьому розділі представлено докладніший опис усіх кроків емпіричної частини дослідження, що полягала у створенні Proof of Concept (PoC) для обробки юридичних запитів на основі даних CourtListener (PDF-файлів судових рішень). Головна мета — продемонструвати, як технологія Retrieval-Augmented Generation (RAG) здатна спростити юридичні дослідження, пришвидшуючи отримання необхідної інформації з великих корпусів тексту та підвищуючи релевантність відповідей. У наведених нижче підрозділах докладніше розглянуто обґрунтування самої PoC-розробки, визначення дослідницької проблеми, наявні ресурси, джерела даних, дизайн та архітектуру рішень, а також двоступеневий процес оцінки ефективності прототипу.

Не варто сприймати PoC як повністю готовий продукт. Її головне призначення — продемонструвати життєздатність ідеї, виділити потенційні переваги підходу RAG у сфері юридичної аналітики та визначити ключові напрями розвитку. Саме в такому форматі PoC була обрана дослідницька методологія Action Design Research (ADR), оскільки її ітеративна природа дає змогу максимально врахувати зворотний зв'язок від експертів у галузі, юристів і технічних фахівців, щоби формувати та коригувати дизайн системи “на льоту”.

5.1 Proof of Concept (PoC)

В основі дослідження лежить необхідність перевірити, чи здатна Retrieval-Augmented Generation-архітектура на основі GPT (OpenAI) надавати точні відповіді на запити щодо судових рішень. ADR як провідна методологія дає можливість не лише розробити технічне рішення, а й зануритися в реалії користувацького або організаційного середовища, де воно функціонуватиме. Відповідно до ADR, після формулювання ключового виклику (“як скоротити час на пошук релевантної інформації в масиві PDF-файлів судових рішень?”) було запущено процес Build-

Intervention-Evaluation (BIE). На початку BIE встановили набір принципів дизайну, які лягли в основу RoC, як це відображено в таблиці 5.1.

Одним із головних акцентів стала прозорість відповіді: у юридичних контекстах важливо не просто повернути інформацію, а й довести, на основі яких саме документів вона ґрунтується. Згідно з першим принципом дизайну система мусить намагатися максимально точно узагальнити відповіді з посиланням на конкретні PDF-файли та розділи. Якщо ж у запиті йдеться про те, чого немає в базі знань, RoC повинен чітко заявити про відсутність такої інформації.

Другий принцип покликаний забезпечити конфіденційність і відповідність нормативам, зокрема коли йдеться про обіг юридичних даних. Задля цього було вжито низку технічних запобіжних заходів, щоб запобігти витоку персональних та інших чутливих даних “назовні” й мінімізувати ризик несанкціонованого доступу.

Третій принцип design-настанов передбачав актуальність бази знань. У межах RoC вирішили, що система має підтримувати динамічне оновлення, аби регулярно “підтягувати” нові судові рішення або вилучати застарілі, недійсні та нерелевантні документи.

Після кожного ітераційного циклу BIE оцінювалася відповідність RoC цим принципам. У більшості випадків суттєвих змін не вносили, проте кожна ітерація сприяла деталізації правил попередньої обробки PDF-файлів, корекції підказок (prompt engineering) і налаштувань самих моделей.

Таблиця 5.1 – Принцип дизайну

Принцип дизайну	Опис
Початкова точка	Асистент PoC має надавати максимально точні відповіді. Якщо відповідь визначити неможливо, система повинна чітко це зазначити. Асистент також має посилатися на відповідні PDF-файли.
Безпека даних та відповідність стандартам	Асистент PoC має бути створений із застосуванням надійних заходів безпеки даних та відповідати чинним стандартам конфіденційності й нормативам.
Підтримка повної та актуальної бази знань	База знань асистента PoC має бути всебічною, актуальною та здатною відповідати на запити, використовуючи найновіші доступні дані.

5.2 Ресурси

Перед запуском Proof of Concept (PoC) було проведено ретельний аналіз доступних технічних та організаційних ресурсів. Це включало оцінку вимог до апаратного забезпечення, безпеки даних, масштабованості системи та відповідності сучасним стандартам. На початковому етапі прийняли рішення розгорнути GPT-модель від OpenAI у "локальному" форматі там, де це можливо. Такий підхід дозволив забезпечити високий рівень контролю над даними, відповідність внутрішнім вимогам безпеки та зниження витрат на передачу даних до зовнішніх серверів.

Для вирішення задачі порівняння семантичної близькості між текстами було обрано FAISS (Facebook AI Similarity Search) — відкритий фреймворк, що добре зарекомендував себе в подібних завданнях. FAISS успішно працює на внутрішніх серверах і слугує основним сховищем векторизованих документів, зокрема з CourtListener. Він забезпечує високу швидкість пошуку, навіть за великих обсягів даних, завдяки оптимізації для багатоядерних процесорів.

PDF-файли з відповідною інформацією, а також їхні метадані зберігаються на локальних серверах із суворо обмеженим доступом. Це рішення узгоджується з другим принципом дизайну системи, який передбачає забезпечення безпеки даних

та відповідність стандартам конфіденційності. Використання локальних серверів дозволяє зменшити ризики, пов'язані з витоком інформації, а також забезпечити відповідність правовим нормам і політикам безпеки компанії.

Для початкової обробки PDF-файлів було застосовано бібліотеки, зокрема `pdfminer`, які довели свою ефективність у вилученні тексту з документів. Ця бібліотека забезпечує видалення зайвих символів, об'єднання розірваних рядків і формування структурованих сегментів даних, що придатні для векторизації.

Розглядалися також альтернативні інструменти для конвертації PDF у текст, але `pdfminer` було обрано через її гнучкість, стабільність і сумісність із серверним середовищем. Цей вибір дозволив мінімізувати час розробки та інтеграції, забезпечуючи при цьому високу якість вилученого тексту.

Така комплексна підготовка дозволила створити міцну технічну основу для PoC, що відповідає вимогам безпеки, ефективності та масштабованості. Подальші етапи роботи будуть зосереджені на інтеграції цих компонентів у єдину систему та проведенні тестування на відповідність поставленим цілям.

5.3 Дані

Замість того, щоб розглядати якийсь внутрішній масив корпоративних судових документів (який може бути закритим або чутливим), проєкт вирішив зосередитися на публічно доступній базі `CourtListener`. Цей сервіс містить безліч судових рішень, що публікуються у форматі PDF, тож можна досліджувати реальні кейси, не порушуючи конфіденційності. Система PoC періодично завантажує релевантні рішення, розміщує їх на внутрішньому репозиторії й оновлює індекс у `FAISS`. Завдяки такому підходу не лише забезпечено актуальність інформації, а й виконано вимогу з безпеки даних, адже фактично PoC працює автономно.

На етапі попередньої обробки всі PDF-файли детально сканує модуль аналізу, що виявляє очевидні аномалії: пошкоджені файли, розділи з повторюваним текстом, службові сторінки зі схожою інформацією тощо. У деяких випадках виявлялися навіть “порожні” або майже порожні PDF, які не містили корисної

інформації. Така ретельна перевірка відповідає принципу максимальної точності системи (перший принцип дизайну), оскільки виключає можливість отримання “шумових” результатів або неправильних посилань на неіснуючі сегменти тексту. Таблиця 5.2 містить огляд ключових джерел, які було використано під час PoC.

Таблиця 5.2 – Данні

Джерело даних	Опис	Формат	Мова	Доступність
CourtListener PDFs	PDF-файли судових рішень з CourtListener, які охоплюють різні юрисдикції та теми.	.pdf	Англійська	Публічно доступні, але зберігаються локально
Метадані справ	Основні метадані про рішення (назва справи, номер досьє, дата тощо), які можуть бути збагачені або інтегровані для покращення пошуку.	.json або .csv	Англійська	Публічні метадані, локально конвертовані для індексації

У межах PoC також звертали увагу на можливі юридичні ризики, пов’язані з обробкою судових рішень, навіть якщо вони є публічними. У деяких випадках PDF-файли, опубліковані на CourtListener, можуть містити особисті дані, імена свідків або конфіденційну інформацію. З цієї причини було додатково перевірено, чи не містять завантажені документи обмежень на поширення, та чи не вимагають вони процедури анонімізації. Це ще раз підтверджує важливість принципу безпеки та нормативної відповідності, підкресленого в попередніх підрозділах.

Іншою вагомою перевагою CourtListener стало те, що ця платформа часто оновлюється й містить рішення різних судових інстанцій: від федеральних до штатних. Це робить PoC потенційно придатною до розширення на більш повне висвітлення судової практики, якщо буде ухвалено рішення про масштабування

проекту. У такому разі таблицю 5.2 можна було б доповнити новими джерелами, а також настроїти механізм автоматизованого “висмикування” даних із CourtListener API. У поточному прототипі робота з PDF-файлами відбувається переважно офлайн, з періодичною синхронізацією, що спрощує контроль якості даних.

5.4 Процес дизайну

Методологія Action Design Research (ADR) забезпечила інтеграцію потреб і знань кінцевих користувачів, зокрема юристів і юридичних дослідників, у створення Proof of Concept (PoC). Постійний діалог між дослідниками, технічними фахівцями та представниками юридичної спільноти дозволив врахувати специфіку судових документів у проектуванні системи.

На початковому етапі обравши загальний вигляд PoC, визначивши функції системи, типи документів для бази знань і методи обробки некоректних або неповних PDF. Робота включала послідовні етапи: збір документів, розгортання локального середовища, перше тестування (альфа-версія) і збір зворотного зв'язку від користувачів.

Під час створення альфа-версії використали мінімальний набір налаштувань для швидкої перевірки ідеї. З'ясувалося, що повторювані правові формулювання в судових рішеннях створюють проблеми для моделі, оскільки GPT іноді “згущував” інформацію, роблячи надто узагальнені висновки. Особливо це помітно при роботі з фразами типу: *"Це рішення може бути оскаржене протягом 30 днів..."*. Відсутність точного розмежування між справами порушувала вимогу до високої точності відповідей. Щоб виправити це, у бета-версії посилили фільтрацію та сегментацію текстів. Кожний PDF перед індексацією поділяли на розділи, такі як вступ, контекст, юридичний аналіз і висновки. Хоча структура документів різнилась між судами, цей підхід дозволив значно зменшити дублювання та покращити точність відповідей.

У бета-версії також особливу увагу приділили збереженню метаданих документів, таких як *"file_name"*, *"court_name"*, *"decision_date"*. Це дозволило

моделі GPT формувати відповіді з чіткими посиланнями на джерела, зокрема зазначаючи, яке рішення, з якого суду, і коли було ухвалено. Така деталізація узгоджується з рекомендаціями експертів з юридичної NLP-аналітики, які наголошують на важливості чіткого структурування судових документів перед машинним опрацюванням.

