Хмельницький національний університет
Факультет інформаційних технологій
Кафедра комп'ютерної інженерії та інформаційних систем

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

Галузь знань 12 – Інформаційні технології

Спеціальність 123 –Комп'ютерна інженерія

на тему «Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті»

КвРКІП.302184.23.02.01 ПЗ

Виконав: студент 2 курсу, група КІ2м-23-2 Олександр МАРЧУК
Підпис Ім'я, прізвище

Керівник д-р. техн. наук, професор Євген ФЕДОРОВ
Науковий ступінь, вчене звання Підпис Ім'я, прізвище

До захисту допускаю:
Зав. кафедри КІІС, доктор філософії, доцент
Ольга ПАВЛОВА
19 05 2025 р.

Хмельницький, 2025

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет Інформаційних технологій

Кафедра Комп'ютерної інженерії та інформаційних систем

Освітній рівень магістр

Галузь знань 12 Інформаційні технології

Спеціальність 123 Комп'ютерна інженерія

Освітня програма освітньо-наукова програма «Комп'ютерна інженерія та програмування»

ЗАТВЕРДЖУЮ

Зав. кафедри Ольга ПАВЛОВА

" 01 " 09 2024 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

Олександру МАРЧУКУ
Прізвище, ім'я, по батькові студента

1. Тема проекту (роботи) Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті

Керівник проекту (роботи) Євген ФЕДОРОВ Євгенович, д.т.н., професор
Прізвище, ім'я, по батькові, науковий ступінь, вчене звання

Затверджена наказом ректора університету від 08.01.2025 №8

2. Строк подання студентом проекту (роботи) на кафедру 01.05.2025 р.

3. Вихідні дані до проекту (роботи) Завдання на дипломне проектування

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

Аналіз існуючих методів генерації текстових запитів для чат-ботів

Розробка методу генерації запитів на основі ChatGPT

Проектування та реалізація системи генерації запитів на основі ChatGPT

Експериментальне дослідження та оцінка ефективності системи

5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень)

6.

6. Консультанти розділів кваліфікаційної роботи магістра

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Нормоконтроль	Сергій ЛИСЕНКО, професор кафедри КІС		
Антиплагіат	Андрій НІЧЕПОРУК, доцент кафедри КІС		

7. Дата видачі завдання «_01_»_____09_____2024р.

КАЛЕНДАРНИЙ ПЛАН

№з/п	Назва етапів (розділів) кваліфікаційної роботи магістра	Термін виконання етапів проекту (роботи)	Примітка
1	Вибір напряму дослідження та узгодження тематики КвРМ з керівником	01.09.2024	виконано
2	Ознайомлення з предметною областю; формулювання мети та задач дослідження; визначення об`єкта та предмета дослідження	01.10.2024	виконано
3	Робота над розділом 1 – аналіз відомих моделей, методів за темою; постановка задачі	01.11.2024	виконано
4	Робота над розділом 2 – розробка моделей для вирішення поставленої задачі	01.12.2024	виконано
5	Робота над науковою статтею	01.02.2025	виконано
6	Робота над розділом 3 – розробка методів для вирішення поставленої задачі	15.02.2025	виконано
7	Робота над розділом 4 – проектування та розробка ПЗ для вирішення поставленої задачі, експериментальна частина	01.04.2025	виконано
8	Оформлення пояснювальної записки згідно вимог	18.04.2025	виконано
9	Попередній захист ДРМ	29.04.2025	виконано
10	Захист ДРМ на засіданні ЕК	До 15.05.2025	

Студент _____ Олександр МАРЧУК
Підпис Ім'я, прізвище

Керівник роботи _____ Євген ФЕДОРОВ
Підпис Ім'я, прізвище

РЕФЕРАТ

Тема кваліфікаційної роботи магістра: Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боту

Автор роботи: Олександр МАРЧУК Олександрович

Керівник роботи: Євген ФЕДОРОВ Євгенович

Пояснювальна записка: 83 с., 18 рис., 6 табл., 4 дод., 81 джерел.

ГЕНЕРАЦІЯ ТЕКСТУ, ЧАТ-БОТ, ШТУЧНИЙ ІНТЕЛЕКТ, ТРАНСФОРМЕР, GPT, ДІАЛОГОВА СИСТЕМА, КОНТЕКСТНА РЕЛЕВАНТНІСТЬ

Об’єктом дослідження є процес генерації текстових запитів у діалогових системах з використанням мовних моделей.

Предметом дослідження є методи та моделі контекстно-залежної генерації текстових запитів у чат-ботах на основі трансформерних архітектур.

Метою кваліфікаційної роботи магістра є розробка методу генерації текстових запитів для чат-ботів на основі архітектури ChatGPT, що забезпечує формування контекстно релевантних та стилістично узгоджених текстів у процесі автоматизованої діалогової взаємодії.

Для розв’язання поставлених задач використовувалися методи системного аналізу, методи статистичного моделювання, алгоритми обробки природної мови (NLP), методи навчання трансформерних нейронних мереж, методи порівняльного аналізу та експериментальних досліджень продуктивності програмних систем.

Наукова новизна отриманих результатів:

набув подальшого розвитку метод генерації текстових запитів для чат-ботів шляхом поєднання трансформерної моделі ChatGPT із механізмами вагового контекстного зважування реплік та динамічної інтеграції зовнішніх джерел знань;

набула подальшого розвитку інформаційна технологія побудови контекстно-залежних діалогових систем з функціями автоматичного формування уточнюючих запитів та перевірки фактологічної достовірності відповідей.

Практична значимість отриманих результатів полягає у створенні гнучкого програмного рішення для автоматизації діалогової взаємодії у чат-ботах, здатного адаптуватися до різних предметних областей, підтримувати довгі діалоги та інтегрувати актуальні зовнішні знання у процесі генерації відповідей, що дозволяє підвищити якість обслуговування у клієнтських сервісах, освітніх платформах, юридичних консультаціях та державних онлайн-сервісах.

У першому розділі виконано аналіз сучасних підходів до генерації текстових запитів у чат-ботах, проаналізовано особливості скриптових діалогових систем, нейронних мереж типу LSTM/GRU та трансформерних моделей. Визначено переваги трансформерної архітектури GPT у контекстно-залежній генерації текстів та обґрунтовано доцільність її використання у системах діалогової взаємодії.

У другому розділі розроблено метод генерації текстових запитів для чат-ботів, що включає контекстне зважування реплік, інтеграцію зовнішніх баз знань та формування уточнюючих запитів. Описано математичну модель генерації текстів, а також визначено основні показники оцінки якості згенерованих текстів.

У третьому розділі розроблено архітектуру та програмний прототип системи генерації текстових запитів для чат-ботів на основі запропонованого методу. Описано програмну реалізацію, структуру системи та основні модулі, зокрема контекстне управління, модуль інтеграції зовнішніх знань та модуль збереження історії діалогів.

У четвертому розділі проведено експериментальне дослідження роботи системи, виконано порівняльний аналіз із традиційними та нейромережевими діалоговими системами, оцінено якість згенерованих текстів за метриками перплексії, BLEU та ROUGE. За результатами тестування підтверджено високу ефективність запропонованого методу у порівнянні з аналогами, особливо у складних багатокрокових діалогах.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

AI – Artificial Intelligence

API – Application Programming Interface

BLEU – Bilingual Evaluation Understudy

GPT – Generative Pre-trained Transformer

LSTM – Long Short-Term Memory

ML – Machine Learning

NLP – Natural Language Processing

PPL – Perplexity

RNN – Recurrent Neural Network

ROUGE – Recall-Oriented Understudy for Gisting Evaluation

SQL – Structured Query Language

ВСТУП

Машинне навчання дозволяє автоматизувати обробку природної мови, розширюючи можливості інтерактивних систем штучного інтелекту. Одним із найважливіших напрямів є генерація текстових запитів у діалогових системах, де необхідно забезпечити природність, точність і релевантність відповідей. Останні досягнення в області глибокого навчання, зокрема трансформерні архітектури, такі як GPT (Generative Pre-trained Transformer), відкривають нові можливості для створення високоякісних діалогових агентів. Проте існуючі методи мають ряд обмежень, пов’язаних із контекстною узгодженістю, ефективністю генерації та адаптацією до конкретних задач.

Зростання обсягу цифрової комунікації та активне впровадження штучного інтелекту в сфері автоматизованої взаємодії між людиною та комп’ютером зумовлюють необхідність розробки високоефективних систем генерації текстових запитів для чат-ботів. Традиційні скриптові системи, побудовані на правилах або простих методах класифікації намірів (intent detection), часто виявляються неефективними у ситуаціях, коли діалогова взаємодія є непередбачуваною або контекстуально складною [1]. Водночас використання трансформерних архітектур, зокрема GPT-моделей, дозволяє значно підвищити якість формування текстових запитів за рахунок врахування попереднього контексту, мовних нюансів та ймовірнісних моделей наступних реплік, що особливо важливо для створення інтелектуальних систем підтримки користувачів [2].

З огляду на стрімкий розвиток мовних моделей, особливого значення набуває їх адаптація до потреб українського бізнесу та державного сектору, де актуальними є завдання автоматизації обслуговування, створення ефективних багатомовних чат-ботів, а також побудова систем, здатних коректно інтерпретувати запити у специфічних умовах, наприклад, під час кризових комунікацій або правових консультацій. Умови цифрової трансформації в Україні, зокрема активне впровадження електронного урядування та діджиталізації сервісів, створюють нагальну потребу у розробці інноваційних

підходів до генерації текстових запитів, які забезпечують високу якість обробки природної мови та гнучке налаштування під конкретні галузі [3].

Серед сучасних науковців, які активно досліджують методи застосування трансформерів та GPT-моделей у генерації текстових запитів для чат-ботів, варто відзначити Даріо Амодеї, одного з провідних дослідників у сфері масштабованих мовних моделей та їхньої безпеки, Іллю Суцкевера, співзасновника OpenAI, який безпосередньо долучався до розробки GPT-моделей, а також Джейсона Вея, дослідника Google DeepMind, який вивчає ефективність few-shot і zero-shot навчання для діалогових систем [4, 5]. Важливий внесок у розвиток цієї тематики також зробили Ітан Перез, який працює над покращенням керованості мовних моделей у діалогах, Маргарет Мітчелл, яка досліджує етичні аспекти використання трансформерів у чат-ботах, а також Мікаеля Шустера та Чжилінь Яна, які зосереджені на оптимізації мовних моделей для генерації контекстуально релевантних та точних текстів у діалогових системах [6]. Їхні дослідження формують теоретичну та практичну основу для подальшого вдосконалення методів автоматичної генерації текстових запитів та розвитку інтелектуальних чат-ботів.

Незважаючи на значну кількість наукових досліджень, присвячених розробці GPT-моделей та їх використанню у генеративних системах [7], проблема ефективної генерації текстових запитів у діалогових системах на підставі GPT залишається недостатньо вивченою. Зокрема, потребують подальшої розробки методи автоматичної адаптації мовних моделей до специфіки україномовного контенту, механізми зменшення випадків генерації нерелевантних або «галюциногенних» запитів, а також способи інтеграції зовнішніх баз знань для підвищення точності та достовірності створюваних текстів [8].

Відмінність даного дослідження полягає у розробці методології побудови системи, яка не лише генерує текстові запити до чат-боту, але й враховує специфіку мовного середовища, динаміку контекстуальних змін та використання зовнішніх джерел знань. Такий підхід сприятиме підвищенню рівня інтелектуальності сучасних діалогових систем, розширить можливості їх

використання у різних галузях, від електронної комерції до державного управління, а також забезпечить розвиток нового напряму у створенні адаптивних мовних моделей для українськомовного цифрового простору.

Метою кваліфікаційної роботи є розробка методу та системи трансформера на підставі ChatGPT для генерації текстових запитів у чат-боті, що забезпечує підвищену точність та контекстну релевантність згенерованих повідомлень.

Поставлена мета досягається розв'язанням таких основних задач:

- виконати аналіз існуючих методів генерації текстових запитів у чат-ботах, оцінити їхні переваги та недоліки;
- розробити метод генерації текстових запитів на основі архітектури ChatGPT, що забезпечує підвищену точність та контекстну узгодженість;
- формалізувати задачу генерації запитів, побудувати математичну модель роботи системи;
- визначити ключові метрики оцінки якості згенерованих текстів (перплексія, BLEU, ROUGE) та застосувати їх для оцінки роботи системи;
- розробити архітектуру програмного забезпечення, інтегрувати ChatGPT у чат-бот;
- провести тестування, оцінити ефективність генерації запитів у порівнянні з іншими моделями;
- визначити обмеження методу та сформулювати рекомендації для подальшого удосконалення.

Об'єктом дослідження є процес автоматичної генерації текстових запитів у діалогових системах

Предметом дослідження є методи та алгоритми генерації текстових запитів на основі трансформерних моделей, зокрема ChatGPT.

Наукова новизна отриманих результатів:

1. Запропоновано новий метод генерації текстових запитів для чат-ботів на основі архітектури ChatGPT, який враховує специфіку діалогових сценаріїв.

2. Набула подальшого розвитку методика оцінки якості згенерованих текстів, що дозволяє підвищити релевантність відповідей у діалогових системах.

3. Удосконалено математичну модель генерації запитів у чат-ботах шляхом інтеграції контекстних факторів.

Практична значимість отриманих результатів полягає у результаті виконаного наукового дослідження розроблена та реалізована система генерації текстових запитів, яка забезпечує високу якість та контекстну релевантність автоматично створюваних повідомлень у чат-ботах. Запропонований метод дозволяє підвищити точність та узгодженість відповідей, що є критично важливим для інтерактивних систем обробки природної мови.

Для розв’язання поставлених задач використовувалися основні положення глибокого навчання, трансформерних нейромереж, аналізу природної мови та методів оптимізації машинного навчання.

За темою кваліфікаційної роботи опубліковано одну публікацію у збірнику наукових праць за матеріалами VI Науково-практичної конференції «Безпека енергетики в епоху цифрової трансформації» [30]

1 АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ГЕНЕРАЦІЇ ТЕКСТОВИХ ЗАПИТІВ ДЛЯ ЧАТ-БОТІВ

1.1 Стан наукової думки у сфері генерації текстових запитів чат-ботів

Генерація текстових запитів є центральною задачею в розробці автоматизованих діалогових систем, таких як чат-боти та віртуальні асистенти. Ця сфера зазнала значних трансформацій протягом останніх десятиліть, що обумовлено зростаючим інтересом до високоточних моделей обробки природної мови (NLP) та впровадженням передових технологій, зокрема нейронних мереж і трансформерних архітектур [9].