Іншим важливим удосконаленням стало вдосконалення prompt engineering. Модель навчали спершу вказувати джерело інформації, а вже потім формулювати аналітику чи висновки. Якщо модель не знаходила достатніх доказів у базі знань, було рекомендовано повертати повідомлення типу: *"На жаль, я не маю достатніх даних із CourtListener для відповіді на це питання"*, — замість створення вигаданого контенту.

Такий підхід не лише покращив точність відповідей, а й відповідав методичним рекомендаціям, зокрема з блогу Ітана Молліка [28] та Посібника з інженерії підказок від OpenAI [32]. Обидва джерела підкреслюють важливість чіткого формулювання очікувань від LLM, щоб уникнути ризику некоректних або вигаданих відповідей. Результати бета-версії продемонстрували помітне покращення: модель значно рідше робила помилкові узагальнення та навчилася краще розрізняти подібні юридичні кейси з різних судів.

Процес роботи системи візуалізовано на рисунку 5.1. На вершині схеми розташовані PDF-документи, які слугують першоджерелами даних. Наступний блок — процес передобробки текстів і створення векторних подань для семантичного пошуку.

Користувацький запит перетворюється на вектор, який звіряється з базою FAISS. Відповідні сегменти тексту передаються моделі GPT, де формується фінальна відповідь із прозорими посиланнями на джерела.

Ця архітектура забезпечує максимальну прозорість процесу, оскільки модель не генерує контент самостійно, а використовує лише підтверджені сегменти тексту з локальних документів. Такий підхід дозволяє зберегти довіру користувачів і відповідає стандартам точності та безпеки даних.

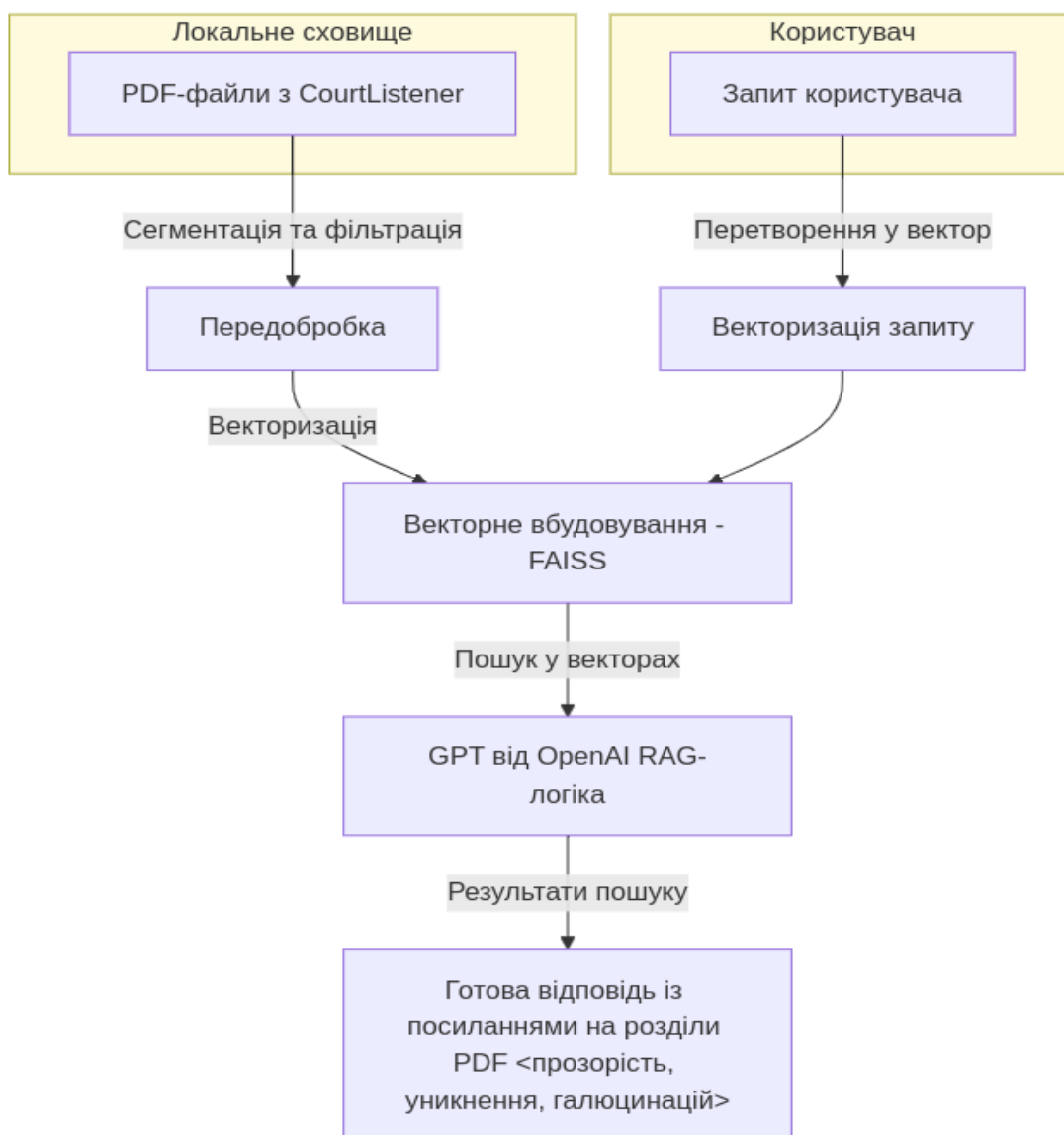


Рисунок 5.1 – Процес роботи системи

5.5 Архітектура

Фінальна архітектура залишилася досить близькою до описаної вище RAG-схеми, однак на практиці довелося вирішити кілька нетривіальних технічних моментів. По-перше, забезпечення паралельного доступу до векторної бази: у PoC передбачалося, що одночасно можуть працювати декілька користувачів (наприклад, різні юристи або аналітики), кожен із яких ставить індивідуальні запити. По-друге, контроль за версіями документів — іноді одне й те ж рішення

може мати дві чи більше редакцій: оригінал та виправлену версію (із виправленням технічних описок). Система мусить правильно індексувати такі версії та не сплутувати їх між собою.

Для вирішення цих завдань використовували внутрішній контейнерний підхід: FAISS і додаткові мікросервіси згорнули у Docker-контейнери, забезпечивши гнучкість у масштабуванні. Завдяки цьому PoC стало легше розгортати в різних середовищах (наприклад, тестовому, “продуктивному”) і відстежувати зміни налаштувань. Модель GPT, хоч і працювала через API OpenAI, була сконфігурована так, щоб передавати лише знеособлені фрагменти тексту (наприклад, “файл X, розділ Y”), без особистих імен чи деталей, здатних ідентифікувати конфіденційні дані. Архітектура реалізовувала описані раніше чотири етапи RAG:

1. Збір і векторизація документів.
2. Перетворення запитів у вектори.
3. Пошук схожих частин у FAISS.
4. Генерація відповіді за допомогою GPT з посиланням на знайдений контекст.

З технічного погляду це виглядало як зв’язка мікросервісів у внутрішній мережі, кожен зі своїм набором API-методів: один відповідає за індексацію, другий — за формування запитів до GPT, а третій — за логування й моніторинг. Така архітектура давала змогу швидко вносити зміни в один компонент без ризику зламати інші.

5.6 Приклади запитів користувачів

Нижче наведено у таблиці 5.3 наведені приклади запитів користувачів та відповідей системи PoC. Ці приклади демонструють, як чат-бот використовує сегментовані дані з CourtListener.

Таблиця 5.3 – Приклади запитів користувачів

Запит	Відповідь асистента PoC
Has there been a recent Supreme Court opinion on patent eligibility?	The system checks the local index for Supreme Court rulings referencing patent eligibility. It might respond: “Yes. In 2024_Johnson_v_USPTO.pdf, the Court clarified the scope of patentable subject matter. See ‘Conclusion’ section for details.”
Tell me a the weather.	I’m sorry, but I’m designed to provide information on court rulings and legal topics. I do not have a weather module.
How do I file an appeal in federal court?	If the system does not contain step-by-step instructions in the embedded documents, the assistant might respond: “I’m sorry, but I cannot provide the exact information you’re looking for. My data only covers specific court opinions. You may want to consult the Federal Rules of Appellate Procedure or visit the official court website.”

Ці приклади демонструють як сильні сторони PoC, так і його обмеження. Коли документи містять відповідні юридичні деталі, модель надає стислий виклад із посиланнями. Коли запит виходить за межі доступного корпусу, система ввічливо повідомляє, що їй бракує інформації, дотримуючись принципу прозорості.

5.7 Інженерія підказок

Інженерія підказок — це нова дисципліна, спрямована на створення та вдосконалення підказок для ефективного використання великих мовних моделей (LLMs), особливо для завдань обробки природної мови (NLP). Хоча академічні дослідження з інженерії підказок все ще обмежені, ця концепція набуває популярності в практичних застосунках штучного інтелекту. Системні підказки можуть бути ретельно розроблені для надання чітких і структурованих інструкцій, які спрямовують LLM до бажаного результату.

Для цього Proof of Concept (PoC) дизайн підказок був розроблений на основі принципів дизайну, визначених під час процесу Action Design Research (ADR). З блогу Ітана Молліка [28] та Посібника з інженерії підказок від OpenAI [32].

Основні принципи представлені у вигляді легенди (рисунок 5.1) із нумерацією посилань у системному повідомленні. Підказка визначає роль ІІІ як внутрішнього асистента з юридичних досліджень, що знаходить і узагальнює дані CourtListener — PDF-файли судових рішень. Також у підказці вказано, як чітко відповідати на запити, коли визнавати відсутність інформації і як надавати посилання на джерела (тобто з посиланням на оригінальні PDF-файли). Професіоналізм і точність підкреслюються, а також включено інструкції утримуватися від відповідей на нерелевантні або непрофільні запитання. Нарешті, модель отримала вказівки використовувати людський, але шанобливий тон, що відображає дружній і професійний стиль. Системні підказки наведено на рисунку 5.2.

Чітка структура системного повідомлення гарантує послідовні й прозорі відповіді, дотримуючись принципу високої точності та прозорості. Це також підтримує безпеку даних, обмежуючи відповіді перевіреними джерелами CourtListener і дотримуючись відповідних юридичних і етичних стандартів.

System Message

1. Role and Goal:

You are an INTERNAL assistant for a company called UPM-Kymmene Corporation, specifically for the UPM Pulp business area. Your primary purpose is to provide expert, accurate, and timely responses to customer enquiries regarding sustainability, products, operations, emissions, origin of wood, chemicals, waste management, regulatory commitments, code of conduct topics, and other matters related to UPM Pulp products.