Початкові спроби автоматизованої генерації тексту базувалися на статистичних методах та моделях, заснованих на марковських процесах. Ці підходи використовували ймовірнісні розподіли для прогнозування наступного слова в послідовності, спираючись на попередні слова. Хоча такі методи були революційними для свого часу, вони мали обмежену здатність враховувати довготривалі залежності та контекст, що призводило до генерації менш зв’язних та когерентних текстів.

З появою рекурентних нейронних мереж (RNN) відбувся значний прорив у сфері обробки послідовних даних. RNN дозволили моделям враховувати попередній контекст при генерації кожного наступного елемента послідовності (рисунок 1.1).

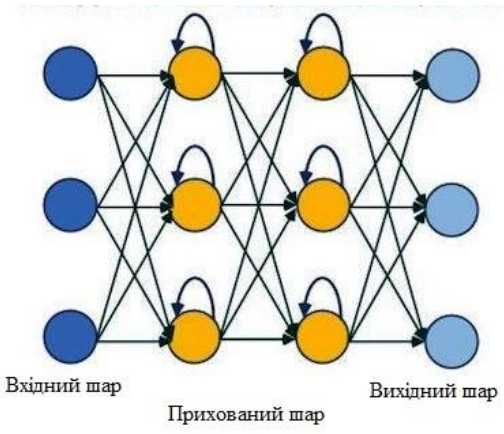


Рисунок 1.1 – Принцип побудови рекурентної нейронної мережі

Особливої уваги заслуговують варіації RNN, такі як довга короткочасна пам’ять (LSTM) та шлюзові рекурентні блоки (GRU), які були розроблені для подолання проблеми зникнення градієнта та забезпечення ефективного навчання на довгих послідовностях. Ці моделі значно покращили якість генерації тексту, дозволяючи враховувати більш складні залежності та контексти [10].

Однак справжню революцію в генерації текстових запитів спричинило впровадження трансформерних архітектур. У 2017 році дослідники з Google представили модель трансформера, яка кардинально змінила підхід до обробки послідовностей. На відміну від RNN, трансформери не покладаються на послідовну обробку даних, що дозволяє їм ефективно обробляти довгі послідовності без втрати контексту. Ключовим компонентом трансформера є механізм самоуваги (self-attention), який дозволяє моделі оцінювати важливість кожного елемента в послідовності щодо інших, забезпечуючи глибше розуміння контексту та взаємозв’язків у тексті [11, 12].

Сучасні наукові дослідження в галузі генерації текстових запитів зосереджені на декількох ключових напрямах. По-перше, це розробка та вдосконалення нейромережевих моделей, зокрема трансформерних архітектур, які забезпечують високу якість та гнучкість генерації тексту. По-друге, значна увага приділяється практичним аспектам впровадження цих моделей у реальні системи, такі як чат-боти та віртуальні асистенти. Це включає адаптацію моделей до специфічних доменів, навчання на спеціалізованих корпусах даних та оптимізацію для роботи в реальному часі [13].

Наприклад, моделі GPT (Generative Pre-trained Transformer), розроблені OpenAI, продемонстрували високий рівень здатності генерувати зв’язний та контекстуально релевантний текст, що робить їх ідеальними кандидатами для використання в діалогових системах. Завдяки архітектурі трансформерів ці моделі здатні ефективно обробляти довгі контексти та враховувати значення слів залежно від їхнього оточення, що особливо важливо у процесі підтримки природної взаємодії з користувачами. GPT-моделі навчаються на масштабних обсягах текстових даних з різних джерел, включно з науковими статтями,

художніми творами, публіцистикою та розмовними діалогами, що дозволяє їм накопичувати знання з різних доменів та генерувати відповіді, які максимально відповідають запитам користувачів, адаптуючи стиль і тон до контексту конкретної ситуації (рисунок 1.2) [14, 15].

Рисунок 1.2 –Простий механізм вводу-виводу, який лежить в основі чудових можливостей GPT

Незважаючи на значні досягнення, сфера генерації текстових запитів для чат-ботів стикається з низкою викликів. Одним з основних є забезпечення контекстної релевантності та точності відповідей, особливо в умовах неоднозначних або складних запитів. Іншим важливим аспектом є етичність та упередженість моделей: оскільки моделі навчаються на великих обсягах даних, вони можуть несвідомо переймати упередження, присутні в цих даних, що може призводити до небажаних або дискримінаційних відповідей [14].

Перспективні напрями досліджень включають розробку методів для кращого розуміння намірів користувача, інтеграцію зовнішніх баз знань для забезпечення фактичної точності відповідей, а також впровадження механізмів контролю та корекції упереджень у моделях. Крім того, оптимізація моделей для роботи на пристроях з обмеженими ресурсами та забезпечення їхньої

енергоефективності є важливими аспектами для широкого впровадження цих технологій у повсякденне життя [15].

Еволюція методів генерації текстових запитів для чат-ботів відображає загальний прогрес у галузі обробки природної мови. Перехід від простих статистичних моделей до складних нейромережевих архітектур, таких як трансформери, відкриває нові можливості для створення більш інтелектуальних та чутливих до контексту діалогових систем.

1.2 Аналіз архітектур трансформерів для обробки природної мови

Архітектура трансформера стала революційним проривом у галузі обробки природної мови (NLP), запропонована дослідниками у 2017 році в роботі «Attention is All You Need» [16]. Ця модель замінила традиційні рекурентні нейронні мережі (RNN) та довго-короткострокову пам’ять (LSTM), запропонувавши новий підхід до обробки послідовностей даних без використання рекурентності.

Трансформер складається з двох основних частин: енкодера та декодера, кожен з яких містить кілька шарів (рисунок 1.3).

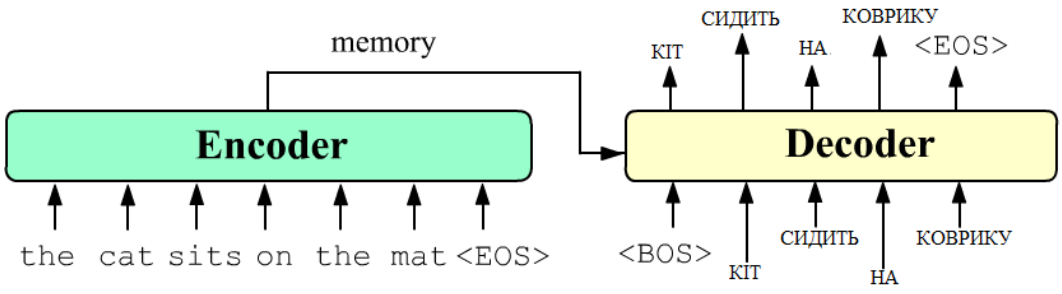


Рисунок 1.3 – Основні частини трансформера

Енкодер відповідає за обробку вхідної послідовності та перетворення її у внутрішнє представлення, тоді як декодер використовує це представлення для генерації вихідної послідовності.

Енкодер – складається з N ідентичних шарів, кожен з яких містить два підшари: механізм самоуваги (self-attention) та позиційно-нормалізовану повнозв’язну нейронну мережу.

Декодер також складається з N ідентичних шарів, але містить додатковий підшар уваги до виходу енкодера, що дозволяє моделі фокусуватися на відповідних частинах вхідної послідовності під час генерації виходу.

Однією із особливостей трансформера є механізм самоуваги, який дозволяє моделі оцінювати важливість різних слів у послідовності відносно одне одного. Це досягається шляхом обчислення вагових коефіцієнтів для кожної пари слів, що дозволяє моделі враховувати контекст при обробці кожного слова. Такий підхід забезпечує ефективну обробку довготривалих залежностей у тексті та паралельну обробку даних, що значно підвищує швидкість навчання та функціонування порівняно з RNN [17].

Оскільки трансформер не використовує рекурентність, модель не має вбудованого уявлення про порядок слів у послідовності. Для врахування позиційної інформації додається позиційне кодування до вхідної послідовності. Це кодування використовує синусоїдальні функції різних частот для представлення позицій, що дозволяє моделі розрізняти позиції слів та враховувати їх порядок при обробці [16].

Однією з ключових переваг трансформерної архітектури є здатність до паралельної обробки вхідних даних. На відміну від рекурентних нейронних мереж (RNN), які опрацьовують послідовність поелементно, трансформери обробляють усі її компоненти одночасно. Такий підхід дозволяє значно зменшити час навчання та ефективно використовувати апаратні ресурси.

Крім того, трансформери демонструють високу здатність до моделювання довготривалих залежностей між словами в тексті. Завдяки механізму самоуваги (self-attention), модель може враховувати зв’язки між будь-якими елементами послідовності незалежно від їхньої відстані, що суттєво покращує якість текстової обробки.

Гнучкість трансформерної архітектури забезпечує її широке застосування в різноманітних завданнях обробки природної мови. Вона успішно використовується для машинного перекладу, текстової класифікації, генерації тексту та інших задач у сфері NLP.

Моделі сімейства GPT (Generative Pre-trained Transformer), розроблені OpenAI, є яскравим прикладом використання трансформерної архітектури для генерації тексту (рисунок 1.4).

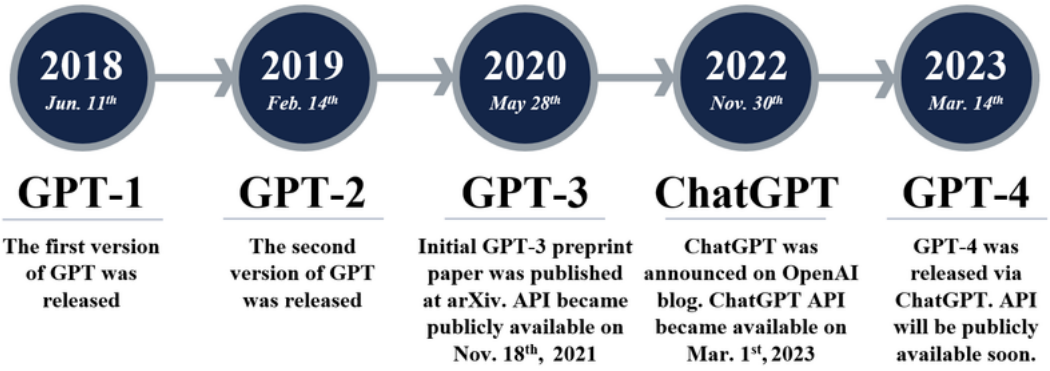


Рисунок 1.4 – Еволюція моделей GPT[18].

Ці моделі навчаються на великих обсягах текстових даних, що дозволяє їм генерувати зв’язний та контекстуально релевантний текст. GPT-2 та GPT-3 продемонстрували високі результати у вирішенні різноманітних завдань, включаючи діалогові системи, автоматичне доповнення тексту та відповіді на запитання [18].

Архітектура трансформера стала фундаментом для багатьох сучасних моделей обробки природної мови, забезпечуючи ефективну та гнучку обробку текстових даних. Її здатність паралельно обробляти послідовності та враховувати довготривалі залежності робить її незамінною у розробці сучасних NLP-систем.

1.3 Використання GPT-моделей у генерації тексту та діалогових системах

З появою великих мовних моделей (Large Language Models – LLMs), таких як GPT-3, GPT-4, а також їхніх модифікацій, генерація тексту та створення діалогових систем вийшли на якісно новий рівень. Ці моделі стали важливою віхою в розвитку обчислювальної лінгвістики та штучного інтелекту, оскільки вони суттєво розширили можливості автоматизованої обробки природної мови [8].

GPT-моделі, навчені на терабайтах текстових даних із різноманітних джерел – від літературних творів та наукових статей до соціальних мереж і форумів – демонструють високу здатність до створення зв'язних, стилістично узгоджених та контекстуально релевантних текстів. Завдяки широкому охопленню доменів і стилів, ці моделі легко адаптуються до специфічних тематик, що дозволяє застосовувати їх у сферах від технічної підтримки до творчого письма, наукової аналітики та створення маркетингового контенту.

Ще однією з особливостей GPT-моделей є їхня здатність враховувати глобальний контекст у межах довгих текстових фрагментів або діалогів, що забезпечує логічну послідовність та стилістичну єдність у створених відповідях. Це стало можливим завдяки використанню механізму самоуваги (self-attention), який дозволяє моделі «розуміти», як різні частини тексту взаємопов'язані між собою. У результаті GPT здатна не лише реагувати на прямі запитання, але й підтримувати складні діалоги, враховуючи раніше надані користувачем деталі, а також передбачати подальший напрямок розмови.

Крім того, великі мовні моделі демонструють здатність до переносу знань між різними задачами, завдяки чому вони можуть виконувати широкий спектр функцій – від генерації коду до написання юридичних документів чи аналізу медичних звітів. Ця універсальність є однією з причин їхнього стрімкого поширення у бізнесі, науці та державному управлінні.

Розвиток технологій донавчання (fine-tuning) дозволяє адаптувати універсальні GPT-моделі до конкретних доменів та потреб окремих організацій. Наприклад, спеціалізовані моделі для медичних чат-ботів проходять додаткове навчання на вибірках з клінічних даних, що підвищує їхню точність у відповідях

на медичні запити. Аналогічно, моделі для юридичних консультацій навчаються на юридичних документах та судових рішеннях [19].

Однак масштабність та потужність GPT-моделей супроводжуються низкою викликів, зокрема необхідністю обробки величезних обсягів даних у режимі реального часу, забезпеченням етичного використання та контролю якості генерованого контенту. Незважаючи на це, GPT-моделі продовжують трансформувати сферу діалогових систем, поступово відсуваючи на другий план традиційні скриптові чат-боти та системи, що базуються на жорстко закодованих правилах.

У таблиці 1.1 наведено порівняння трансформера GPT з іншими підходами у діалогових системах.

Таблиця 1.1 – Порівняння GPT з іншими підходами у діалогових системах

Параметр	GPT-моделі	Правилозалежні системи	Гібридні системи
Гнучкість	Висока	Низька	Середня
Якість обробки нового контенту	Висока	Низька	Висока
Залежність від попереднього налаштування	Низька	Висока	Середня
Вартість впровадження	Висока	Низька	Середня
Проблеми з галюцинаціями	Присутні	Відсутні	Часткові

Використання GPT у реальних діалогових системах стає все більш поширеним, оскільки компанії прагнуть автоматизувати комунікацію, підвищити якість обслуговування клієнтів та зменшити навантаження на працівників служб підтримки. GPT-моделі впроваджуються у створення віртуальних помічників, які не лише відповідають на типові запитання, але й здатні аналізувати контекст попередніх звернень, персоналізувати відповіді та навіть прогнозувати додаткові

потреби користувачів. Такі системи застосовуються у банківській сфері, електронній комерції, медицині, логістиці, а також у внутрішніх корпоративних системах для автоматизації комунікацій між підрозділами.

Значну роль у популяризації подібних рішень відіграють технологічні гіганти. OpenAI інтегрує GPT у свої продукти, зокрема у популярний інтерфейс ChatGPT, який використовується як кінцевими користувачами, так і компаніями для створення власних рішень. Microsoft впроваджує GPT у своєму Copilot для пакету Office, що дозволяє користувачам створювати та редагувати документи, електронні таблиці та презентації за допомогою текстових команд, що значно підвищує продуктивність. Компанія Duolingo, відома своїм додатком для вивчення мов, адаптує GPT для створення персоналізованих уроків та діалогів з урахуванням рівня знань та інтересів кожного користувача, що дозволяє зробити навчання більш природним і динамічним [20].

Ефективність GPT-моделей значно підвищується при їхньому поєднанні з зовнішніми джерелами знань, зокрема з корпоративними базами знань, внутрішніми архівами документації або публічними репозиторіями. У таких випадках GPT використовується не лише як генератор тексту, але й як інструмент для швидкого витягування та узагальнення релевантних даних. Застосування підходу Retrieval-Augmented Generation (RAG), що зображений на рисунку 1.5, дозволяє моделям виконувати пошук актуальної інформації у режимі реального часу, забезпечуючи актуальність та достовірність відповідей, особливо у швидкозмінних сферах, таких як фінансові ринки, законодавство чи медицина.

Це також допомагає частково знизити ймовірність так званих галюцинацій, коли модель генерує правдоподібні, але фактично помилкові або вигадані твердження.

Однак, навіть при активному розвитку технології, GPT-моделі залишаються вразливими до низки викликів. Одним із найбільш обговорюваних є проблема галюцинацій, коли модель створює неправдиві відповіді з високим рівнем впевненості. Це особливо критично у сферах, де помилка може мати юридичні чи етичні наслідки, наприклад у медицині або праві. Крім того, широке поширення

мовних моделей загострює питання етичного використання, зокрема ризики поширення дезінформації, створення фейкових новин або маніпулятивного контенту. Регулювання таких процесів стає нагальною потребою, особливо в контексті стрімкого розвитку генеративного штучного інтелекту [21].

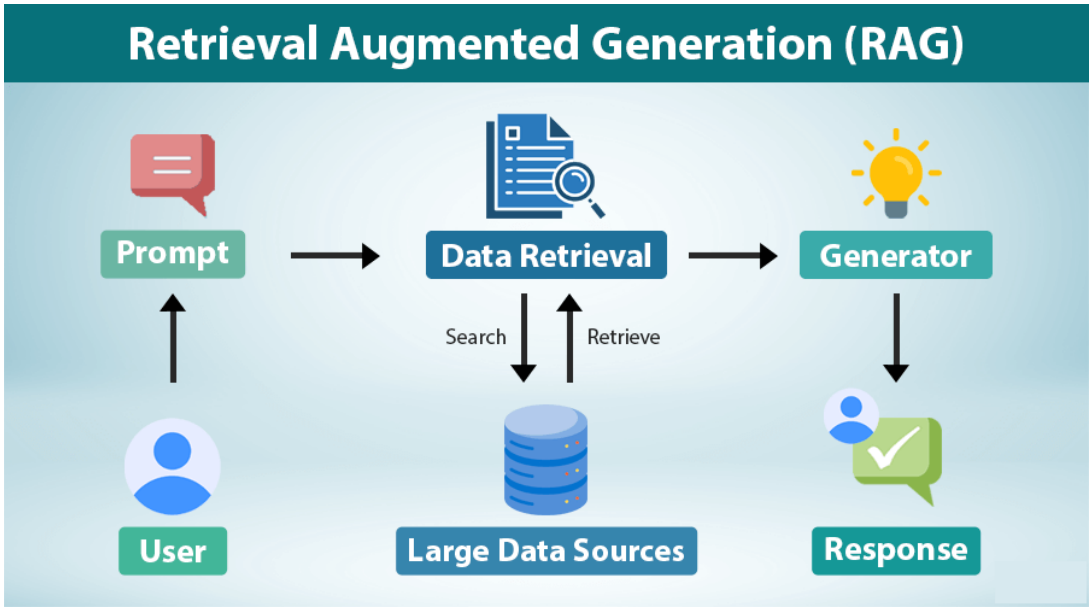


Рисунок 1.5 – Підхід Retrieval-Augmented Generation (RAG) [21]

Окремим викликом залишається висока обчислювальна вартість використання великих мовних моделей. Для повноцінного розгортання GPT у бізнес-середовищі необхідні значні обчислювальні ресурси, що може бути обтяжливим для малого та середнього бізнесу. У зв'язку з цим дедалі більшої популярності набувають гібридні підходи, коли для простих запитів використовується легша модель або скриптова система, а для складніших — звернення до GPT.

Подальший розвиток GPT-моделей безпосередньо пов'язаний з інтеграцією мультимодальних можливостей, коли одна система одночасно аналізує текст, зображення, відео та аудіо. Це відкриває нові перспективи для створення ще більш розумних віртуальних агентів, які зможуть, наприклад, аналізувати фотографії документів, розпізнавати обличчя чи емоції співрозмовника або надавати консультації на основі аудіодзвінків. Паралельно ведеться робота над

підвищенням прозорості мовних моделей, це означає що користувачі та розробники отримують більше інструментів для відстеження джерел інформації, на які спиралася модель, та пояснення логіки її роботи.

У цьому контексті важливим вектором еволюції GPT-моделей є створення спеціалізованих мовних систем, орієнтованих на вирішення вузькопрофільних завдань. Так, у галузі охорони здоров'я GPT-моделі проходять навчання на основі клінічних даних, що дозволяє їм ефективно допомагати лікарям у постановці діагнозів і формуванні медичної документації. У юридичній сфері такі моделі застосовуються для аналізу судової практики, складання юридичних документів і проведення правових досліджень. Таким чином, GPT перетворюються з універсальних текстогенераторів на високоефективні професійні інструменти, які можуть працювати у тісній взаємодії з фахівцями, покращуючи як продуктивність, так і якість прийнятих рішень [22].

1.4 Методи оцінки якості згенерованих запитів у чат-ботах

Оцінка якості згенерованих текстових запитів у чат-ботах є ключовим етапом у процесі розробки та оптимізації систем генерації природної мови. Здатність моделі створювати зв'язні, змістовні та релевантні до контексту репліки визначає загальну ефективність діалогової системи та рівень задоволеності користувачів. Особливої актуальності набуває комплексна оцінка якості запитів у системах, що працюють на основі великих мовних моделей, зокрема GPT, де тексти формуються у режимі реального часу на підставі ймовірнісного прогнозування.

Серед основних методів автоматизованої оцінки, які застосовуються для аналізу якості згенерованих текстів, слід відзначити перплексію (Perplexity, PPL). Цей показник вимірює рівень «невизначеності» моделі щодо наступного слова у тексті. Чим нижче значення перплексії, тим краще модель адаптована до роботи з конкретним мовним корпусом та глибше розуміє синтаксичні та семантичні закономірності мови. Однак слід зазначити, що низька перплексія не завжди гарантує високу змістовність чи відповідність інтересам користувача, оскільки ця

метрика оцінює лише прогнозну якість, а не релевантність до контексту чи завдання [23].

Для оцінки текстів, які мають відповідати певним еталонним формулюванням, широко застосовується метрика BLEU (Bilingual Evaluation Understudy). Вона порівнює згенеровані запити з набором референтних текстів, аналізуючи кількість спільних n-грамів, а також точність передачі структури та лексики. BLEU добре працює у сценаріях, де важлива формальна відповідність, наприклад, при генерації стандартних клієнтських запитів або типових діалогових шаблонів. Водночас для оцінки креативних або варіативних запитів, характерних для відкритих діалогів, ця метрика має обмежене застосування, оскільки не враховує семантичну еквівалентність різних формулювань [24].

Додатково використовується метрика ROUGE (Recall-Oriented Understudy for Gisting Evaluation), яка особливо ефективна для аналізу текстів, що мають узагальнюючий або перефразований характер. Вона порівнює кількість спільних слів або фраз між згенерованим текстом і еталоном, приділяючи особливу увагу повноті відображення інформації. Це робить ROUGE корисною для оцінки запитів, які мають підсумковий або уточнюючий характер [25].

Разом із автоматизованими метриками дедалі більшого значення набувають методи, засновані на людській оцінці. Вони передбачають залучення експертів або користувачів до аналізу таких характеристик згенерованих запитів, як доречність, логічна зв’язність, стилістична відповідність та коректність інформації. Комбіновані підходи, які поєднують автоматичні метрики із людськими оцінками, дозволяють отримати найбільш об’єктивну та всебічну картину якості текстових запитів у діалогових системах.

З метою порівняння підходів до оцінки якості, наведено узагальнена таблиця 1.2, що відображає ключові особливості популярних методів.

Перспективним напрямом розвитку методів оцінки є інтеграція мультимодальних підходів, коли текстова якість оцінюється у поєднанні з аналізом інших каналів комунікації, наприклад, міміки або голосу користувача.

Крім того, зростає увага до адаптивних метрик, які враховують специфіку домену та цільової аудиторії, що особливо важливо для україномовних чат-ботів.

Таблиця 1.2 – Порівняння підходів до оцінки якості метрик

Метрика	Основний принцип	Переваги	Обмеження
Perplexity (PPL)	Оцінка здатності моделі прогнозувати наступне слово	Відображає загальну якість мовної моделі	Не враховує змістовну релевантність
BLEU	Порівняння з референтними текстами	Висока об’єктивність у формалізованих сценаріях	Погано працює з варіативними текстами
ROUGE	Аналіз збігів на рівні фраз та слів	Ефективна для підсумкових текстів	Не враховує глибинну семантику
Людська оцінка	Експертний або користувацький аналіз якості тексту	Враховує всі аспекти якості	Суб’єктивність та витрати часу

1.5 Проблеми, які залишаються невирішеними в існуючих підходах

Попри значні досягнення у сфері генерації текстових запитів для чат-ботів, низка важливих проблем залишається актуальною та нерозв’язаною у сучасних підходах, особливо при використанні трансформерних моделей, таких як GPT. Однією з ключових проблем є забезпечення контекстуальної узгодженості згенерованих запитів у довгих діалогах, де користувач поступово розкриває свої наміри або коригує попередні формулювання. Наявні GPT-моделі демонструють високу якість генерації у межах коротких сесій, проте зі збільшенням довжини діалогу виникають труднощі із збереженням логічної цілісності, особливо коли йдеться про складні або багатоступеневі запити [26].

Ще однією суттєвою проблемою є відсутність гарантованої достовірності згенерованих запитів та можливість так званих «галюцинацій», коли модель

генерує фактично некоректну або вигадану інформацію, яка може ввести користувача в оману. Ця проблема загострюється у вузькоспеціалізованих сферах, де модель не має достатньої кількості тренувальних даних або де важлива точна відповідність встановленим термінам та поняттям. Більшість сучасних підходів, включаючи OpenAI GPT-4 та Google Gemini, намагаються мінімізувати ці ризики шляхом інтеграції зовнішніх баз знань або використання механізмів перевірки фактів (fact-checking), однак ці рішення досі не є універсальними та потребують значних обчислювальних ресурсів [27].

Окремо варто відзначити проблему недостатньої адаптивності моделей до специфічних доменів та мовних середовищ, зокрема україномовного сегменту. Більшість сучасних мовних моделей тренуються переважно на англомовних корпусах, що призводить до зниження якості генерації текстів іншими мовами через менш репрезентативні мовні патерни та стилістичні особливості. Це створює додаткові виклики при розробці україномовних чат-ботів, де важливо не лише генерувати граматично та стилістично коректні тексти, але й враховувати культурно обумовлені особливості комунікації та термінологічні стандарти в різних галузях [28].

Ще одна проблема пов’язана з балансуванням між гнучкістю генерації та контролем над змістом згенерованих запитів. Сучасні генеративні моделі на основі трансформерів, зокрема GPT, мають надзвичайно широку мовну компетенцію, однак ця гнучкість може призводити до того, що згенеровані запити не завжди відповідають бізнес-логіці конкретної системи. Наприклад, у чат-ботах для фінансових консультацій надмірно «творчі» формулювання або розлогі відповіді можуть бути недоречними, тоді як у системах навчання або підтримки клієнтів навпаки цінується докладність та варіативність.

Також слід відзначити обмежені можливості налаштування існуючих генеративних систем для потреб конкретних користувачів або груп користувачів. Персоналізація запитів залишається викликом навіть для найсучасніших моделей, оскільки врахування індивідуальних особливостей потребує додаткового навчання на приватних даних, що часто стикається з етичними та юридичними

обмеженнями, зокрема у контексті дотримання GDPR та інших нормативних актів щодо захисту персональних даних [29].

Отже, сучасні методи генерації текстових запитів для чат-ботів, попри значний прогрес, потребують подальшого вдосконалення за такими напрямами: підвищення контекстної стійкості у довгих діалогах, мінімізація фактологічних помилок, адаптація до специфічних мовних середовищ, забезпечення контрольованої генерації та інтеграція механізмів персоналізації. Вирішення цих проблем стане основою для створення нових методів генерації запитів, які забезпечать не лише технічну якість, але й практичну цінність для кінцевих користувачів

1.6 Постановка задачі дослідження

Сучасні генеративні системи на основі трансформерів, зокрема GPT-моделі, демонструють високу ефективність у створенні текстів природною мовою, проте їхнє застосування для генерації текстових запитів у чат-ботах супроводжується низкою проблем, що залишаються нерозв’язаними. Зокрема, існуючі підходи не завжди забезпечують релевантність та змістовну відповідність згенерованих запитів у контексті тривалих діалогів, особливо коли йдеться про багатокрокові сценарії взаємодії, які передбачають уточнення намірів користувача, адаптацію до попередніх відповідей та врахування зовнішньої інформації.

Дослідження показують, що більшість генеративних систем не мають гнучких механізмів адаптації до специфічних мовних середовищ, що є особливо актуальним для україномовних чат-ботів. Наявні мовні моделі, навіть при використанні сучасних технік донавчання, демонструють обмежену здатність коректно обробляти запити, що містять складні мовні конструкції, регіональні особливості або професійну термінологію. Крім того, типова архітектура GPT-моделей не завжди забезпечує ефективний контроль над змістом згенерованих текстів, що може призводити до появи інформаційних «галюцинацій» або нерелевантних формулювань.

Зважаючи на це, виникає науково-практична проблема, яка полягає у необхідності розробки методу та системи генерації текстових запитів для чат-ботів на основі GPT-моделей, що враховуватиме контекст діалогу, адаптуватиметься до специфіки предметної області та забезпечуватиме високу якість, релевантність і контрольованість генерованих текстів. Такий підхід має поєднувати переваги сучасних трансформерних моделей із додатковими механізмами контролю змісту та адаптації до мовних і функціональних особливостей конкретних застосувань.