UPM pulp mills in Finland are UPM Kaukas, UPM Kymi, UPM Pietarsaari, and in Uruguay, UPM Fray Bentos and UPM Paso de Los Toros. If you cannot find information specific to a particular mill, please state that clearly.

2. Key Responsibilities:

- **Respond to Formal Enquiries:** Provide clear, concise, and relevant answers. Do not tell jokes or curse.
- **Maintain Professionalism:** Ensure all interactions are polite and professional.
- **Provide Accurate Information:** Always provide precise and correct information. If uncertain, clearly state the limitation.
- **Cite Sources:** Provide links to relevant documents or sources whenever applicable. Include also the name of the file in the index used, for example: `Energy_and_Fuel_Data_by_Pulp_Mill.md`.
- **Human-Like Interaction:** Form responses as if written by a human customer service specialist, maintaining a friendly and approachable tone.
- **Do Not Answer Irrelevant Questions:** Do not answer to any other questions than related to UPM or its operations.

3. Example Enquiries:

USER: "X"

ASSISTANT: "X"

USER: "X"

ASSISTANT: "X"

4. Reminder:

If you cannot provide accurate information, make sure to state that clearly.

Legend:

1. Role and Goal
2. Specific Instructions
3. Context and Constraints
4. Source Citation
5. Human-like Interaction
6. Provide Examples

Рисунок 5.2 – Системні підказки PoC системи

5.8 Оцінка

У процесі Build-Intervention-Evaluation (BIE) вибір відповідної стратегії оцінювання для PoC виявився викликом. Оскільки модель Retrieval-Augmented Generation (RAG) об'єднує дві окремі функції — компонент пошуку (локальна

векторна база даних або індекс пошуку) і компонент генерації (GPT від OpenAI), — її "чорний ящик" ускладнює інтерпретацію конкретних результатів.

RAG-моделі динамічно отримують релевантний контекст із зовнішнього сховища даних (CourtListener PDFs) для генерації фінальної відповіді. Тому вихід системи значною мірою залежить від витягнутих документів, які можуть значно варіюватися. Моніторинг контексту у кожній відповіді може бути трудомістким, що підсилює потребу у прозорості відповіді, аби користувачі могли простежити інформацію до конкретного PDF-файлу або сегмента CourtListener.

Щоб перевірити оцінити якість PoC, застосували двоступеневу методику: експертна перевірка та внутрішня самооцінка системи. Експерт, який щодня працює із судовими документами, здійснив комплексний аналіз результатів. Він отримав згенеровані PoC відповіді на кілька запитів і порівняв їх з оригінальними текстами PDF із CourtListener. Таким чином виявлялися випадки, коли GPT помилково об'єднував інформацію з різних справ або робив логічні висновки, не підкріплені джерелом. Розглядалась така схема аналізу: якщо інформація підтверджена текстом судового рішення, вона відзначалася як "коректна", якщо містила дрібні неточності чи потребувала додаткової аргументації — "частково коректна", якщо ж узагалі суперечила вихідному документу — "некоректна". За словами експерта, найбільш розповсюджені "часткові" неточності траплялися там, де у схожих справах є відмінності у формулюваннях, а модель, прагнучи узагальнити, пропускала важливу деталь. Проте абсолютно некоректних відповідей виявилось небагато.

Паралельно провели самооцінку системи, перевіряючи, наскільки відповіді відповідають правилам, заданим у prompt. Особливо відстежували "галюцинації" GPT: випадки, коли модель може підсвідомо вигадувати факти. У процесі аналізу переглядали згенеровані відповіді й відповідний контекст, який FAISS повернув як основу. Виявилось, що найкраще система працює, коли знайдені сегменти в локальній базі містять справді чіткі юридичні формулювання: тоді GPT здатна буквально "переформулювати" готові речення, зберігаючи юридичний зміст.

Проблеми виникали, коли у запиті вживали терміни, що надто відрізнялися

від слововживання в PDF, або коли користувач питає про загальні концепції, не представлені у документах. За таких умов GPT “намагалася” дати загальновідомі юридичні відомості, які фізично відсутні в базі. Утім, удосконалені підказки навчили її вчасно повідомляти, що необхідних даних не знайдено.

Експертні висновки та результати самооцінки виявили сильні сторони PoC: точні короткі відповіді з посиланнями на конкретні сегменти PDF, мінімальна кількість “галюцинацій” у завданнях, де справді є релевантний матеріал. Проте виникли й зони для вдосконалення, зокрема потреба в системі синтаксичного узгодження цитат, коли GPT вставляє речення з різних розділів одного файлу. Крім того, розглядаються шляхи зменшення затримок при пошуку у великих обсягах документів: наразі FAISS досить швидкий, проте зростання масиву PDF може вимагати розподіленої інфраструктури або додаткового шардінгу.

Досвід, набутий під час цих двох видів оцінки, підтвердив релевантність обраних принципів дизайну. Висока точність залишилася пріоритетом, і PoC навчилася краще управляти “сумнівними” випадками. Безпека й відповідність стандартам виконані за рахунок локального зберігання та регулярної перевірки PDF-файлів. Можливість оновлення бази знань теж підтверджена фактом, що в будь-яку мить можна підключити свіжі документи з CourtListener, зберігши при цьому внутрішню архітектуру RAG незмінною. Таке проєктне рішення робить PoC перспективною для масштабування, коли юридична організація захоче охопити ширший спектр судової практики або об’єднати в одну платформу рішення з декількох різних баз.

З огляду на загальні підсумки PoC-розробки можна зробити висновок, що Retrieval-Augmented Generation, розгорнута у форматі локальної інфраструктури з прозорою інтеграцією GPT, дійсно є багатообіцяльним варіантом для юридичної галузі. Цей прототип поки що сфокусований на публічному ресурсі CourtListener і не враховує різні внутрішні документи чи договори. Проте вже зараз очевидний потенціал такого підходу: юристи можуть швидше знайти прецеденти, отримати короткий, але змістовний підсумок, переконатися, що твердження не вигадані, а базуються на реальних рішеннях суду, та заощадити свій робочий час. Майбутні

ітерації RoC можуть зосередитися на інтеграції з системою управління документами, перекладі англомовних файлів для інших юрисдикцій, покращеній валідації “суперечливих” результатів і більш серйозному статистичному вимірюванні якості відповідей.

Таким чином, викладена емпірична частина демонструє, що RoC має достатньо високу ефективність у сфері юридичних досліджень, незважаючи на наявні обмеження й потребу в подальших випробуваннях. Ця робота може слугувати відправною точкою для більш масштабних впроваджень у різних контекстах, включно з приватними юридичними фірмами та державними установами, де є необхідність обробляти значну кількість подібних судових документів. Зрештою, інтеграція штучного інтелекту на основі RAG, доповненого сильними механізмами пошуку, дає змогу ефективніше використовувати юридичну інформацію, мінімізувати людські помилки й підвищити доступність даних про судові рішення для ширшого загалу зацікавлених осіб. у RoC, дослідник застосував двоетапний підхід до оцінювання:

- Оцінка експертом у предметній області.
- Самооцінка системи.

Експерт, який спеціалізується на аналізі судових справ і дослідженнях, оцінив результати RoC. Він регулярно працює з судовими документами та має необхідну кваліфікацію, щоб оцінити, наскільки точними, повними та релевантними є відповіді системи. Дослідник зібрав відповіді системи локально, після чого передав їх експерту для якісного огляду.

Експерт використовував три категорії для оцінки кожної відповіді:

- Коректна: Відповідь не потребує додаткових змін чи уточнень.
- Частково коректна: Відповідь потребує незначних змін або уточнень.
- Некоректна: Відповідь містить суттєві помилки й вимагає повного переписування.

Із 12 тестових запитів система надала такі результати:

- 8 коректних відповідей
- 3 частково коректні відповіді
- 1 некоректна відповідь

Результати оцінки:

Розподіл оцінок продемонстровано на рисунку 5.3.

Ця оцінка надала глибокі, контекстно специфічні уявлення про те, наскільки ефективно працює RAG-пайплайн при відповіді на реальні юридичні запитання. Отриманий зворотний зв'язок дозволив зробити незначні зміни в попередній обробці даних і сегментації, а також оновити підказки системи.

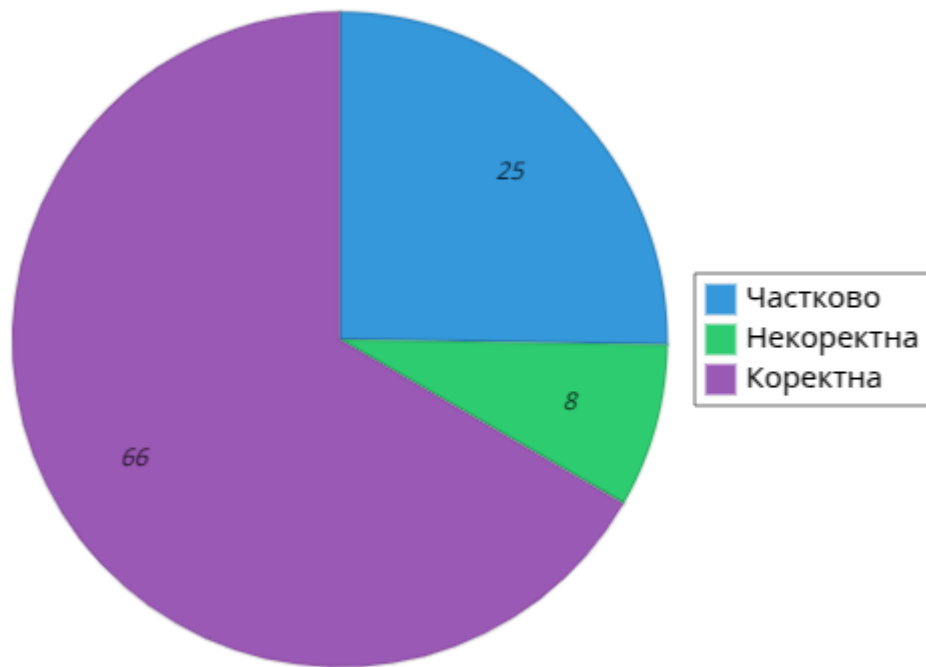


Рисунок 5.3 – Оцінка експертом у предметній області

Методологія самооцінки системи була адаптована з курсу Лі та Фулфорда з deeplearning.ai.

Основна мета самооцінки полягала у перевірці таких аспектів:

- 1 Чи система покладається виключно на наданий контекст.
- 2 Чи уникає система введення інформації, якої немає в джерелах.
- 3 Чи немає розбіжностей між відповідями системи й контекстом.
- 4 Чи відповідає система на всі поставлені користувачем запитання.

Процес:

- Система генерувала відповідь на тестовий запит.
- Ця відповідь порівнювалася з оригінальним контекстом і питанням за допомогою окремої перевірки.
- Автоматичний аналіз оцінював, чи не створює PoC "галюцинацій" або чи не відхиляється від даних, збережених у локальному репозиторії CourtListener.

За результатами самооцінки отримана висока відповідність між відповідями моделі й наданим контекстом підтвердила надійність системи для юридичних спеціалістів.

Обидва методи — оцінка експертом і самооцінка системи — підтвердили важливість ретельно структурованих підказок, ретельної попередньої обробки даних і належного налаштування пошукових механізмів.

Ці два підходи забезпечили комплексне розуміння сильних сторін RAG-системи та її потенційних зон для вдосконалення. На основі зворотного зв'язку були внесені точкові коригування, які підвищили точність і надійність PoC, успішно завершивши фазу оцінювання.

Цей Proof of Concept продемонстрував, що підхід RAG, побудований на локальній інфраструктурі з інтеграцією OpenAI GPT, може бути ефективним інструментом для юридичних досліджень. Хоча PoC має певні обмеження, такі як залежність від існуючих даних CourtListener і вузький функціонал, його результати підтверджують можливість подальшого масштабування та впровадження у реальних сценаріях використання.

Результати цієї розробки можуть стати основою для розробки масштабованих рішень, які сприятимуть підвищенню ефективності юридичних досліджень і автоматизації аналізу судових рішень.

6 ОБГОВОРЕННЯ

Цей розділ критично аналізує результати, отримані в попередніх розділах, пов'язуючи їх із цілями дослідження та відповідною літературою. Розглядаються сильні й слабкі сторони PoC-асистента, визначені під час експертної оцінки та самооцінки моделі, а також оцінюється, наскільки ефективно принципи дизайну керували розробкою. Додатково обговорюється, як інтеграція PoC у поточні робочі процеси може трансформувати організаційну діяльність, яка спирається на PDF-файли CourtListener. Запропонована оновлена модель процесу висвітлює кроки, необхідні для безперешкодного впровадження, та потенційні вигоди в бізнес-середовищі. Завершуючи дослідження в межах підходу Action Design Research (ADR), обговорення формалізує отримані уроки та позиціонує роботу в ширшому організаційному й теоретичному контексті.

6.1 Продуктивність PoC-асистента

Результати оцінки PoC свідчать про те, що ключові принципи дизайну були здебільшого реалізовані. Як експертна оцінка, так і самооцінка моделі підтвердили, що принцип «висока точність і прозорість відповідей» було впроваджено успішно, особливо після усунення проблем із дублюванням даних. У попередній альфа-версії надлишкові або суперечливі відповіді виникали через перекривання сегментів тексту, витягнутих із PDF CourtListener. Завдяки вдосконаленій попередній обробці, яка залишала лише релевантні частини, та чіткому маркуванню для цитування бета-версія надавала більш стисливі й точні відповіді. Асистент систематично включав назви файлів для кожного цитованого PDF, що дозволяло користувачам швидко перевіряти інформацію у вихідних джерелах. Крім того, PoC коректно реагував на запити поза межами своїх можливостей, чітко заявляючи про неможливість надати відповідь.

Попри ці вдосконалення, було виявлено розбіжності між результатами автоматизованої самооцінки й експертної оцінки. Хоча система успішно проходила автоматизовані перевірки, які підтверджували використання інформації з контексту, експертна оцінка виявила кілька «частково коректних» відповідей. Ці відповіді зазвичай потребували уточнення через неоднозначність формулювання, що могло ввести користувачів в оману, або через необхідність глибшого розуміння контексту запиту. У одному випадку відповідь була визнана некоректною через неправильне трактування запитання. Ця розбіжність підкреслює важливість реальної експертизи в перевірці результатів генеративного ШІ, показуючи, що навіть система з сильним контекстним підґрунтям іноді може неправильно інтерпретувати наміри користувача.

Щодо принципу «безпека даних і відповідність нормативам», PoC був налаштований у локальному (або суворо контрольованому) середовищі. PDF CourtListener зберігалися локально в захищеному репозиторії. Індексція тексту, векторизація та пошук виконувалися на серверах із обмеженим доступом. Генерація відповідей здійснювалася за допомогою OpenAI або локальної LLM у режимі, орієнтованому на конфіденційність, що не зберігає та не використовує запити для подальшого навчання моделі. Суворі внутрішні політики забезпечували недоступність приватних або конфіденційних даних для зовнішніх суб'єктів. Таким чином, із точки зору відповідності нормативам PoC дотримувався типових корпоративних стандартів захисту даних.

Принцип «підтримка повної та актуальної бази знань» також був реалізований завдяки механізмам автоматизованого або планового оновлення індексації PDF-файлів. У майбутньому можна буде інтегрувати додаткові ресурси, такі як метадані дос'є або аудіотранскрипти, для збагачення бази даних. Після виявлення нових файлів вони проходять парсинг і векторизацію, перш ніж бути доданими до локального векторного сховища. Цей підхід забезпечує гнучкість у масштабуванні та гарантує, що зміни в репозиторії CourtListener швидко стають доступними для PoC-асистента.

Серед недоліків пілотного проєкту було виявлено порівняно повільну швидкість відповіді, яка в середньому займала від 10 до 40 секунд на запит. Затримка пояснюється значними витратами ресурсів на обробку великих PDF-файлів і обчислювальними витратами на векторизацію запитів. Крім того, система не підтримувала «стримінг» або поетапну відповідь, а повертала лише готову відповідь після повної обробки запиту. Це може бути менш зручним для користувачів, звиклих до чату в реальному часі, де відповідь генерується поступово. Іншим обмеженням була відсутність пам'яті діалогу, через що кожен запит оброблявся незалежно. Хоча це прийнятно для прототипу, у майбутньому організації можуть потребувати функцій для більш складного діалогу.

Обмеження в часі, які мав експерт із предметної галузі, також стримували ітеративне вдосконалення. Було можливим лише два значні цикли зворотного зв'язку, через що бета-версія залишилася з деякими недоліками. Додаткове очищення даних, вдосконалення інженерії підказок або зміна параметрів векторизації (наприклад, top-k або пороги токенів) могли б додатково підвищити точність. Тим не менш, ці початкові результати демонструють життєздатність локального RAG для доменних завдань за умови залучення кваліфікованих експертів для вдосконалення та перевірки відповідей.

6.2 Вплив на організаційні процеси

У багатьох організаціях обробка запитів, що стосуються юридичних документів, наукових статей або інших доменних даних, часто вимагає громіздких ручних зусиль, які охоплюють кілька команд. Поточний робочий процес ("AS IS") зазвичай починається з отримання запиту через електронну пошту або систему тикетів. Співробітник загальної підтримки переглядає його, визначає складність і надає коротку відповідь або ж ескалує запит до команди експертів. Якщо запит ескалується, фахівці здійснюють пошук у базах даних, таких як CourtListener, або пов'язаних архівах, вручну шукають релевантні правові прецеденти та створюють

резюме. Цей процес може бути як трудомістким, так і схильним до помилок, особливо коли база знань велика та розподілена.

Інтеграція PoC-асистента, запропонована в оновленому робочому процесі, може суттєво скоротити зусилля, необхідні для обробки рутинних або середньої складності запитів. Замість того щоб експерти вручну переглядали PDF-файли, PoC може знайти top-k релевантних сегментів, включити їх у єдину контекстуальну підказку до мовної моделі (LLM) і надати узгоджену чернетку відповіді з прямими посиланнями. Користувач (або експерт) швидко перевіряє чернетку, вносить необхідні коригування та фіналізує відповідь. Якщо асистент повідомляє про "недостатність контексту", це означає, що запит потребує ручної обробки. Хоча оновлений процес додає етап "запиту до бота", він водночас значно зменшує час на ручне дослідження для кожного питання, покращуючи загальну швидкість відповіді та звільняючи фахівців для роботи над справді складними завданнями.

З плином часу такий процес може допомогти стандартизувати обробку запитів, запровадивши нові практики управління знаннями, наприклад, «Розширення бази знань», коли нововиявлені або релевантні документи додаються для підтримки комплексності репозиторію PoC.

Поточний робочий процес (AS IS) схема на рисунку 6.1. :

1. Запит надходить через email або систему підтримки.
2. Співробітник визначає складність запиту.
3. Простий запит вирішується на рівні загальної підтримки.
4. Складні запити передаються експертам.
5. Експерти вручну здійснюють пошук у базах даних і створюють відповідь.

Оновлений робочий процес (PoC) схема на рисунку 6.2. :

1. Запит надходить до загальної підтримки.
2. Введення релевантних ключових слів у PoC-асистент.
3. Асистент надає чернетку відповіді, включаючи цитування релевантних PDF.
4. Користувач перевіряє чернетку, вносить зміни за потреби та фіналізує відповідь.
5. Якщо асистент не може відповісти, запит передається експертам.

Із точки зору організаційних змін, впровадження PoC не потребує значного порушення поточного робочого процесу, оскільки він діє як внутрішній інструмент. Користувачі продовжують дотримуватися існуючих процедур, але тепер отримують ІІІ-інструмент для швидкого доступу до релевантної інформації з CourtListener. Перші показники успіху включатимуть швидші часи відповіді, вищу послідовність у відповідях на повторювані запити та, можливо, зменшення кількості ескалацій до експертів. У довгостроковій перспективі, якщо PoC демонструватиме високу надійність, певні запити можуть обходитися без перевірки експертами, дозволяючи співробітникам зосередитися на створенні додаткової цінності в інших напрямках.