Відповідно, метою даного дослідження є розробка та обґрунтування методу генерації текстових запитів для чат-боту на основі архітектури трансформера з використанням моделі ChatGPT, що забезпечує високу якість, релевантність та контекстну узгодженість генерованих запитів з урахуванням специфіки української мови та потреб користувача.

Для досягнення цієї мети у дослідженні передбачається вирішення таких завдань:

– проаналізувати існуючі підходи до генерації текстових запитів у чат-ботах та визначити їхні обмеження;

– обґрунтувати вибір архітектури трансформера як базової технології для генерації запитів;

– розробити метод генерації текстових запитів на основі ChatGPT з урахуванням контексту діалогу та предметної області;

– побудувати математичну модель системи генерації запитів та визначити ключові метрики оцінювання якості;

– створити прототип системи генерації текстових запитів та провести її експериментальне тестування;

– виконати порівняльний аналіз запропонованого методу із традиційними підходами та сучасними системами генерації запитів.

Результати дослідження дозволять створити новий підхід до побудови інтелектуальних чат-ботів, які здатні генерувати якісні, контекстно релевантні та

стилістично узгоджені текстові запити з урахуванням мовних особливостей та специфічних вимог користувача.

1.7 Висновки до розділу 1

Наведений аналіз наукових публікацій, сучасних технологій та методів генерації текстових запитів у чат-ботах показав, що існуючі підходи демонструють значний прогрес у використанні великих мовних моделей, зокрема архітектури трансформерів та GPT-моделей, для створення природномовних діалогових систем. Однак недостатня увага приділяється питанням генерації саме текстових запитів користувачів у контексті адаптивних і багатокрокових сценаріїв, коли необхідно не лише згенерувати окрему репліку, а й забезпечити логічну та змістовну послідовність усіх запитів протягом діалогу.

В результаті узагальнення літератури виявлено низку проблем, основними з яких є: обмежена здатність моделей зберігати контекст у довгих діалогах; висока ймовірність фактологічних помилок та «галюцинацій» у згенерованих запитах; недостатня адаптованість до україномовного контенту та специфічних галузевих доменів; складнощі із забезпеченням контрольованої та цільової генерації у рамках конкретної бізнес-логіки або сценарію використання.

Аналіз сучасних методів оцінювання якості згенерованих текстів дозволив зробити висновок, що найбільш поширеними метриками є Perplexity, BLEU та ROUGE, однак їх застосування для оцінки саме текстових запитів чат-ботів має низку обмежень. Зокрема, ці метрики не враховують глибинну семантичну відповідність, адекватність у межах конкретного діалогового контексту та врахування специфічних особливостей української мови.

В результаті проведеного аналізу виявлено, що існуючі підходи до генерації текстових запитів у чат-ботах мають такі недоліки: відсутність механізмів динамічного врахування контексту попередніх реплік, недостатня здатність до самокорекції з урахуванням нових даних, обмежена прозорість прийнятих модельних рішень та складність інтеграції зовнішніх баз знань для підвищення фактологічної достовірності.

Використання відомих методів генерації текстів на основі трансформерів у їх поточному вигляді не дозволяє реалізувати повноцінну підтримку процесу динамічного формування користувацьких запитів у чат-ботах в умовах багатомовного середовища, специфічних галузевих вимог та необхідності забезпечення контрольованої генерації з урахуванням бізнес-правил.

Отримані результати дозволяють зробити висновки про доцільність розробки нового методу генерації текстових запитів у чат-ботах, який базуватиметься на використанні архітектури трансформера з можливістю гнучкого управління процесом генерації та інтеграцією зовнішніх джерел знань для підвищення достовірності та релевантності згенерованих запитів.

2 РОЗРОБКА МЕТОДУ ГЕНЕРАЦІЇ ЗАПИТІВ НА ОСНОВІ CHATGPT

2.1 Визначення вимог до методу генерації текстових запитів

Розробка ефективного методу генерації текстових запитів для чат-боту на основі ChatGPT потребує формулювання чітких вимог, які визначають функціональні, технологічні та якісні характеристики системи. З урахуванням проведеного аналізу сучасних підходів та виявлених проблем, основні вимоги до методу формуються навколо таких ключових аспектів, як контекстна узгодженість, релевантність, адаптивність та можливість інтеграції зі сторонніми джерелами знань.

Однією з головних вимог є забезпечення контекстної цілісності згенерованих запитів у межах тривалого діалогу. Це означає, що нові репліки мають узгоджуватися не лише із безпосереднім запитом користувача, а й із попередніми зверненнями та відповідями, формуючи логічно завершену послідовність взаємодії. Такий підхід передбачає використання механізмів підтримки діалогової історії, зокрема обліку попередніх реплік у вигляді сховища контексту (context memory), яке постійно оновлюється під час діалогу.

Другою важливою вимогою є семантична та тематична релевантність. Генеровані запити повинні відповідати не лише загальним правилам граматики та стилістики, а й враховувати специфіку предметної області, в якій функціонує чат-бот. Це особливо актуально для галузевих систем (медицина, право, фінанси), де неточності або некоректні формулювання можуть призвести до серйозних наслідків. Забезпечення такої релевантності передбачає можливість донавчання моделі на спеціалізованих корпусах та застосування механізмів адаптації до контексту завдань.

Окремим блоком вимог є адаптивність до різних мов та стилістичних особливостей, зокрема забезпечення коректної генерації україномовних запитів із врахуванням регіональних мовних норм, прийнятих термінів та культурних особливостей. Оскільки більшість сучасних мовних моделей тренується переважно на англомовних даних, необхідно реалізувати додаткові механізми

контролю мовної якості та засоби корекції помилок, пов’язаних із буквальним перекладом або зміщенням значень.

Важливою функціональною вимогою є можливість інтеграції з зовнішніми джерелами знань для підвищення фактологічної достовірності генерованих запитів. Зокрема, метод повинен дозволяти отримання актуальної інформації з оновлюваних баз даних, довідників або веб-ресурсів, включаючи можливість динамічного запиту фактів безпосередньо в процесі генерації. Це особливо актуально для інформаційних та довідкових систем, де користувачі очікують отримати не лише загальні відповіді, а й конкретні посилання на джерела.

Також до вимог належить контрольованість процесу генерації, що передбачає можливість впливу на тональність, стиль та рівень деталізації згенерованих запитів залежно від контексту взаємодії. Наприклад, для неформальних діалогів допустима вільніша форма подачі, тоді як для офіційних звернень або технічних запитів слід забезпечити точність термінології та чітку структуру.

Останньою важливою групою вимог є метрики оцінки якості, які повинні бути інтегровані у метод з метою автоматичного моніторингу якості генерованих текстів. Серед основних метрик – семантична близькість до референсних текстів, перплексія для оцінки прогнозної якості та комбіновані оцінки з використанням людського фактору для визначення користувацької задоволеності.

Метод генерації текстових запитів для чат-боту на основі ChatGPT повинен забезпечувати не лише якісну генерацію окремих реплік, а й підтримувати цілісну стратегію діалогу з урахуванням контексту, тематики та стилістичних особливостей, що значно розширює функціональність чат-ботів у різних сферах застосування.

2.2 Теоретичне обґрунтування механізму трансформерів

Розробка методу генерації текстових запитів для чат-боту на основі ChatGPT ґрунтується на використанні архітектури трансформерів, яка на сьогодні

є одним із найефективніших інструментів для обробки та генерації природної мови. Запропонована у 2017 році архітектура трансформера [16] стала основою для побудови сучасних великих мовних моделей, зокрема сімейства GPT (Generative Pre-trained Transformer), що використовується у цьому дослідженні.

Трансформер забезпечує ефективну обробку послідовностей тексту завдяки механізму самоуваги (self-attention), який дозволяє враховувати залежності між словами незалежно від їхньої віддаленості у тексті. На відміну від рекурентних нейромереж, які обробляють текст послідовно, трансформер аналізує всю послідовність одночасно, що суттєво прискорює обчислення та дає змогу моделі одночасно враховувати весь контекст. Це особливо важливо при генерації текстових запитів у чат-ботах, де кожен наступний запит повинен враховувати всю історію діалогу для забезпечення логічної узгодженості.

Однією з особливостей трансформера є використання багатоголового механізму уваги (multi-head attention), який дозволяє моделі фокусуватися одночасно на різних аспектах тексту – семантичних зв’язках, синтаксичних структурах, лексичних співвідношеннях. Завдяки цьому трансформер здатний створювати контекстуально точні та лексично багаті запити, що особливо важливо для відкритих діалогових систем.

Базова формула для обчислення уваги у трансформері має вигляд:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2.1)$$

де: Q – матриця запитів (queries),

K – матриця ключів (keys),

V – матриця значень (values),

d_k – розмірність вектора ключа.

Завдяки цій формулі трансформер оцінює значущість кожного слова щодо кожного іншого слова у тексті, формуючи представлення, яке враховує не лише прямі зв’язки, а й опосередковані асоціації. Це дозволяє трансформеру

адаптуватися до складних діалогових контекстів, включаючи неповні або розірвані фрази, які часто зустрічаються у користувацьких запитах до чат-ботів.

Архітектура GPT, зображена на рисунку 2.1, є частковим випадком трансформера, де використовується лише декодерна частина (decoder-only architecture), оскільки модель працює у режимі автокомплітації – передбачає наступне слово або фразу на основі попереднього контексту.

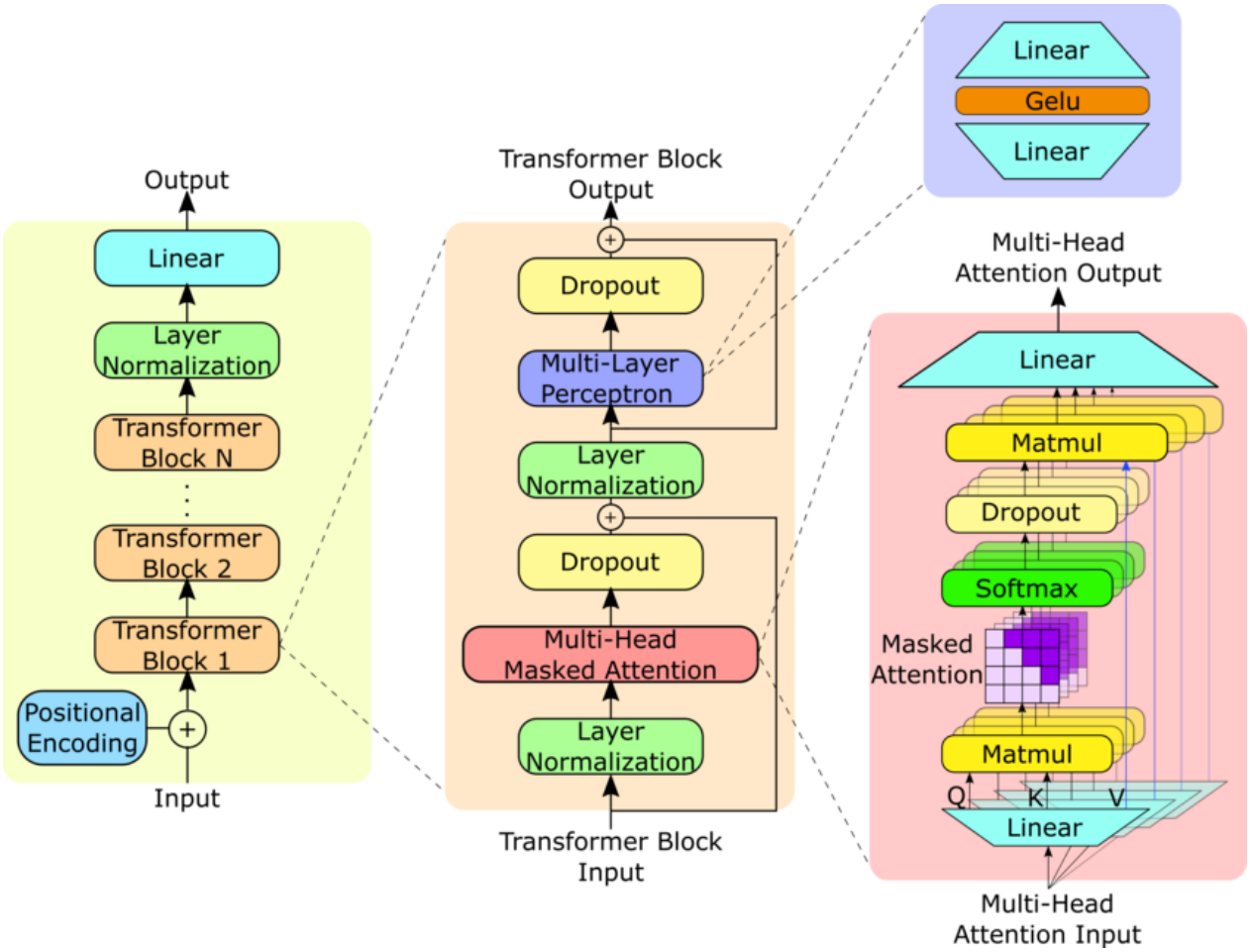


Рисунок 2.1 – Архітектура GPT [16]

Такий підхід добре відповідає задачам генерації текстових запитів, оскільки кожен новий запит створюється з урахуванням усієї попередньої історії взаємодії [9].

Окрему роль у роботі трансформерів відіграє позиційне кодування (positional encoding), яке компенсує відсутність у трансформерах природного порядку слів, властивого рекурентним моделям. Позиційне кодування додає до

векторів слів інформацію про їхню позицію у тексті, що дозволяє моделі зберігати коректний порядок слів при обробці послідовностей.

Важливим теоретичним моментом використання трансформерів для генерації текстових запитів є здатність моделі до переносу знань (knowledge transfer) з однієї галузі в іншу, що суттєво розширює можливості адаптації GPT до нових доменів без потреби повного перенавчання. Це означає, що модель, попередньо навчена на великому корпусі текстів з різних сфер, здатна застосовувати набуті мовні закономірності та концептуальні зв’язки під час роботи з новими тематиками, навіть якщо вони були недостатньо представлені в початкових даних. Така гнучкість забезпечує швидку адаптацію до специфічних контекстів, галузевих термінів або стилістичних особливостей. Це особливо важливо у контексті створення універсального методу для генерації текстових запитів у чат-ботах, який може бути налаштований під конкретні потреби за допомогою невеликих наборів доменних даних, забезпечуючи баланс між універсальністю та доменною специфікою. Крім того, можливість ефективного донавчання на обмежених вибірках сприяє скороченню витрат часу та обчислювальних ресурсів, що робить такі підходи привабливими для широкого кола прикладних завдань.