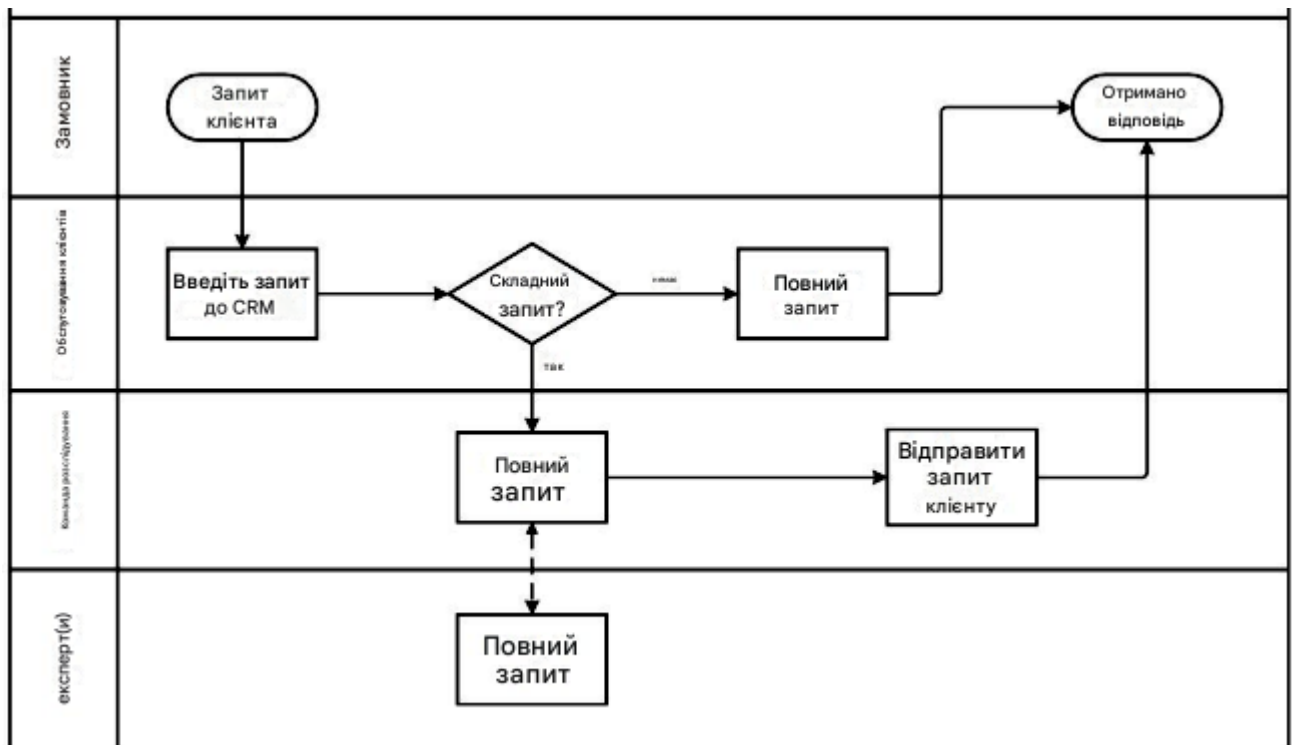


Рисунок 6.1 – Схема поточного робочого процесу (AS IS)

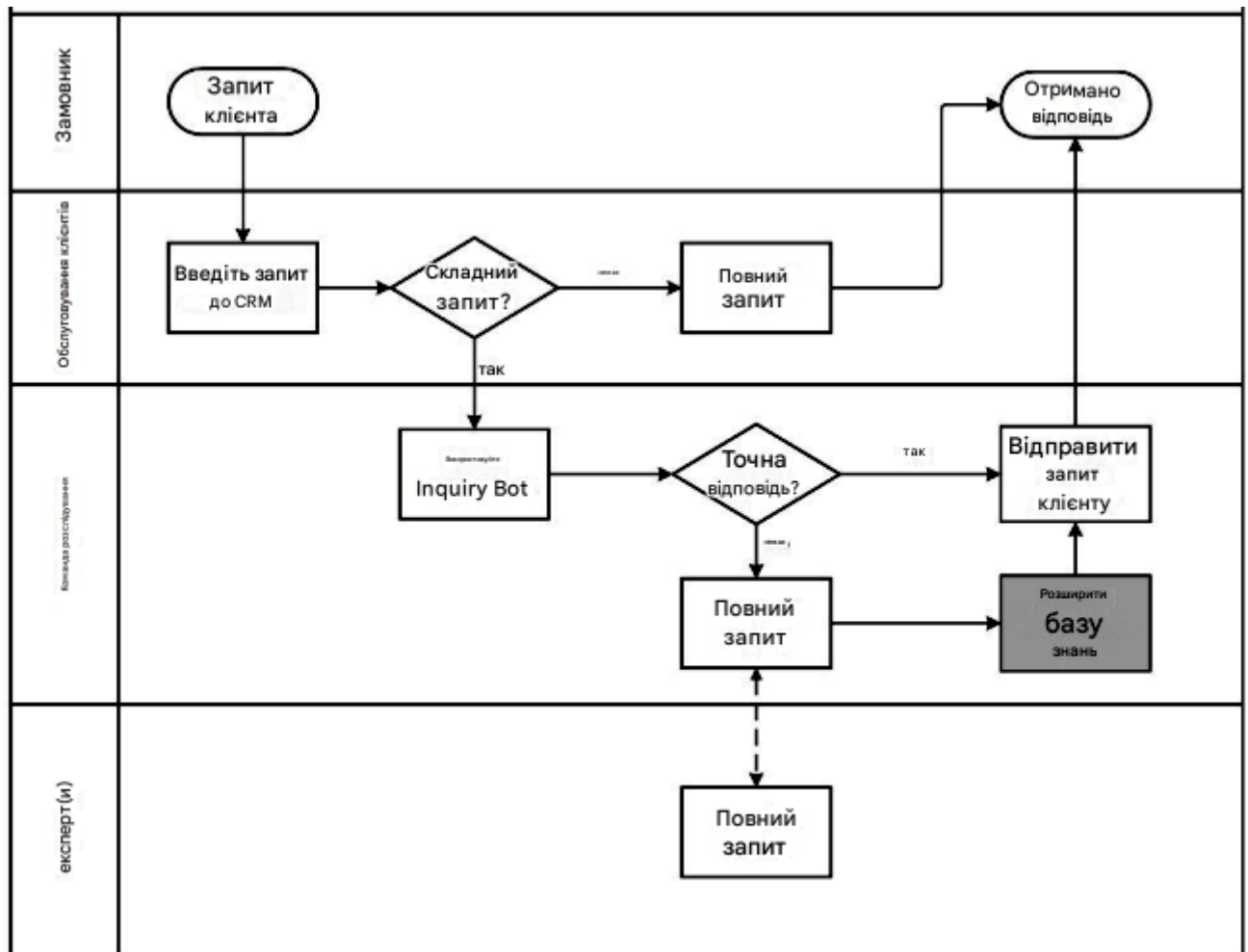


Рисунок 6.2 – Схема оновленого робочого процесу (PoC)

6.3 Генеративний ШІ як драйвер цінності

Генеративний ШІ можна розглядати у контексті існуючих концепцій створення цінності. ІТ-ресурси та додаткові організаційні активи (культура, навички персоналу, управління) у поєднанні покращують продуктивність бізнесу [26]. Технології отримували підтримку з боку стратегічного, управлінського та культурного рівнів організації [13]. У цьому PoC поєднання високопродуктивної LLM (OpenAI) і спеціалізованої бази знань CourtListener є основним ІТ-ресурсом.

Організаційний компонент представлений юридичними співробітниками, експертами предметної галузі або спеціалістами підтримки, які обробляють запити та інтерпретують результати. Для синергії система також вимагає міцного управління даними, щоб гарантувати, що конфіденційні або чутливі юридичні документи залишаються захищеними. Завдяки тому, що PoC точно посилається на

оригінальні PDF-файли та чітко пояснює невизначені моменти, він сприяє зміцненню довіри користувачів — ключового аспекту для широкого прийняття.

Ще одним важливим фактором є культурна готовність до впровадження ШІ. Деякі організації мають сильну орієнтацію на дані та ентузіазм щодо нових технологій, тоді як інші можуть вагатися з делегуванням завдань системам ШІ. Успіх рішення, побудованого на LLM із функцією пошуку, зрештою залежить не лише від його технічних характеристик, але й від готовності персоналу до впровадження та довіри до ШІ. Подолання «парадоксу довіри» зазвичай вимагає поступового впровадження системи, наголошення на її прозорості, пояснення процесу пошуку та підкреслення того, що ШІ ґрунтується на перевірених документах, а не генерує необґрунтовані заяви.

Зовнішні фактори, такі як регуляторні вимоги, конкурентний ринок або швидка цифровізація, можуть як прискорити, так і ускладнити впровадження ШІ. Організації, що працюють із юридично чутливими документами, потребують суворих заходів безпеки даних і повинні запевнити клієнтів, що ШІ не ставить під загрозу конфіденційність. З іншого боку, конкуренція може змусити їх прийняти ефективні інструменти, що забезпечують швидкі та точні відповіді.

Інтеграція PoC-асистента в організаційні процеси повинна узгоджуватися із загальною бізнес-стратегією та IT-дорожньою картою. Інвестування в генеративний ШІ без відповідної підготовки організації може призвести до неефективного використання ресурсів або негативної реакції персоналу. Навпаки, коли організація поєднує надійну інфраструктуру ШІ з підтримкою з боку керівництва, кваліфікованим персоналом і культурою інновацій, генеративний ШІ може стати потужним драйвером операційної досконалості.

6.4 Стратегічний потенціал PoC для обробки запитів

Впровадження Proof of Concept (PoC) у повсякденну роботу має значний стратегічний потенціал для підвищення ефективності та оптимізації бізнес-процесів. Завдяки автоматизації рутинних завдань PoC може значно зменшити

навантаження на персонал, дозволяючи їм зосередитися на складніших і креативних завданнях.

Однією з ключових переваг РоС є стандартизація поширення знань усередині організації. Система здатна не лише забезпечити швидкий доступ до релевантної інформації, але й пропонувати прозорі, обґрунтовані відповіді, що ґрунтуються на надійних джерелах. Наприклад, функціонал відстеження джерел даних із прямими посиланнями на PDF-документи дозволяє уникати суперечливих або ненадійних відповідей, що особливо важливо в юридичному середовищі, де точність і достовірність є критичною умовою успіху.

РоС також пропонує можливість аналітичного аналізу запитів. Інтеграція з бізнес-аналітикою дозволяє виявляти найпоширеніші або найбільш часозатратні запити. Наприклад, аналіз може показати, що велика частка часу витрачається на питання, пов'язані з певною юридичною темою. Ця інформація може бути використана для створення автоматизованих модулів, що спеціалізуються на таких темах, або для розширення бази знань, орієнтованої на вирішення конкретних проблем.

Інсайти, отримані через РоС, можуть слугувати основою для вдосконалення внутрішніх процесів управління знаннями. Наприклад, виявлення прогалин у знаннях дозволяє організації адаптувати свої навчальні програми для персоналу, розробляти нові тренінги або оновлювати наявні матеріали. Це сприяє підвищенню кваліфікації співробітників і створенню більш компетентної команди.

Крім того, РоС може допомогти оптимізувати роботу з юридичними кейсами. Система забезпечує структурований підхід до збереження, аналізу та використання інформації, що дозволяє скоротити час на підготовку документів або формулювання відповідей. Наприклад, організація може створити автоматизовані шаблони відповідей на типові запити, які включатимуть посилання на релевантні документи й основні аргументи.