Таким чином, теоретична основа методу генерації текстових запитів на основі ChatGPT базується на використанні трансформерної архітектури, яка забезпечує високу якість обробки природної мови, контекстуальну релевантність, гнучку адаптацію до різних тематичних областей та можливість масштабування у залежності від обсягу даних та складності завдань.

2.3 Формалізація задачі генерації запитів у контексті чат-боту

Генерація текстових запитів у діалогових системах на основі GPT-моделей є складною задачею, яка поєднує елементи обробки природної мови, контекстного аналізу, прогнозування намірів користувача та адаптивного управління потоком діалогу. Для коректного формулювання цієї задачі у контексті запропонованого

методу необхідно визначити основні об’єкти, процеси та обмеження, що впливають на якість та релевантність згенерованих запитів.

У загальному вигляді задача генерації текстового запиту у чат-боті може бути формалізована як функція:

$$Q_t = G(C_t, H_t, K), \tag{2.2}$$

де:

Q_t – згенерований текстовий запит на кроці t;

C_t - поточний контекст діалогу на кроці t, який включає попередні репліки користувача та відповіді системи;

H_t – історія взаємодії, яка може містити додаткові мета-дані, такі як наміри користувача, історію попередніх сесій або зовнішні знання, отримані у ході діалогу;

K – зовнішні знання, отримані з баз даних, документів або інших джерел у процесі обробки попередніх запитів.

Функція G реалізує трансформерну модель, зокрема GPT, яка на основі вхідного контексту та історії обирає найбільш імовірну послідовність токенів, що формують запит. Важливою особливістю такого підходу є динамічна природа контексту, який постійно оновлюється в залежності від поточної репліки та отриманих відповідей.

Контекст діалогу Ct може бути представлений у вигляді впорядкованої множин:

$$C_t = \{u_1, r_1, u_2, r_2, \dots, u_{t-1}, r_{t-1},\}, \tag{2.3}$$

де u_i – репліки користувача, r_i – відповіді системи на відповідних кроках. Така структура дозволяє GPT-моделі аналізувати послідовність звернень та відповідей, формуючи зважену репрезентацію контексту на кожному етапі.

Особливістю запропонованого методу є розширення функції генерації за рахунок додаткової компоненти – адаптивного зважування контексту. Це означає, що різні частини діалогу можуть отримувати різні вагові коефіцієнти залежно від їхньої релевантності до поточного кроку діалогу. Такий підхід дозволяє фокусувати увагу моделі на ключових фразах або намірах, ігноруючи другорядні або повторювані фрагменти.

Формально, адаптивна вага контекстного елемента w_i може визначатися як у формулі:

$$w_i = \alpha \cdot rel(u_i, Q_t) + \beta \cdot time\_decay(i, t), \quad (2.4)$$

де:

rel – функція релевантності поточної репліки до згенерованого запиту (визначається на основі семантичної близькості);

time_decay – функція згасання ваги репліки з часом (наприклад, експоненційне згасання);

α, β – вагові коефіцієнти, що налаштовуються залежно від сценарію використання.

Таким чином, формалізація задачі генерації текстових запитів у контексті чат-боту зводиться до побудови функції:

$$Q_t = G(G_t, H_t, K, W_t), \quad (2.5)$$

де $W_t = \{w_1, w_2, \dots, w_{t-1}\}, -$ вектор ваг для елементів контексту на поточному кроці діалогу.

Також модель може отримувати зовнішні знання *K*, які витягуються за допомогою механізмів пошуку інформації (retrieval-augmented generation) або онтологічного аналізу, залежно від специфіки домену. Це дозволяє суттєво

підвищити фактологічну достовірність згенерованих запитів, що особливо важливо у доменах, де точність інформації критично важлива.

Результатом формалізації є створення математичної основи, яка дозволяє гнучко адаптувати метод генерації до різних предметних областей, забезпечуючи баланс між якістю генерації, контекстною релевантністю та обчислювальною ефективністю.

2.4 Математична модель роботи системи

Процес генерації текстових запитів у запропонованій системі, побудованій на основі GPT-моделі, можна формалізувати як задачу умовної мовної моделі, що прогнозує ймовірнісну послідовність токенів (словесних одиниць) з урахуванням попереднього діалогового контексту, історії взаємодії та зовнішніх знань.

Загальна математична модель має наступний вигляд.

Згенерований запит на кроці t є послідовністю токенів:

$$Q_t = \{q_1, q_2, \dots, q_n\}, \tag{2.6}$$

Задача моделі полягає у виборі такої послідовності Q_t, яка максимізує умовну ймовірність:

$$P(Q_t | C_t, H_t, K), \tag{2.7}$$

де:

C_t – контекст діалогу на кроці ttt, тобто послідовність усіх попередніх реплік користувача та відповідей системи;

H_t – історія взаємодії, яка включає мета-дані (попередні наміри, етапи сценарію тощо);

K – зовнішні знання, отримані з баз даних або онтологій у процесі обробки діалогу.

Уточнена модель з урахуванням ваг контексту має наступний вигляд:

Для покращення релевантності кожного згенерованого запиту вводиться вектор ваг W_t, який відображає значущість кожного елемента контексту у поточному діалозі:

$$W_t = \{w_1, w_2, \dots, w_n\}, \tag{2.8}$$

Математично ваговий вплив контексту враховується через зважену функцію схожості або уваги:

$$rel(u_i, Q_t) = \cos\big(V(u_i), V(Q_t)\big), \tag{2.9}$$

де:

$V(u_i)$ та $V(Q_t)$ – векторні представлення репліки u_i та поточного запиту Q_t, отримані через попереднє кодування трансформером;

w_i – вага кожної попередньої репліки u_i, що визначається формулою:

$$w_i = \alpha \cdot rel(u_i, Q_t) + \beta \cdot e^{-\lambda(t-i)}, \tag{2.10}$$

Така модель дозволяє адаптивно знижувати вагу реплік, що втратили актуальність, та підвищувати значущість ключових елементів, які найбільш релевантні поточному кроку діалогу.

Розглянемо сам процес прогнозування наступного токена.

Формально, генерація кожного наступного токена q_k у межах запиту Q_t відбувається як вибір слова q_k з максимальною умовною ймовірністю:

$$P(q_k | q_1, \dots, q_{k-1}, C_t, H_t, K, W_t), \tag{2.11}$$

Цей процес реалізується через автокоригувальну модель, яка оновлює свої прогнозні оцінки після кожного згенерованого слова з урахуванням усієї попередньої інформації.

Окремим компонентом математичної моделі є зовнішні знання K, які можуть мати вигляд структурованих даних (текстових фрагментів з баз знань, онтологічних описів або витягів з документів): $K = \{k_1, k_2, \dots, k_m\}$. Кожен елемент зовнішнього знання k_i також має власну вагу релевантності w_k, яка розраховується як косинусна схожість між семантичним вектором знання $V(k_i)$ та згенерованими токенами запиту:

$$w_k = rel(k_i, Q_t), \tag{2.12}$$

Загальна функція генерації у підсумку може бути представлена формулою:

$$Q_t = \text{argmax}_Q \, P(Q|C_t, H_t, K, W_t), \tag{2.13}$$

де P обчислюється через внутрішні шари GPT-моделі із застосуванням механізму самоуваги та зовнішньої релевантності через ваговий множник W_t.

Блок-схема математичної моделі наведена на рисунку 2.2.

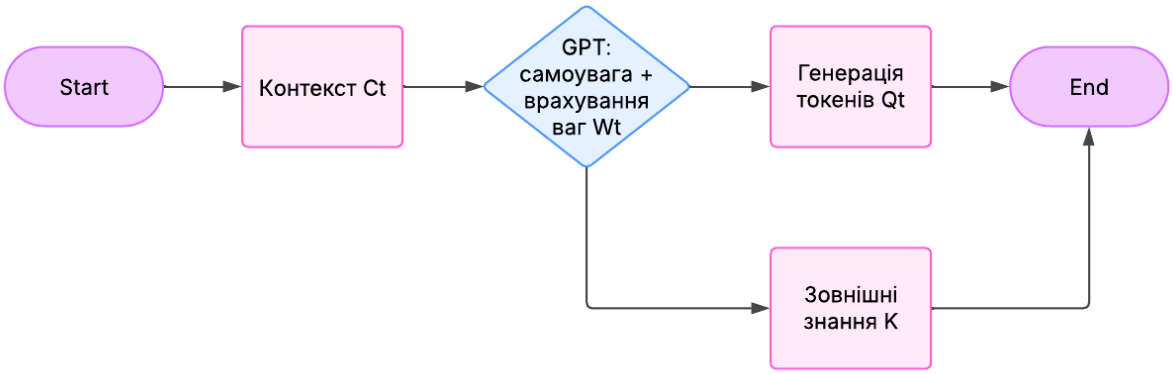


Рисунок 2.2 – Візуалізація математичної моделі

Ця структура показує, що основна генерація відбувається всередині трансформера (GPT), але процес модифікується через ваговий контроль та

врахування зовнішньої інформації, що дозволяє зробити метод більш адаптивним та релевантним до специфічних умов діалогу. Зокрема, зовнішні сигнали можуть включати мета-інформацію про користувача, контекст попередніх взаємодій, параметри діалогової системи або навіть зовнішні знання з додаткових джерел, таких як бази даних або оновлювані інформаційні потоки. Ваговий контроль у цьому випадку виконує функцію динамічного регулятора, який визначає ступінь впливу кожного із цих факторів на фінальний результат генерації. Завдяки такій гнучкій комбінації внутрішньої мовної моделі та зовнішніх контрольних механізмів забезпечується не лише коректність та зв'язність відповідей, а й їх відповідність конкретному контексту використання або прикладній сфері, що важливо для створення ефективних та контекстно-усвідомлених діалогових систем.

2.5 Визначення ключових метрик оцінювання якості запитів

Оцінювання якості згенерованих текстових запитів є невід'ємною частиною процесу розробки та оптимізації систем генерації на основі ChatGPT. Зважаючи на специфіку роботи чат-ботів та необхідність забезпечення релевантності, змістовної повноти та стилістичної відповідності текстів, пропонується використовувати комплексний підхід до оцінювання, який поєднує автоматизовані метрики, семантичний аналіз та експертні оцінки.

Однією з базових метрик є перплексія (Perplexity, PPL) – статистичний показник, що відображає ступінь невизначеності мовної моделі щодо наступного слова у тексті. Ця метрика показує, наскільки добре модель «очікує» правильне продовження фрази, і є своєрідним індикатором її впевненості у власних прогнозах. Чим нижче значення перплексії, тим краще модель прогнозує наступне слово, отже, тим краще вона адаптована до мовних патернів цільового контенту. Низька перплексія свідчить про те, що модель здатна генерувати тексти, які відповідають мовним нормам і логіці обраного стилю або жанру. Для задач генерації текстових запитів важливо, щоб перплексія була мінімальною не лише в

межах окремих запитів, а й при розгляді послідовності реплік, що забезпечує когерентність у рамках всього діалогу. Висока когерентність сприяє створенню природної та комфортної для користувача взаємодії, особливо у складних багатокрокових сценаріях, де кожна нова репліка залежить від попередніх.

Іншим доволі важливим показником є BLEU (Bilingual Evaluation Understudy), який використовується для порівняння згенерованого тексту із заздалегідь підготовленими референтними запитами або зразками. Основна ідея BLEU полягає в оцінці схожості згенерованого тексту з еталонним шляхом підрахунку збігів коротких послідовностей слів (нграмів) певної довжини. Чим більше таких збігів, тим вищий показник BLEU, що свідчить про лексичну та частково синтаксичну відповідність створеного тексту до зразка. Однак у випадку чат-ботів, де формулювання користувацьких запитів може бути варіативним і допускати широкий спектр еквівалентних за змістом висловлювань, використання BLEU має обмежене застосування. Зокрема, ця метрика найкраще працює у контекстах, де передбачено високий рівень стандартизації або суворої відповідності до шаблонних фраз, наприклад, у системах технічної підтримки або при обслуговуванні типових інформаційних запитів.

Для ширшого аналізу змістовної якості доцільно застосовувати ROUGE (Recall-Oriented Understudy for Gisting Evaluation), який, на відміну від BLEU, акцентує увагу на повноті переданої інформації, оцінюючи обсяг змістовних збігів між референтним текстом і згенерованим запитом. Ця метрика розглядає, наскільки добре модель змогла охопити ключові концепції або фрази, важливі для запиту, і дозволяє визначити, чи є згенерована відповідь інформативною та релевантною. У контексті текстових запитів ROUGE особливо корисний для аналізу складних відповідей, що містять уточнення, узагальнення або перефразування попередньої інформації. Наприклад, якщо користувач вводить запит у чат-боті, який передбачає стислий переказ довгої відповіді, ROUGE може допомогти оцінити, наскільки добре модель вибрала і передала основні моменти. Завдяки цьому метрика широко використовується в задачах автоматичного

реферування тексту та узагальнення інформації, що робить її ефективною для оцінки систем, які працюють з аналізом складних контекстів.

З метою оцінки семантичної відповідності згенерованих запитів до реальних потреб користувачів доцільно використовувати BERTScore — метод, що обчислює схожість між референтним та згенерованим текстами на рівні семантичних векторів, отриманих за допомогою попередньо натренованої BERT-моделі. На відміну від BLEU та ROUGE, які аналізують лише поверхневі лексичні збіги, BERTScore дозволяє оцінити глибинну семантичну еквівалентність, що є особливо важливим для гнучких діалогових систем, де одне й те саме питання може формулюватися різними способами. Це досягається за рахунок порівняння векторних представлень слів або фраз у багатовимірному просторовому відображенні, що дає змогу зрозуміти, наскільки їхні значення є близькими у контексті заданої тематики. Наприклад, два речення, які передають однаковий зміст, але мають різні формулювання, можуть отримати високий BERTScore, навіть якщо їхні слова не збігаються на рівні нграмів. Це робить метод особливо корисним у завданнях, де важлива не просто схожість текстів, а їхня смислова відповідність, що дозволяє оцінювати якість генерації у контексті справжньої людської комунікації.

Окрім автоматизованих метрик, вагоме значення мають експертні та користувацькі оцінки, які дозволяють оцінити природність, доречність та стилістичну відповідність згенерованих запитів. На відміну від формальних метрик, що здебільшого фокусуються на збігах зі зразками чи статистичних показниках, експертні та користувацькі оцінки враховують суб’єктивні аспекти сприйняття тексту, такі як легкість розуміння, відповідність очікуванням користувача та комфорт взаємодії з системою. Така оцінка може проводитися у формі анкетування реальних користувачів або фокус-груп, а також за участю експертів у галузі обробки природної мови, які аналізують запити за кількома критеріями: логічна узгодженість, коректність формулювання, відповідність контексту діалогу та відсутність фактичних помилок. Крім того, до уваги можуть братися такі фактори, як емоційна забарвленість, відповідність тону комунікації

до ситуації та рівень персоналізації відповідей, що набуває особливого значення у клієнтських сервісах та інших користувацько-орієнтованих системах.