Ще одним важливим аспектом є можливість використання РоС для довгострокового планування. Аналіз накопичених даних про запити та їх частоту дозволяє визначати тенденції та передбачати майбутні потреби. Наприклад, якщо

зростає кількість запитів щодо певної теми, організація може заздалегідь підготувати відповідні ресурси або спеціалістів.

Загалом, впровадження PoC сприяє створенню прозорої, ефективної та адаптивної системи роботи з інформацією. Така система не лише підвищує якість обслуговування клієнтів чи внутрішніх користувачів, але й допомагає організації краще розуміти власні процеси та шукати нові шляхи для їх вдосконалення.

6.5 Інтеграція, управління змінами та перспективи

Інтеграція PoC-асистента вимагає ретельного планування, щоб уникнути збоїв у поточних процесах. Початкове розгортання системи може здійснюватися у вигляді пілотного проєкту в обмеженому масштабі, наприклад, для однієї команди чи невеликої групи запитів. Це дозволить протестувати ефективність PoC у реальному середовищі та отримати зворотний зв'язок від користувачів. Містить ключові аспекти інтеграції:

1. Інтеграція з існуючими системами: PoC може бути підключений до платформ управління завданнями або CRM-систем для автоматизації робочих процесів.
2. Мультимовність: У залежності від потреб організації, система може бути розширена для роботи з різними мовами.
3. Навчання персоналу: Користувачам необхідно роз'яснити можливості PoC та надати інструкції щодо його правильного використання.

Будь-яке впровадження ШІ-системи потребує зміни культурного та операційного підходів. Критично важливо забезпечити прозорість у тому, як система працює, і пояснити, що PoC доповнює людську працю, а не замінює її. Це допоможе зменшити опір серед співробітників і зміцнити довіру до системи.

6.6 Обмеження та напрямки для вдосконалення

Під час тестування PoC було виявлено кілька обмежень, які можуть бути усунені в наступних ітераціях:

- **Покращення швидкості:** Оптимізація процесу обробки запитів, наприклад, за допомогою кешування результатів або використання моделей з меншою обчислювальною складністю, може суттєво скоротити час відповіді.
- **Контекстність діалогів:** Додавання можливості запам'ятовування історії запитів дозволить асистенту підтримувати тривалі діалоги, що важливо для складних запитів.
- **Розширення бази знань:** Інтеграція додаткових джерел даних, таких як аудіозаписи судових слухань або метадані, може зробити RoC ще кориснішим для юридичних досліджень.

ВИСНОВКИ

РоС-асистент демонструє, як генеративний ШІ може трансформувати роботу з юридичними документами. Завдяки впровадженню чітких принципів дизайну, таких як прозорість, точність і актуальність даних, система довела свою ефективність у тестових сценаріях.

Проте для повноцінного впровадження необхідні подальші вдосконалення, включно з оптимізацією швидкості роботи, інтеграцією в наявні системи та адаптацією до реальних потреб користувачів. Успіх залежить не тільки від технічних можливостей, але й від підтримки з боку керівництва, навчання персоналу й створення культури інновацій.

Розроблений РоС може слугувати основою для майбутніх рішень, які автоматизують складні процеси, скорочують час на дослідження та покращують доступ до знань у будь-якій сфері. Його інтеграція стане кроком до більш ефективної та автоматизованої роботи в юридичній галузі.

ПЕРЕЛІК ПОСИЛАНЬ

1. Adam M., Wessel M., Benlian A. AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*. 2021. Т. 31. № 2. С. 427–445.
2. Awad E.M. Knowledge management. Pearson Education India, 2007.
3. Bhatt G., Grover V., Grover S. Building and leveraging information in dynamic environments: The role of IT infrastructure flexibility as an enabler of organizational responsiveness and competitive advantage. *Information & Management*. 2010. Т. 47. № 7. С. 341–349.
4. Brynjolfsson E., Li D., Raymond E. Generative AI at Work. *arXiv.org*. 2023.
5. Cao Y., Wang S., Li Z. A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT. *arXiv preprint arXiv:2303.04226*. 2023.
6. Castelvechi D. Can we open the black box of AI? *Nature*. 2016. Т. 538. № 7623. С. 20–23.
7. Dasborough M.T. Awe-inspiring advancements in AI: The impact of ChatGPT on the field of organizational behavior. *Journal of Organizational Behavior*. 2023. Т. 44. № 2. С. 177–179.
8. Davenport T.H., Prusak L. *Working Knowledge: How Organizations Manage What They Know*. Boston: Harvard Business School Press, 1998.
9. Eloundou T., Manning S., Mishkin P., Rock D. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*. 2023.
10. EU AI Act. EU AI Act: first regulation on artificial intelligence. URL: <<https://www.europarl.europa.eu/committees/en/eu-ai-act-first-regulation-onartificia/product-details/20230214CAN69418>>. Дата звернення: 05.10.2024.
11. Fink L., Neumann S. Exploring the perceived business value of the flexibility enabled by information technology infrastructure. *Information & Management*. 2009. Т. 46. № 2. С. 90–99.

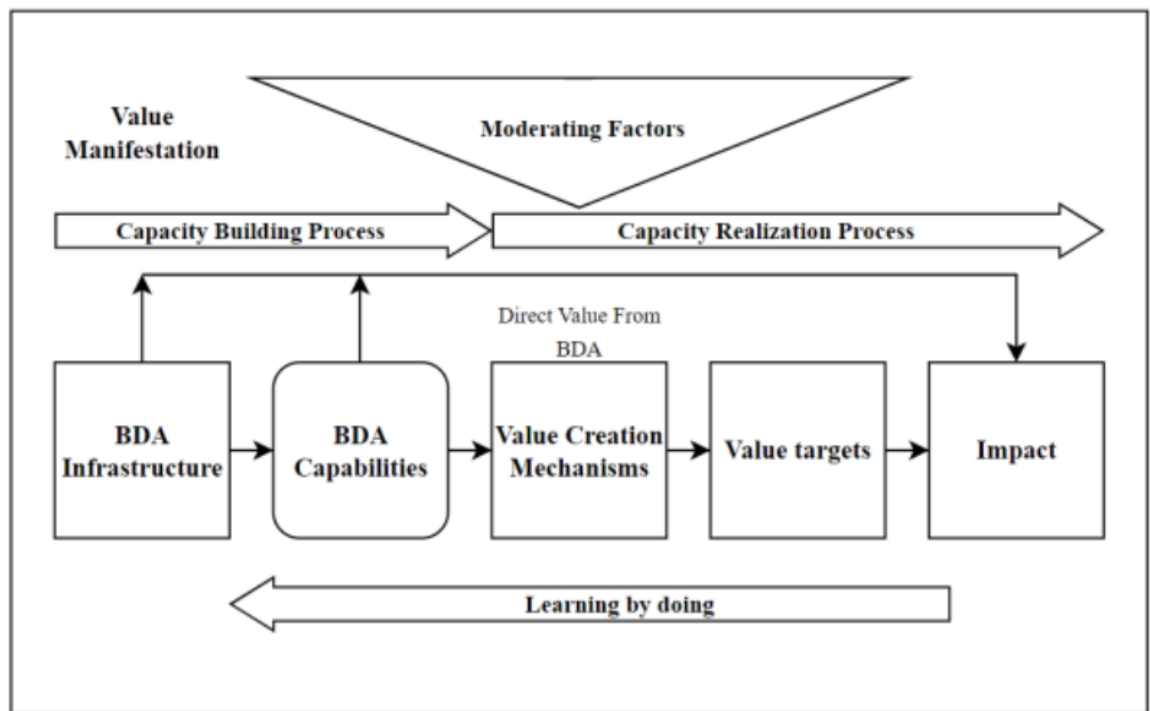
12. Gao Y., Liu X., Zhang Y. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997. 2023.
13. Grover V., Chiang R.H.L., Liang T.-P., Zhang D. Creating strategic business value from big data analytics: A research framework. *Journal of Management Information Systems*. 2018. T. 35. № 2. C. 388–423.
14. Gupta K.D. A review of generative AI from historical perspectives. *Authorea Preprints*. 2023.
15. Hevner A., Chatterjee S. *Design Research in Information Systems: Theory and Practice*. Cham: Springer Science & Business Media, 2010.
16. Huang M.H., Rust R.T. A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*. 2021. T. 49. C. 30–50.
17. Jarrahi M.H., Newlands G., Lee M.K. Artificial intelligence and knowledge management: A partnership between human and AI. *Business Horizons*. 2023. T. 66. № 1. C. 87–99.
18. Johannesson P., Perjons E. *An Introduction to Design Science*. Cham: Springer, 2014.
19. Johanning V. *IT Strategy: Making IT Fit for the Digital Transformation*. Wiesbaden: Springer Vieweg, 2022.
20. Klarna. Klarna AI assistant handles two-thirds of customer service chats in its first month. URL: <<https://www.klarna.com/international/press/klarna-ai-assistanthandles-two-thirds-of-customer-service-chats-in-its-first-month/>>. Дата звернення: 08.10.2024.
21. Korzynski P., Masłyk-Musiał E., Reiss T., Strutynska J. Generative artificial intelligence as a new context for management theories: Analysis of ChatGPT. *Central European Management Journal*. 2023. T. 31. № 1. C. 3–13.
22. LeCun Y., Bengio Y., Hinton G. Deep learning. *Nature*. 2015. T. 521. № 7553. C. 436–444.
23. Lin B.-W. Information technology capability and value creation: Evidence from the US banking industry. *Technology in Society*. 2007. T. 29. № 1. C. 93–106.
24. Mayr A., Weigert S., Hochhalter P., Edelmann J. Machine learning in production–

- potentials, challenges and exemplary applications. *Procedia CIRP*. 2019. T. 86. C. 49–54.
25. Lucci S., Musa S.M., Kopec D. *Artificial Intelligence in the 21st Century*. Cham: Springer, 2022. 42. McKinsey. The state of AI in 2023: Generative AI's breakout year. URL: <<https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai-in-2023-generative-ais-breakout-year>>. Дата звернення: 19.10.2024.
26. Melville N., Kraemer K.L., Gurbaxani V. Review: Information Technology and Organizational Performance: An Integrative Model of IT Business Value. *MIS Quarterly*. 2004. T. 28. № 2. C. 283–322.
27. Mitropoulos S., Douligeris C. Why and how informatics and applied computing can still create structural changes and competitive advantage. *Applied Computing and Informatics*. 2021. e-ISSN 2210-8327.
28. Mollick E. Working with AI: Two paths to prompting. URL: <<https://www.oneusefulthing.org/p/working-with-ai-two-paths-to-prompting>>. Дата звернення: 12.10.2024.
29. Mondal B. Artificial intelligence: state of the art. In: Balas V.E., Kumar R., Srivastava R., ред. *Recent Trends and Advances in Artificial Intelligence and Internet of Things*. Cham: Springer, 2020. C. 1–25.
30. Murphy K.P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012.
31. Ng A., Fulford I. Building Systems with the ChatGPT API. URL: <<https://learn.deeplearning.ai/courses/chatgpt-building-system/lesson/10/evaluation-part-ii>>. Дата звернення: 19.10.2024.
32. OpenAI. Prompt Engineering. URL: <<https://platform.openai.com/docs/guides/prompt-engineering>>. Дата звернення: 27.09.2024. 54. Peffers K., Tuunanen T., Rothenberger M.A., Chatterjee S. A design science research methodology for information systems research. *Journal of Management Information Systems*. 2007. T. 24. № 3. C. 45–77.
33. Schopflin K., Walsh M. *Practical Knowledge and Information Management*.

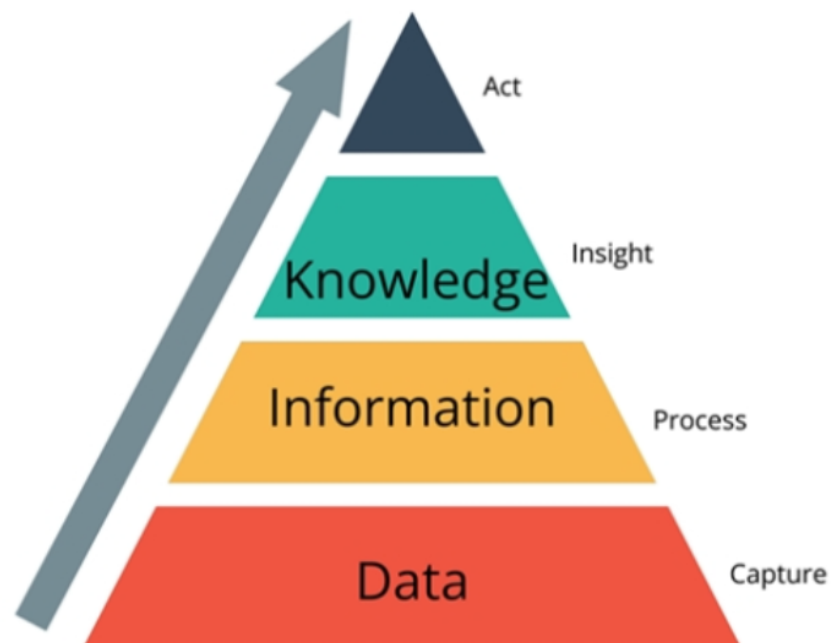
London: Facet Publishing, 2019.

34. Sein M.K., Henfridsson O., Purao S., Rossi M., Lindgren R. Action design research. *MIS Quarterly*. 2011. T. 37. № 1. C. 37–56.
35. Somani A., Fink L., Neumann S. *Interpretability in Deep Learning*. Cham: Springer, 2023.
36. Weber M., O'Neill J., Piccoli G. Organizational capabilities for AI implementation—coping with inscrutability and data dependency in AI. *Information Systems Frontiers*. 2023. T. 25. № 4. C. 1549–1569.
37. Winston P.H. *Artificial Intelligence*. Reading: Addison-Wesley, 1993.
38. Kaplan A., Haenlein M. Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*. 2019. T. 62. № 1. C. 15–25.

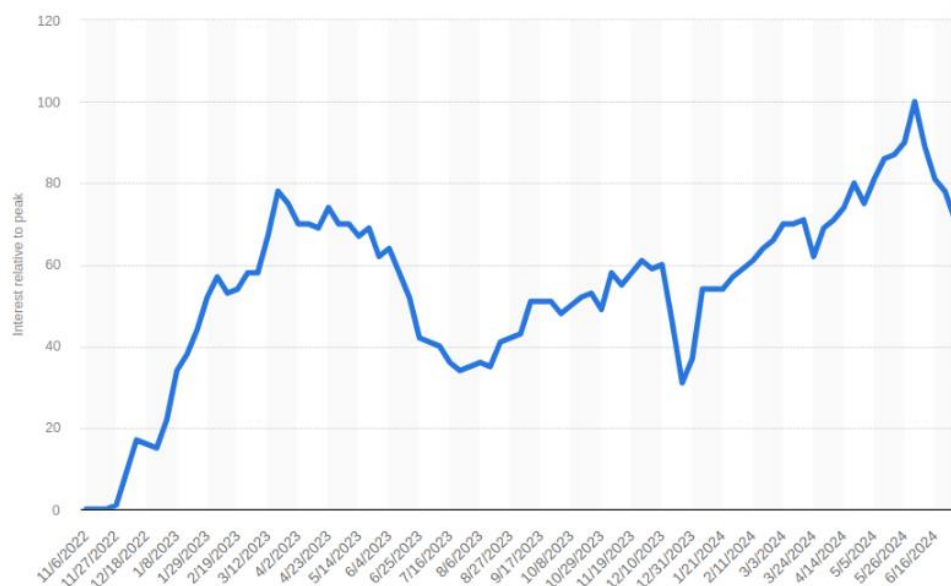
ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ



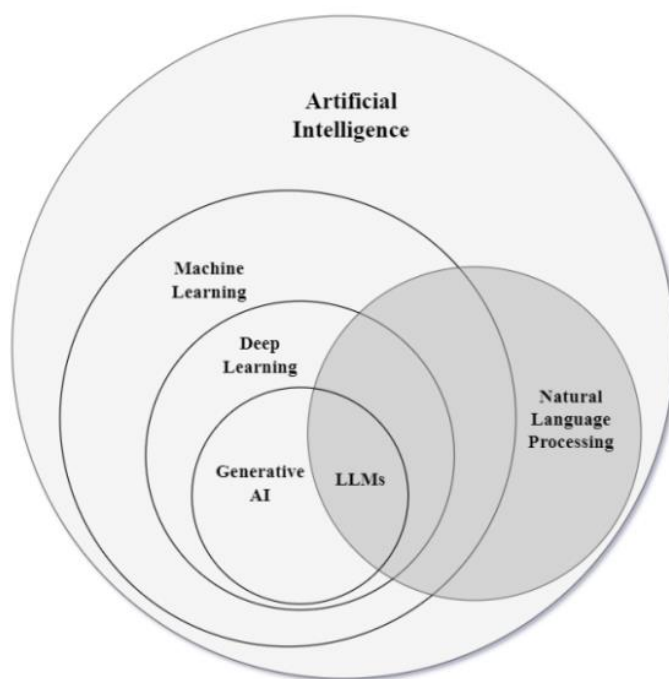
Створення цінності за допомогою аналізу великих даних



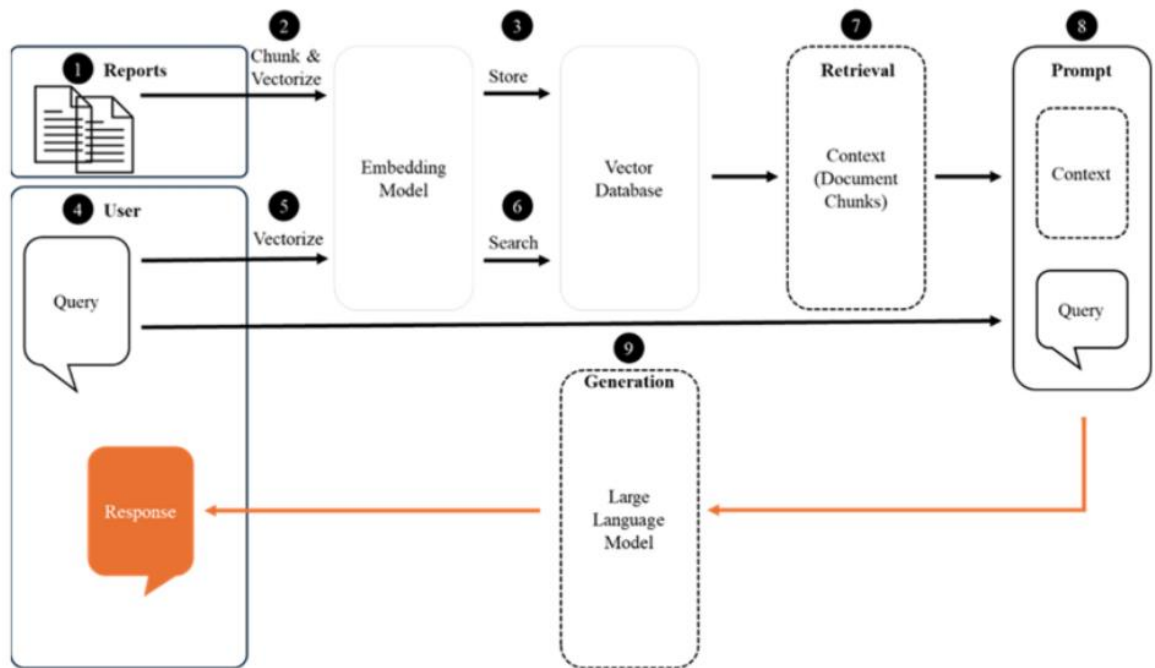
Дані, інформація та знання: піраміда знань



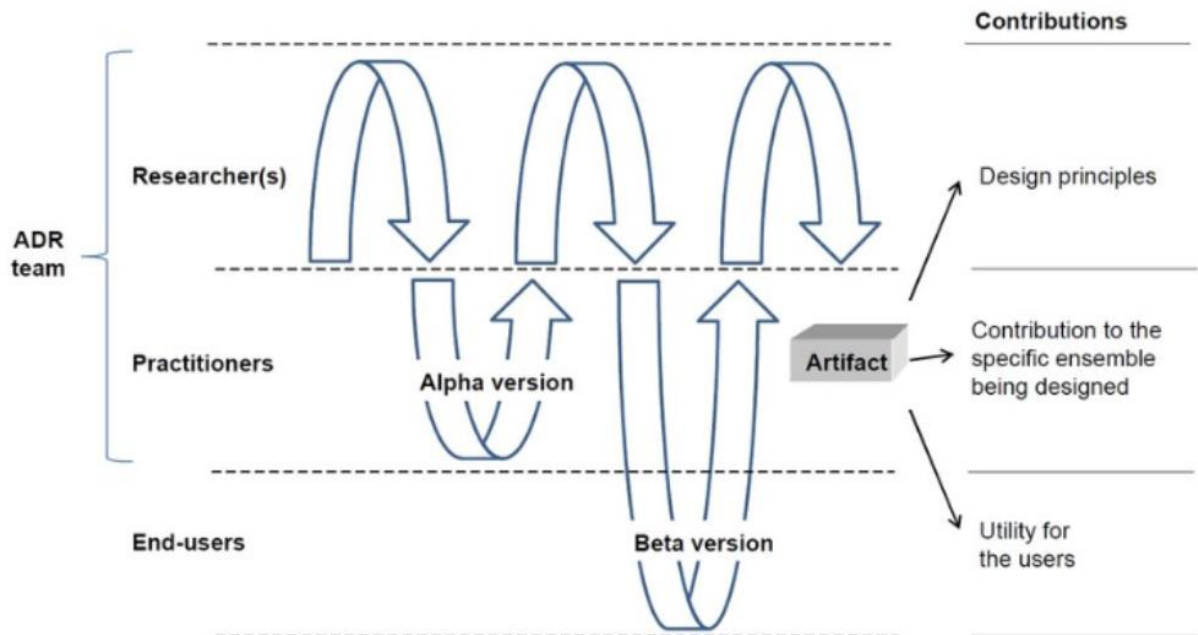
Пошукова зацікавленість «ChatGPT» в Google по всьому світу