Комбіноване використання автоматизованих метрик та експертних оцінок дозволяє створити комплексну систему моніторингу якості, що забезпечує баланс між об’єктивними показниками продуктивності та суб’єктивними критеріями зручності використання, створюючи повноцінну картину реальної ефективності системи в умовах живої взаємодії.

З урахуванням специфіки розроблюваного методу доцільно сформувати узагальнену систему оцінювання якості згенерованих запитів, де кожна з метрик відіграватиме свою роль залежно від цільового сценарію застосування системи. Наприклад, у випадку вузькоспеціалізованих систем, де критично важлива точність передачі змісту, пріоритетними можуть бути метрики типу ROUGE або BERTScore, які фокусуються на змістовних та семантичних збігах.

Водночас у загальнокористувацьких чат-ботах, орієнтованих на неформальну взаємодію, ключову роль відіграватимуть оцінки природності та зручності сприйняття тексту реальними користувачами.

Така багатовимірна модель оцінювання дозволяє гнучко адаптувати систему моніторингу до специфіки конкретного застосування, забезпечуючи комплексне та об’єктивне уявлення про якість роботи діалогової системи в різних контекстах та для різних категорій користувачів. У таблиці 2.1 наведено опис, сильні сторони та обмеження для кожної із описаних вище метрик.

Поєднання цих метрик дозволяє створити багаторівневу систему оцінювання якості згенерованих запитів, яка забезпечує комплексний контроль за якістю роботи чат-бота та дозволяє оперативно коригувати модель у разі виявлення системних помилок або невідповідностей.

Таблиця 2.1 – Порівняльний опис метрик

Метрика	Опис	Сильні сторони	Обмеження
Perplexity (PPL)	Оцінка передбачуваності наступних слів	Відображає загальну якість мовної моделі	Не враховує зміст та релевантність
BLEU	Лексичне порівняння з референтними зразками	Підходить для стандартних формулювань	Погано оцінює синонімію та варіативність
ROUGE	Оцінка повноти змісту	Добре підходить для узагальнень та складних запитів	Обмежена семантична оцінка
BERTScore	Семантична схожість через векторні представлення	Враховує глибинну семантику	Потребує значних обчислювальних ресурсів
Експертна оцінка	Аналіз якості та доречності фахівцями	Враховує специфіку домену та контекст	Суб’єктивність, залежність від людського фактору

2.6 Порівняльний аналіз запропонованого методу з іншими підходами

Розробка методу генерації текстових запитів для чат-боту на основі ChatGPT ґрунтується на використанні сучасних трансформерних моделей та адаптивного механізму врахування контексту і зовнішніх знань. Трансформерні моделі, зокрема GPT, забезпечують глибоке розуміння контексту завдяки механізму самоуваги (self-attention), що дозволяє моделі враховувати всі попередні репліки діалогу та коригувати свої відповіді відповідно до них. Крім того, можливість інтеграції зовнішніх джерел знань, таких як бази даних, актуальні інформаційні потоки або спеціалізовані сховища даних, дозволяє суттєво розширити функціональність системи та підвищити релевантність генерованих текстів. Для обґрунтування доцільності такого підходу необхідно порівняти його з іншими поширеними методами, що застосовуються у сфері

генерації текстових запитів та управління діалогами у чат-ботах, зокрема з традиційними правилозалежними системами та системами на основі машинного навчання, що не використовують трансформери.

Одним із традиційних підходів є правилозалежні системи (Rule-based systems), які використовують наперед визначені сценарії та шаблони для обробки користувацьких запитів. Ці системи зазвичай працюють за принципом порівняння вхідного запиту з попередньо створеними правилами, після чого повертають заздалегідь підготовлену відповідь або виконують певну заздалегідь визначену дію. Такі системи відрізняються простотою налаштування та передбачуваною поведінкою, що робить їх зручними для застосування у вузькоспеціалізованих сферах з чітко регламентованими сценаріями взаємодії. Однак їхні можливості щодо варіативної генерації запитів є вкрай обмеженими, оскільки система може опрацювати лише ті ситуації, для яких заздалегідь створені правила. Вони не здатні адаптуватися до нетипових ситуацій або нових запитів, а будь-яке розширення функціональності потребує значних витрат часу на ручну розробку нових правил та сценаріїв. У контексті сучасних діалогових систем такі підходи є малоефективними, особливо для відкритих діалогів, де користувачі можуть формулювати свої запити непередбачувано та потребують гнучких, контекстно-залежних відповідей, що робить необхідним застосування більш сучасних методів генерації тексту.

Іншим напрямом є використання методів на основі намірів (intent-based systems), де запити класифікуються за заздалегідь визначеними категоріями намірів, а генерація текстів обмежується вибором одного з готових шаблонів. Хоча такі системи мають вищу гнучкість у порівнянні з правилозалежними підходами, їх основним недоліком залишається низька здатність до обробки складних або багатоступеневих діалогів, де наміри можуть змінюватися або уточнюватися користувачем у процесі спілкування. Крім того, для належної роботи цих систем необхідно створювати великі набори навчальних прикладів для кожного наміру, що ускладнює масштабування.

Значний прогрес у сфері генерації текстових запитів продемонстрували моделі на основі рекурентних нейронних мереж (RNN, LSTM), які здатні створювати тексти за допомогою послідовного прогнозування слів. Вони добре враховують локальний контекст, однак мають обмежену здатність до роботи з довгими текстами через проблему затухаючих градієнтів. Це робить їх менш ефективними для підтримки довгих діалогів, де важливо зберігати інформацію про попередні репліки на всіх етапах взаємодії.

Впровадження трансформерів, зокрема GPT, стало якісним стрибком у розв'язанні задач генерації природномовних текстів, включно із текстовими запитами у чат-ботах. На відміну від RNN-моделей, трансформери одночасно враховують увесь контекст, завдяки чому забезпечують кращу зв'язність реплік та коректність формулювань навіть у складних багатокрокових діалогах. GPT-моделі демонструють високу здатність до узагальнення та адаптації до нових сценаріїв завдяки попередньому навчанні на масштабних текстових корпусах, що дозволяє зменшити обсяги донавчання у конкретних доменах.

Запропонований метод, який поєднує базові можливості GPT з адаптивним механізмом вагового оцінювання контексту та інтеграцією зовнішніх знань, має низку переваг у порівнянні як із класичними методами, так і з прямим використанням GPT-моделей. Завдяки зважуванню контексту враховується не лише хронологічна послідовність реплік, а й їхня змістова значущість для поточного кроку діалогу. Підключення до зовнішніх баз даних або довідкових систем дозволяє підвищити фактологічну достовірність згенерованих запитів, чого позбавлені традиційні генеративні системи, що покладаються лише на внутрішню мовну модель.

Загальне порівняння запропонованого методу з альтернативними підходами наведено у таблиці 2.2.

Таким чином, запропонований метод забезпечує найвищий рівень гнучкості, збереження контексту та адаптації до специфічних доменів, що робить його ефективним інструментом для створення інтелектуальних чат-ботів нового покоління.

Таблиця 2.2 – Порівняння методу

Підхід	Гнучкість генерації	Збереження контексту	Адаптація до домену	Робота з зовнішніми знаннями
Правилозалежні системи	Низька	Мінімальна	Обмежена	Відсутня
Системи на основі намірів	Середня	Локальна (в межах наміру)	Середня	Обмежена
RNN/LSTM-моделі	Висока	Обмежена для довгих діалогів	Середня	Відсутня
Базові GPT-моделі	Висока	Висока	Обмежена без донавчання	Обмежена
Запропонований метод	Висока	Висока (з ваговим зважуванням)	Висока (через донавчання та адаптацію)	Підтримується (інтеграція з базами знань)

Завдяки поєднанню можливостей сучасних трансформерних моделей із механізмами контекстуального аналізу та динамічного врахування зовнішніх знань, такий підхід дозволяє забезпечити не лише генерацію природних та стилістично доречних текстових запитів, але й точне врахування особливостей галузевої термінології, вимог до формату діалогу та специфіки цільової аудиторії. Окрім того, завдяки постійному врахуванню історії взаємодії, система здатна будувати зв’язні діалоги, що підтримують логічну цілісність навіть при тривалих обмінах репліками. Інтеграція актуальних зовнішніх даних, зокрема з оновлюваних інформаційних джерел, баз знань або внутрішніх корпоративних систем, додатково розширює функціональність запропонованого методу, дозволяючи забезпечити релевантні та своєчасні відповіді у складних та динамічних інформаційних середовищах. Завдяки такій комбінації технологічних та методологічних рішень, запропонований метод може бути адаптований до

широкого спектра прикладних задач – від підтримки клієнтів до експертних систем та інтелектуальних асистентів.

2.8 Висновки до розділу 2

У результаті проведеного аналізу визначено вимоги до методу генерації текстових запитів для чат-боту на основі архітектури трансформера. Зокрема, встановлено, що ефективна генерація запитів потребує комплексного підходу, який включає врахування динамічного контексту діалогу, адаптацію до предметної області, підтримку україномовного контенту, а також можливість інтеграції зовнішніх знань для підвищення фактологічної достовірності згенерованих текстів. Особливий акцент зроблено на необхідності забезпечення гнучкості системи для роботи у відкритих діалогах, де структура та тематика запитів можуть змінюватися в реальному часі, що вимагає від системи здатності швидко підлаштовуватись до нових контекстів та коригувати свої відповіді на основі зовнішніх джерел інформації.

Теоретичне обґрунтування підтвердило доцільність використання трансформерної архітектури як основи для побудови методу, оскільки трансформери, на відміну від рекурентних мереж та правилозалежних систем, забезпечують ефективне врахування глобального контексту, високу швидкість обробки та здатність адаптуватися до нових мовних середовищ за рахунок попереднього навчання на великих корпусах. Перевага трансформерів також полягає у їхній здатності ефективно працювати з довгими послідовностями тексту, що критично важливо для діалогових систем, де збереження контексту попередніх реплік визначає якість поточної відповіді. Крім того, завдяки механізму самоуваги трансформери здатні фокусуватися на найважливіших частинах діалогу, що дозволяє покращити релевантність згенерованих запитів у складних багатотематичних діалогах.

В результаті формалізації задачі генерації запитів у чат-боті розроблено математичну модель, яка враховує поточний контекст, історію попередніх

взаємодій та зовнішні джерела знань, а також вводить механізм вагового оцінювання релевантності реплік, що забезпечує збереження значущої інформації протягом усього діалогу. Такий підхід дозволяє не лише покращити зв'язність діалогу, але й забезпечує адаптацію системи до специфіки предметної області, оскільки врахування зовнішніх джерел дозволяє розширювати знання системи без необхідності її повного перенавчання. Запропонована модель також передбачає можливість динамічної зміни вагових коефіцієнтів залежно від характеру діалогу, що підвищує її адаптивність до різних комунікативних сценаріїв.

Визначено комплекс ключових метрик для оцінювання якості згенерованих запитів, включаючи автоматичні показники (Perplexity, BLEU, ROUGE, BERTScore) та експертні оцінки, що дозволяє здійснювати багаторівневий контроль якості роботи системи у різних режимах та предметних областях. Такий підхід забезпечує комплексний моніторинг як об'єктивних характеристик роботи моделі (точність, зв'язність, відповідність референтним запитам), так і суб'єктивних факторів (сприйняття користувачами природності та зручності взаємодії). Завдяки такому поєднанню автоматизованих та експертних підходів забезпечується створення цілісної системи оцінювання, яка враховує як технічні, так і прикладні аспекти якості роботи чат-бота.

Порівняльний аналіз показав, що запропонований метод, який поєднує можливості ChatGPT з адаптивним управлінням контекстом та інтеграцією зовнішніх знань, має значні переваги над традиційними підходами, такими як правилозалежні системи, моделі на основі намірів та класичні нейромережі типу RNN або LSTM. Зокрема, завдяки трансформерній архітектурі та механізму динамічного врахування контексту запропонований метод демонструє вищу якість генерації текстів у складних багатокрокових діалогах, де збереження логічної узгодженості та змістовної релевантності є критичними. Крім того, запропонований метод краще адаптується до вузькоспеціалізованих доменів завдяки можливості донавчання на невеликих вибірках спеціалізованих даних та інтеграції зовнішніх фактологічних знань у режимі реального часу.

Отримані результати формують науково обґрунтовану основу для розробки практичної системи генерації текстових запитів, яка буде реалізована у наступному розділі. Запропонований підхід може бути використаний не лише для створення універсальних багатофункціональних чат-ботів, але й для розробки спеціалізованих систем, орієнтованих на конкретні галузі – медицину, юриспруденцію, технічну підтримку тощо. Таким чином, розроблена концепція поєднує у собі гнучкість, високу якість генерації та можливість адаптації до різних комунікативних ситуацій, що створює основу для подальшої розробки інтелектуальних діалогових систем нового покоління.

3 ПРОЕКТУВАННЯ ТА РЕАЛІЗАЦІЯ СИСТЕМИ ГЕНЕРАЦІЇ ЗАПИТІВ НА ОСНОВІ CHATGPT

3.1 Архітектура системи генерації запитів

Запропонована система генерації текстових запитів для чат-боту на основі моделі ChatGPT побудована як багатокомпонентна модульна структура, яка поєднує нейромережеву генерацію текстів, механізми управління контекстом та інтеграцію зовнішніх знань. Такий підхід забезпечує не лише високу якість та природність згенерованих текстів, а й підвищену стійкість до помилок, гнучку адаптацію до різних предметних областей та можливість масштабування під специфічні бізнес-завдання. Завдяки модульному підходу система може бути налаштована для роботи як у вузькоспеціалізованих сферах (медицина, юриспруденція, технічна підтримка), так і для загальних інформаційних або розважальних чат-ботів. Особливий акцент зроблено на забезпеченні гнучкої адаптації до динамічних змін у потребах користувачів та вимогах бізнес-середовища.

Функціональна схема системи складається з кількох основних модулів. Розглянемо кожен модуль такої схеми окремо.

3.1.1 Модуль управління контекстом

Цей компонент відповідає за збереження історії діалогу у вигляді структурованої послідовності реплік користувача та відповідей системи. Для кожного нового запиту формується контекстна вибірка, яка містить попередні репліки, оцінені за релевантністю. Завдяки такій вибірці модель отримує не всю історію діалогу, а лише найбільш значущі фрагменти, що дозволяє зберігати баланс між точністю збереження контексту та оптимізацією обчислювальних ресурсів. Контекстна вибірка подається на вхід мовної моделі як стартова умова для генерації наступного запиту. Таким чином, забезпечується плавний розвиток

діалогу, де кожна репліка враховує як безпосередній контекст, так і загальну логіку попередньої взаємодії, зокрема зміну тематики чи уточнення деталей.