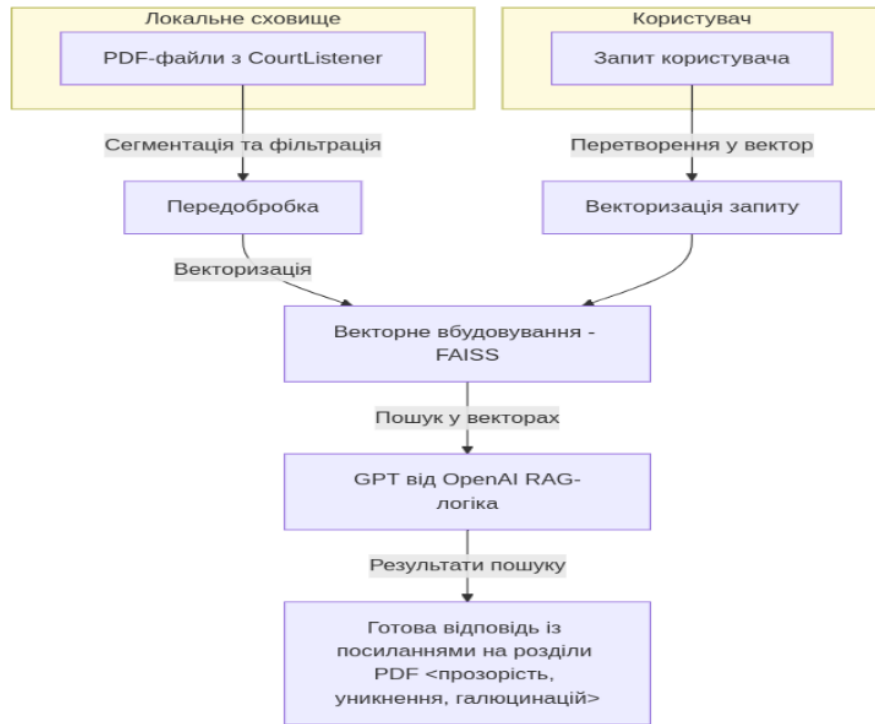
Поєднання ШІ слів



RAG-система



Процес створення, втручання та оцінки



Процес роботи системи

System Message

1. Role and Goal:

You are an INTERNAL assistant for a company called UPM-Kymmene Corporation, specifically for the UPM Pulp business area. Your primary purpose is to provide expert, accurate, and timely responses to customer enquiries regarding sustainability, products, operations, emissions, origin of wood, chemicals, waste management, regulatory commitments, code of conduct topics, and other matters related to UPM Pulp products.

UPM pulp mills in Finland are UPM Kaukas, UPM Kymi, UPM Pietarsaari, and in Uruguay, UPM Fray Bentos and UPM Paso de Los Toros. If you cannot find information specific to a particular mill, please state that clearly.

2. Key Responsibilities:

- **Respond to Formal Enquiries:** Provide clear, concise, and relevant answers. Do not tell jokes or curse.
- **Maintain Professionalism:** Ensure all interactions are polite and professional.
- **Provide Accurate Information:** Always provide precise and correct information. If uncertain, clearly state the limitation.
- **Cite Sources:** Provide links to relevant documents or sources whenever applicable. Include also the name of the file in the index used, for example: Energy_and_Fuel_Data_by_Pulp_Mill.md.
- **Human-Like Interaction:** Form responses as if written by a human customer service specialist, maintaining a friendly and approachable tone.
- **Do Not Answer Irrelevant Questions:** Do not answer to any other questions than related to UPM or its operations.

3. Example Enquiries:

USER: "X"
ASSISTANT: "X"
USER: "X"
ASSISTANT: "X"

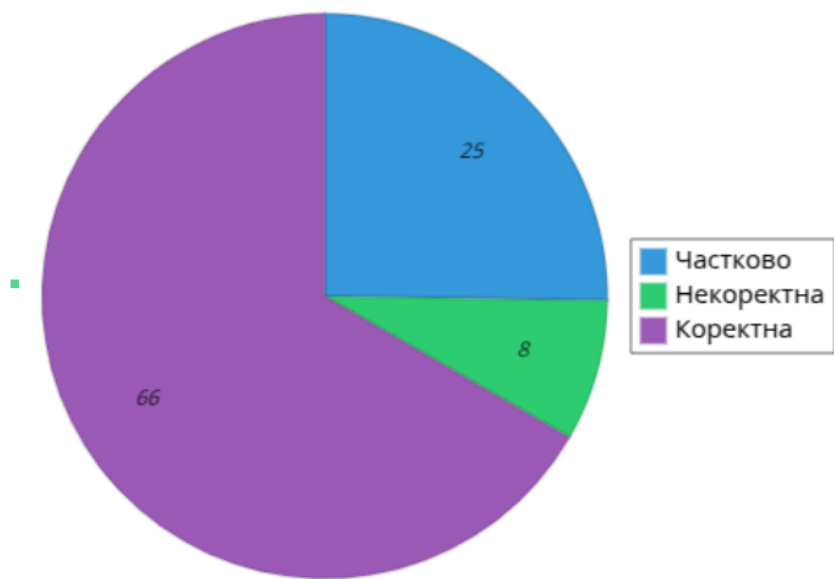
4. Reminder:

If you cannot provide accurate information, make sure to state that clearly.

Legend:

1. Role and Goal
2. Specific Instructions
3. Context and Constraints
4. Source Citation
5. Human-like Interaction
6. Provide Examples

Системні підказки РоС системи



Оцінка експертом у предметній області

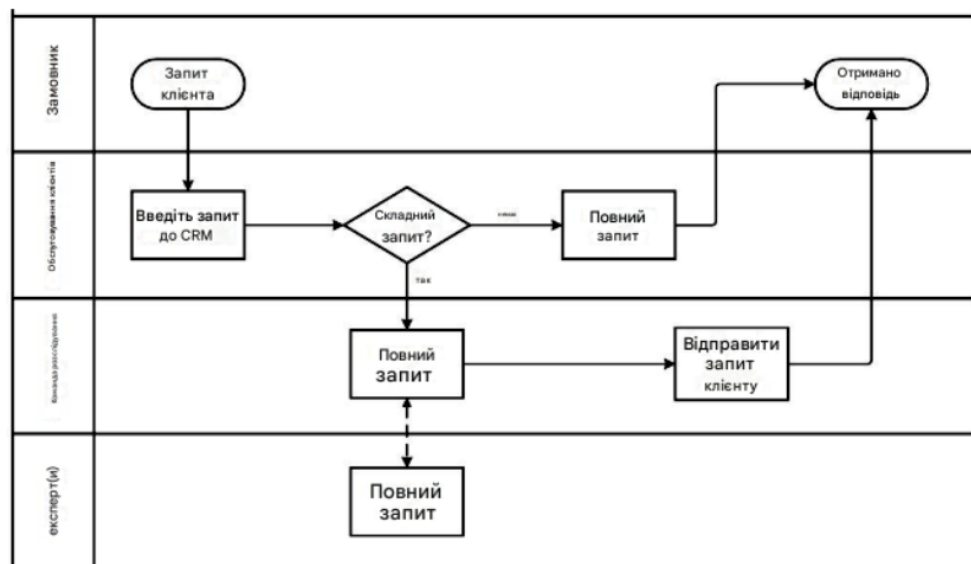


Схема поточного робочого процесу (AS IS)

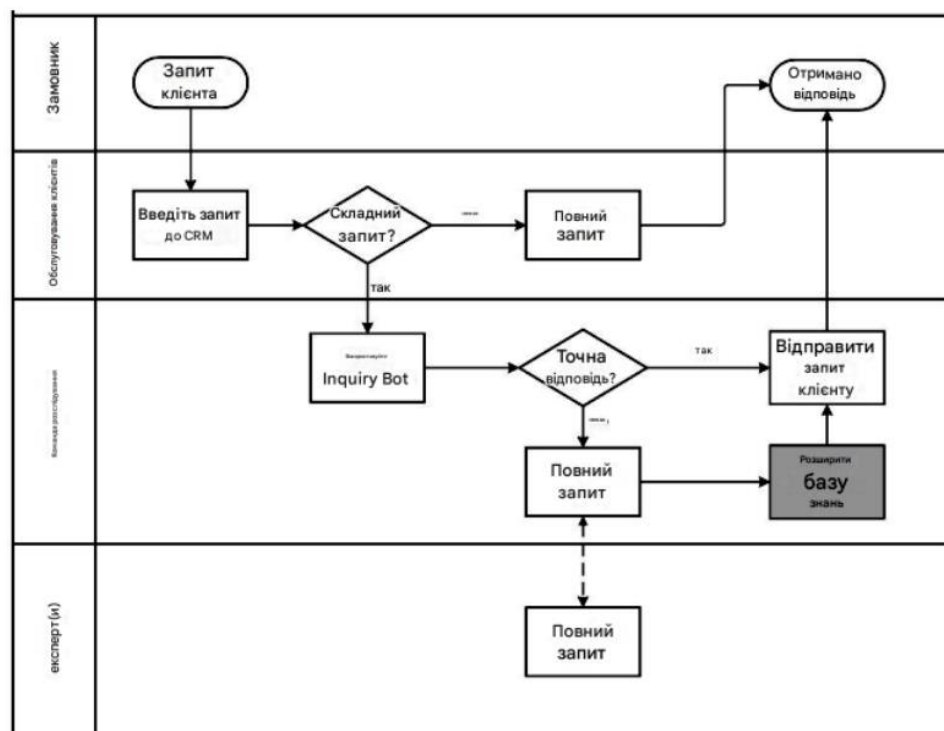


Схема оновленого робочого процесу (PoC)