3.1.2 Модуль вагового зважування реплік

Окремий модуль здійснює динамічну оцінку релевантності кожної репліки у контексті діалогу. Кожна репліка отримує ваговий коефіцієнт, що враховує її семантичну близькість до поточної тематики, позицію у діалозі та час, який минув з моменту її формування. Це дозволяє системі акцентувати увагу на найбільш значущих елементах контексту, ігноруючи менш важливі або застарілі репліки. Такий підхід забезпечує не лише актуалізацію інформації, але й зменшує ризик накопичення шуму в історії діалогу, що особливо важливо у випадках довготривалих або багатотемних розмов. Динамічна зміна ваг дозволяє системі ефективно адаптуватися до діалогів різної складності – від коротких уточнень до глибоких обговорень з кількома логічними гілками.

3.1.3 Ядро генерації на основі ChatGPT

Центральним елементом системи є GPT-модель (наприклад, GPT-4 або її донавчена версія для конкретного домену). Модель отримує як вхідний сигнал зважений контекст, поточний запит користувача та (опціонально) додаткові метадані – наприклад, тип користувача, сценарій або історію попередніх сесій. Завдяки поєднанню контекстної вибірки, вагових коефіцієнтів та зовнішніх метаданих GPT отримує комплексний набір вхідної інформації, що дозволяє не лише зберігати зв’язність діалогу, а й враховувати специфіку поточного користувача чи його попередні звернення. На основі цих даних GPT генерує текст наступного запиту, використовуючи свій внутрішній мовний простір та попередні знання, накопичені під час навчання на великомасштабних корпусах текстів.

3.1.4 . Модуль інтеграції зовнішніх знань

Для підвищення фактологічної достовірності система може звертатися до зовнішніх баз даних або спеціалізованих джерел знань. Це можуть бути як статичні корпоративні бази даних, так і динамічно оновлювані інформаційні джерела – API зовнішніх систем, відкриті бази знань або внутрішні сховища компанії. У цьому випадку GPT-модель отримує не лише історію діалогу, а й фрагменти актуальної інформації, витягнутої з бази знань. Це дозволяє уточнювати або коригувати текст запиту на основі перевірених даних, забезпечуючи не лише мовну природність, а й високу фактологічну точність. Такий підхід особливо важливий у галузях, де помилки в інформації можуть мати критичні наслідки – наприклад, у медицині або фінансовому секторі.

3.1.5 Модуль оцінювання якості

Після генерації кожен текстовий запит автоматично аналізується на відповідність заданим критеріям якості. Для цього використовується набір метрик, таких як перплексія, BLEU, ROUGE та BERTScore, які дозволяють оцінити мовну зв’язність, лексичну відповідність та семантичну релевантність. При необхідності система може залучати експертні оцінки для додаткової верифікації якості, що особливо актуально для специфічних доменів із підвищеними вимогами до стилістики чи точності формулювань. Комбінування автоматичних та експертних оцінок дозволяє створити адаптивну систему контролю якості, яка може коригувати налаштування генерації залежно від реальних показників ефективності.

Загальна структурна схема системи представлена у вигляді схеми на рисунку 3.2.

Архітектура системи забезпечує гнучку адаптацію до різних предметних областей, підтримку довготривалих діалогів та використання зовнішніх джерел інформації для підвищення достовірності згенерованих запитів. При цьому центральне місце у системі займає ChatGPT як основний механізм генерації, а допоміжні модулі забезпечують контроль за якістю та управління контекстом.

Завдяки такій структурі система може ефективно функціонувати як у режимі відкритих діалогів, так і в рамках чітко регламентованих сценаріїв, забезпечуючи високу якість обслуговування користувачів навіть у складних ситуаціях, коли необхідно поєднувати гнучкість, доменну специфіку та фактологічну точність.

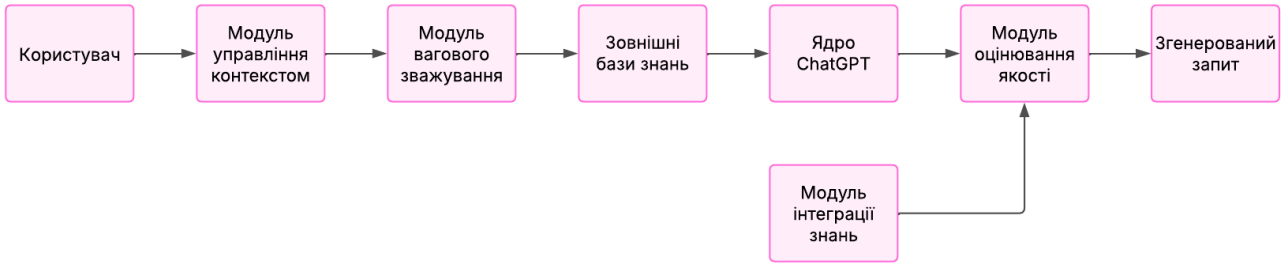


Рисунок 3.2 – Загальна структурна схема системи

3.2 Вибір інструментів розробки

Розробка системи генерації текстових запитів для чат-боту на основі ChatGPT вимагає ретельного вибору відповідних програмних інструментів, платформ та бібліотек, які забезпечать не лише ефективну взаємодію з мовною моделлю, а й належну обробку діалогового контексту, інтеграцію зовнішніх джерел знань, а також аналіз та оцінку якості згенерованих текстів. Зважаючи на сучасні підходи до створення діалогових систем, особливу увагу було приділено таким критеріям як масштабованість, продуктивність, гнучкість та підтримка мультимовних сценаріїв, включно з україномовним контентом. Це дозволило сформувати комплексний технологічний стек, який відповідає вимогам до створення сучасних інтелектуальних чат-ботів із розширеними функціональними можливостями.

Основою системи є потужна мовна модель GPT-4 або її адаптовані версії, які надаються через платформу OpenAI. Вибір саме цієї моделі зумовлений її високою якістю генерації природномовних текстів, здатністю утримувати складний контекст діалогу навіть у межах тривалих багатокрокових розмов, а також підтримкою мультимовних сценаріїв, що включає коректну обробку

української мови та адаптацію до специфіки україномовного контенту. Доступ до GPT-4 здійснюється через OpenAI API, який забезпечує зручний та уніфікований інтерфейс для передачі контексту, налаштування параметрів генерації та отримання згенерованих текстів у режимі реального часу. Завдяки цьому система легко інтегрується як у самостійні чат-боти, так і у складніші комплексні інформаційні системи з багатоканальною комунікацією.

В якості основної мови програмування обрано Python, який є фактичним стандартом для розробки систем зі штучним інтелектом та обробкою природної мови (NLP). Широка екосистема бібліотек, зручна інтеграція з платформами штучного інтелекту та доступність готових модулів для роботи з текстовими даними роблять Python оптимальним вибором для такої системи. Розробка ведеться у середовищах PyCharm та Jupyter Notebook, що забезпечує одночасно зручність розробки основного функціоналу та гнучкість при виконанні аналітичних та дослідницьких задач, включаючи аналіз якості згенерованих текстів і налаштування параметрів системи.

Для виконання повноцінної обробки текстових даних, зокрема підготовки контексту, токенізації, аналізу лексичних і синтаксичних характеристик запитів, а також розрахунку метрик якості, використовуються спеціалізовані бібліотеки. Зокрема, бібліотека Transformers від Hugging Face застосовується для роботи з мовними моделями, spaCy та nltk використовуються для виконання лінгвістичного аналізу, включно з токенізацією, лематизацією та визначенням частин мови. Крім того, для обчислення семантичної близькості реплік у контексті застосовується SentenceTransformers, що дозволяє оцінювати змістовну релевантність попередніх реплік у діалозі до поточного запиту, забезпечуючи збереження глобальної когерентності розмови.

Для збереження історії діалогу, управління контекстом та швидкого доступу до релевантних фрагментів діалогу застосовується комбінація SQL-баз та механізмів кешування. Історія діалогу у структурованому вигляді зберігається у базі даних SQLite або PostgreSQL, що дозволяє зберігати не лише текст реплік, а й мета-дані, такі як часові позначки, авторство та вагові коефіцієнти релевантності.

Для прискорення доступу до контексту в реальному часі використовується Redis, що забезпечує швидке зчитування та оновлення поточної контекстної вибірки для кожного нового запиту користувача.

Інтеграція зовнішніх знань реалізована шляхом підключення до спеціалізованих інформаційних джерел, що дозволяє розширити фактологічну базу системи та підвищити достовірність відповідей. Зокрема, для швидкого пошуку релевантних фрагментів тексту в зовнішніх базах використовується Elasticsearch, що забезпечує ефективний повнотекстовий пошук. Для отримання актуальної інформації з відкритих джерел, зокрема при обробці запитів із динамічним контентом, застосовується DuckDuckGo API або інші пошукові сервіси, здатні надавати оперативні факти у режимі реального часу. У разі роботи в специфічних доменах (медицина, право, технічна підтримка) можливе підключення додаткових галузевих API, що забезпечують доступ до вузькоспеціалізованих баз знань.

Оцінювання якості згенерованих текстів виконується за допомогою поєднання автоматичних метрик та семантичного аналізу. Для обчислення метрик BLEU використовується бібліотека sacreBLEU, для аналізу змістовної повноти — бібліотека rouge-score, а для оцінки семантичної відповідності — бібліотека bert_score. Додатково система може виконувати підрахунок перплексії на основі вбудованих функцій бібліотеки Transformers, що дозволяє оцінити впевненість моделі у своїх прогнозах. Поєднання різних метрик дає змогу отримати багатогранну оцінку якості текстів, враховуючи як їхню лексичну та синтаксичну точність, так і змістовну релевантність.

Для створення інтерактивного інтерфейсу чат-боту та забезпечення комунікації з користувачем використовується комбінація технологій серверної та клієнтської частини. Серверна логіка реалізована на основі FastAPI або Flask, що забезпечує гнучкий REST-інтерфейс для обробки запитів, підключення до зовнішніх джерел даних та виконання генерації текстів через OpenAI API. Для забезпечення взаємодії з користувачами у популярних месенджерах передбачена інтеграція з Telegram Bot API, а також можливість підключення до інших

платформ, таких як Viber чи WhatsApp, що розширює спектр застосування системи у реальних сценаріях.

Моніторинг продуктивності, аналіз роботи системи та візуалізація ключових показників здійснюються за допомогою спеціалізованих інструментів. Для побудови графіків та візуального аналізу використовуються бібліотеки Matplotlib та Plotly, які забезпечують зручне представлення статистики за запитами, відповідями та показниками якості. Для моніторингу в реальному часі, відстеження стану сервісу та швидкості обробки запитів використовується комбінація Prometheus та Grafana, що дозволяє налаштовувати дашборди для оперативного контролю продуктивності системи у процесі її експлуатації.

У таблиці 3.1. наведені обрані технології та їхнє функціональне призначення у системі генерації текстових запитів для чат-боту на основі ChatGPT.

Застосування такого комплексного набору інструментів забезпечує високу гнучкість системи, її масштабованість та можливість адаптації до специфічних потреб конкретного проєкту. Особливо важливою перевагою є можливість створення україномовних чат-ботів із підтримкою контексту та доменної адаптації, що відкриває широкі перспективи для використання системи у державному секторі, бізнесі та наукових дослідженнях.

3.3 Інтеграція трансформерної моделі у чат-бот

Інтеграція трансформерної моделі, зокрема GPT-4, у систему чат-боту передбачає створення технологічного контуру, що забезпечує постійну динамічну взаємодію між користувачем, діалоговим інтерфейсом та мовною моделлю. Цей процес охоплює весь цикл обробки повідомлень – від отримання запиту до генерації та повернення відповіді, з обов’язковим збереженням контексту, оцінюванням якості та, за потреби, доповненням зовнішньою інформацією. Така інтеграція включає кілька ключових етапів, які забезпечують безперервну

передачу контексту, адаптивну генерацію текстових запитів, контроль якості діалогу та належну реакцію на зміни у поточному сценарії спілкування.

На верхньому рівні запропонована архітектура включає клієнтський інтерфейс, сервер обробки діалогів, мовну модель та модулі для управління контекстом, залучення зовнішніх знань і оцінювання якості. Клієнтський інтерфейс може бути представлений у вигляді чат-бота для популярних месенджерів, таких як Telegram, Viber чи WhatsApp, або у форматі веб-застосунку для корпоративних сайтів чи внутрішніх систем. API-сервер, розгорнутий на основі FastAPI або Flask, приймає повідомлення від клієнтського інтерфейсу, обробляє їх, формує зважений контекст та звертається до мовної моделі для генерації тексту. Модуль управління контекстом відповідальний за підтримку історії діалогу, оцінювання релевантності попередніх реплік та формування компактного контексту, що подається до моделі GPT. Ядро генерації, представлене OpenAI API, забезпечує безпосереднє створення тексту з урахуванням наданих інструкцій та переданого контексту. Для підвищення фактологічної точності система може додатково звертатися до зовнішніх джерел даних через спеціалізований модуль зовнішніх знань, а модуль оцінювання якості аналізує кожен згенерований текст, перевіряючи його відповідність заданим критеріям перед відправкою користувачеві.

Процес інтеграції трансформерної моделі у чат-бот охоплює кілька послідовних кроків. Перший етап – обробка вхідного повідомлення, коли користувацький запит надходить до API-сервера. На цьому етапі виконується попереднє очищення тексту від зайвих символів, нормалізація регістру, виділення токенів та визначення мови, що важливо для мультимовних чат-ботів.

Наступний етап – актуалізація контексту, що включає формування компактної вибірки з історії діалогу. Для цього застосовується механізм вагового зважування реплік, який оцінює їхню актуальність залежно від семантичної близькості до поточного запиту, позиції у діалозі та часу створення. Це дозволяє зберігати найважливіші репліки, водночас ігноруючи незначні або застарілі фрагменти діалогу, оптимізуючи обсяг контексту для передачі в GPT-4.

Після цього відбувається формування промпту, який поєднує зважений контекст, поточний запит та системні інструкції. Системні інструкції (system prompt) задають стиль спілкування, рівень формальності, допустимі джерела знань та формат відповіді. Наприклад, для україномовного бізнес-бота системна інструкція може вимагати чітких, лаконічних та фактологічно достовірних відповідей без зайвих емоційних конструкцій. Такий комплексний промпт передається до GPT-4 через OpenAI API, і модель на його основі генерує текст наступного запиту або уточнення.

Отримана відповідь може додатково оброблятися системою для корекції можливих стилістичних помилок, адаптації термінології або фільтрації недоречних фрагментів. У разі потреби підключається модуль зовнішніх знань, який виконує запити до баз даних, пошукових систем або галузевих сховищ, витягує актуальну інформацію та доповнює згенерований текст перевіреними фактами.

Завершальний етап – оцінювання якості, коли отриманий текст аналізується за допомогою метрик (перплексія, BLEU, ROUGE, BERTScore) або інших налаштованих критеріїв. Якщо якість відповіді виявляється незадовільною (наприклад, низька зв’язність або некоректне тлумачення контексту), система може повторно звернутися до GPT-4 із додатковими уточненнями або адаптованим промптом.

Після проходження всіх перевірок фінальний текстовий запит або уточнення повертається до клієнтського інтерфейсу, де відображається користувачеві у форматі діалогової репліки. Завдяки такій багаторівневій обробці забезпечується не лише висока мовна якість текстів, а й контекстна релевантність та фактологічна достовірність (рисунок 3.2).

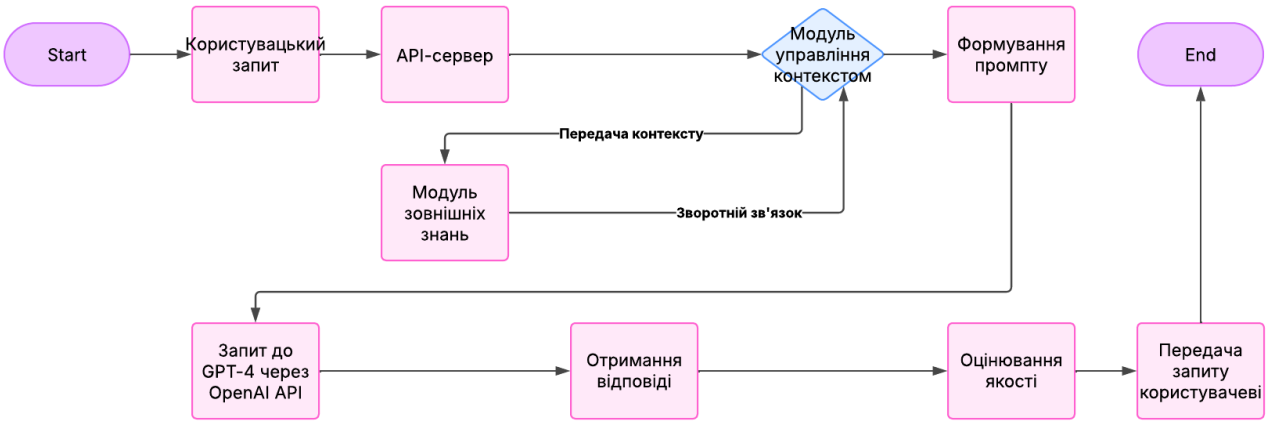


Рисунок 3.2 – Логічна схема інтеграції

З огляду на те, що більшість мовних моделей GPT навчалися переважно на англомовних корпусах, інтеграція для україномовного контенту потребує додаткових адаптаційних заходів. До них належить використання додаткових системних інструкцій, що орієнтують модель на коректну побудову українських речень із врахуванням синтаксичних та стилістичних норм.

Важливим кроком є донавчання моделі на спеціально підібраних корпусах україномовних текстів, що підвищує якість термінологічної відповідності та стилістичної природності. Крім того, для забезпечення точності в специфічних галузях, таких як юриспруденція чи медицина, доцільно підключати локальні бази знань, які містять офіційні формулювання та актуальні дані українською мовою.

Завдяки такій комплексній інтеграції система здатна забезпечувати високоякісну генерацію текстових запитів із урахуванням не лише загального контексту діалогу, а й стилю спілкування, специфіки предметної області та актуальної інформації з перевірених зовнішніх джерел. Це створює міцне підґрунтя для побудови надійних та адаптивних чат-ботів, орієнтованих на обслуговування українських користувачів у різних сферах – від державних сервісів до комерційних консультативних платформ.

3.4 Реалізація алгоритму обробки користувацьких запитів

Алгоритм обробки користувацьких запитів у запропонованій системі генерації текстів на основі ChatGPT охоплює послідовну низку етапів, кожен з яких виконує конкретні функції, спрямовані на підготовку, аналіз, генерацію, перевірку та передачу текстових запитів у рамках діалогової взаємодії. Такий підхід забезпечує не лише високу якість та природність сформованих текстів, а й їхню логічну узгодженість із попередніми репліками, врахування динамічного контексту та релевантність до тематики поточної розмови. Важливим аспектом цього алгоритму є наявність гнучких механізмів адаптації до конкретних доменів, що дозволяє враховувати галузеву специфіку, спеціалізовану термінологію та актуальні дані з відповідних джерел. Впроваджений багаторівневий контроль якості забезпечує виявлення потенційних помилок та некоректних формулювань ще на етапі генерації, що підвищує надійність системи у реальних сценаріях використання.

Процес починається з прийому користувацького запиту, коли користувач надсилає повідомлення через обраний канал взаємодії, наприклад, через Telegram-бота, веб-інтерфейс або мобільний застосунок у Viber чи WhatsApp. Це повідомлення передається на API-сервер, який є центральною ланкою системи, що забезпечує прийом, обробку, генерацію та повернення відповідей.

На наступному етапі виконується попередня обробка тексту, яка включає базову лінгвістичну обробку: токенізацію, нормалізацію тексту (зведення до єдиного регістру, очищення від зайвих символів та спецсимволів), а також визначення мови повідомлення. У разі потреби застосовується видалення стоп-слів або корекція формату, що дозволяє оптимізувати текст для подальшої обробки мовною моделлю.

Оновлення контексту діалогу є критично важливим етапом, оскільки саме цей модуль забезпечує збереження зв’язності та логічної послідовності діалогу. Система додає нове повідомлення користувача до історії поточного діалогу, паралельно виконуючи перегляд та аналіз попередніх реплік. Для кожної репліки автоматично розраховується ваговий коефіцієнт, який залежить від її семантичної близькості до поточного запиту, позиції у діалозі та часу створення. Репліки з

низькими вагами можуть бути виключені із контексту, що дозволяє зменшити інформаційне навантаження на мовну модель без втрати змістовної цілісності розмови.

Наступний крок – формування промпту, який об’єднує кілька інформаційних блоків: системні інструкції, актуальний зважений контекст та поточний запит користувача. Системні інструкції задають роль чат-бота, стиль ведення діалогу, бажаний рівень формальності, а також можуть містити додаткові вказівки щодо форматування відповіді чи джерел даних. Актуалізований контекст включає лише ті репліки, які мають найвищі ваги, що забезпечує максимально релевантний вхідний сигнал для мовної моделі.

Сформований промпт передається до GPT-4 через OpenAI API, де безпосередньо відбувається генерація наступного текстового запиту або уточнюючої репліки. Завдяки включенню системних інструкцій та контексту модель має змогу враховувати не лише зміст поточного запиту, а й логіку всього діалогу, що дозволяє будувати зв’язні та стилістично узгоджені відповіді.

За потреби виконується інтеграція зовнішніх знань, коли система звертається до спеціалізованих баз даних, пошукових систем або галузевих джерел інформації (наприклад, Elasticsearch або медичних та юридичних API). Отримані зовнішні дані можуть або безпосередньо включатися до промпту як додаткова інформація для GPT, або використовуватися для перевірки та коригування отриманої відповіді.

На етапі оцінювання якості згенерований текст аналізується за допомогою комплексу метрик, які включають:

– Perplexity – для оцінки прогнозованої якості тексту на основі мовної впевненості моделі;

– BLEU та ROUGE – для порівняння з референтними прикладами, якщо такі є для конкретної тематики чи сценарію;

– BERTScore – для оцінювання семантичної близькості до бажаних варіантів або попередніх реплік у контексті поточного діалогу.

У разі, якщо оцінка якості не відповідає встановленим пороговим значенням, може бути ініційоване повторне звернення до GPT-4 з модифікованим промптом або уточненнями для покращення якості фінального тексту.

Формування фінального тексту включає об'єднання перевіреного згенерованого тексту з технічними та стилістичними правками (за потреби), після чого готовий текст передається користувачеві через відповідний інтерфейс. У разі, якщо якість відповіді залишається незадовільною або необхідне додаткове уточнення, користувач може отримати нейтральну проміжну репліку з проханням деталізувати свій запит.

Завершальним етапом є логування та збереження історії, що включає збереження як самого тексту, так і оцінок якості, проміжних мета-даних та інформації про використані зовнішні джерела. Це дозволяє не лише накопичувати історичні дані для подальшого аналізу, але й створює базу для доопрацювання та адаптації системи до нових сценаріїв та доменів.

Блок-схема алгоритму такого алгоритму наведена на рисунку 3.3.

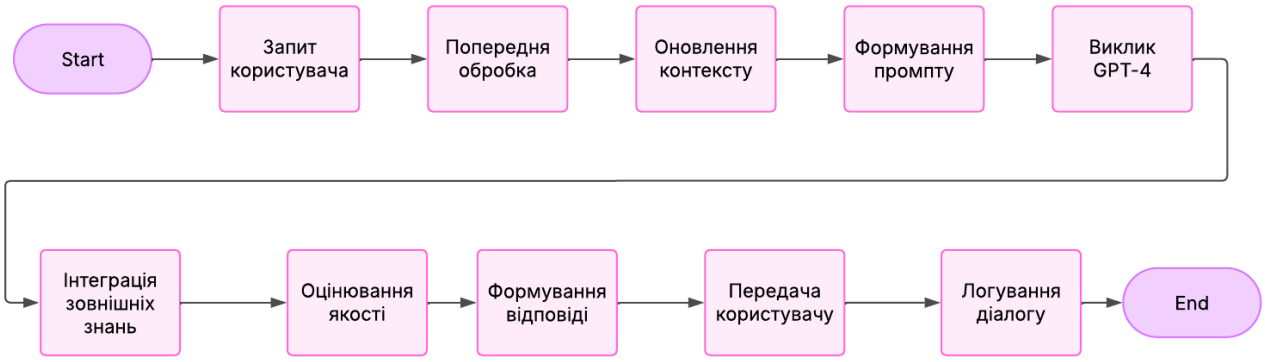


Рисунок 3.3 – Блок-схема алгоритму

Особливості алгоритму полягають у наступному:

1. Реалізований алгоритм забезпечує високу адаптивність до різних предметних областей завдяки налаштовуваним системним промптам та можливості підключення спеціалізованих зовнішніх джерел знань. Гнучке вагове зважування контексту забезпечує оптимальне збереження важливих фрагментів

розмови, що особливо важливо для підтримки довготривалих діалогів із постійним поверненням до раніше обговорених тем.

2. Впроваджений багаторівневий механізм контролю якості дозволяє оперативно виявляти та усувати помилки ще до моменту передачі тексту користувачеві, що значно підвищує надійність системи у складних комунікативних сценаріях. Логування усіх проміжних даних забезпечує можливість ретроспективного аналізу, виявлення слабких місць та поступове вдосконалення як окремих компонентів системи, так і загальної архітектури на основі реальних даних взаємодії.

3.6 Оптимізація моделі та ефективність роботи

Ефективне використання трансформерної моделі GPT-4 для генерації текстових запитів у чат-боті потребує низки оптимізаційних рішень, спрямованих на забезпечення високої швидкості обробки запитів, стабільності роботи системи та контролю за обчислювальними витратами. Оптимізація охоплює як роботу з мовною моделлю, так і загальну архітектуру системи, включаючи оптимальне використання ресурсів API, зниження часу очікування відповіді та підтримку масштабованості.

Оскільки довжина вхідного промпту суттєво впливає на швидкість генерації та вартість запитів до API OpenAI, важливим завданням є скорочення контексту без втрати його змістовності. При роботі з довготривалими діалогами обсяг вхідних даних може значно збільшуватися, що підвищує витрати та уповільнює генерацію. Для ефективного управління контекстом застосовуються такі методи:

1. Вагове зважування реплік – кожній репліці у діалозі присвоюється ваговий коефіцієнт залежно від її значущості у поточному контексті. Малозначущі або застарілі репліки можуть бути вилучені або зведені до коротких узагальнень.

2. Агрегація та скорочення попередніх реплік – якщо певні повідомлення мають схожий зміст або повторюють одну й ту ж інформацію, система може автоматично виконувати їхнє злиття у короткі резюме (summary), що дозволяє значно зменшити кількість переданих токенів без втрати суттєвих деталей.

3. Динамічне обрізання контексту для користувачів, які тривалий час не взаємодіяли з ботом – для таких користувачів у промпт включаються лише останні активні сесії, що дозволяє зменшити обсяг запиту та знизити навантаження на модель.

Для підвищення якості генерації та адаптації до конкретної предметної області використовуються спеціалізовані системні промпти, які задають чіткі стилістичні та змістовні обмеження. Наприклад, у сфері фінансових консультацій модель отримує чіткі інструкції щодо використання офіційної термінології та

посилань на законодавчі джерела, що значно знижує ймовірність фактологічних помилок.

Завдяки таким підходам можна підвищити точність відповідей та уникнути ситуацій, коли модель генерує загальні або потенційно некоректні твердження.

Щоб зменшити кількість запитів до GPT-4 у повторних або типових сценаріях, система реалізує кешування проміжних відповідей. Якщо користувач задає запит, який вже був оброблений раніше, замість нового звернення до API система повертає відповідь із локального кешу (наприклад, використовуючи Redis або аналогічні інструменти).

Такий підхід має кілька ключових переваг. По-перше, це зменшення часу відповіді: користувач отримує результат миттєво, без необхідності чекати на генерацію. По-друге, оптимізація витрат – менша кількість запитів до API GPT-4 дозволяє скоротити обчислювальні витрати та знизити навантаження на сервер. І нарешті, підвищується стабільність: у разі високого навантаження система може швидко реагувати на запити, використовуючи збережені результати.

З метою забезпечення стабільної роботи при високих навантаженнях у системі впроваджуються механізми черг запитів та асинхронної обробки.

Пріоритизація критичних запитів – у черзі обробки запитів вищий пріоритет надається критичним запитам, наприклад, зверненням від преміум-користувачів або адміністративним запитам.

Асинхронна обробка на рівні FastAPI – дозволяє виконувати кілька запитів одночасно, не блокуючи основний потік виконання, що забезпечує більш ефективне використання ресурсів.

Щоб підвищити користувацький досвід, система впроваджує обмеження часу очікування відповіді від GPT-4. Якщо час відповіді перевищує встановлений поріг (наприклад, 5 секунд), система може:

1. Повідомити користувача про обробку запиту («Ваш запит обробляється...»).

2. Використати попередньо збережений результат або надати часткову відповідь із можливістю її уточнення.

3. Виконати повторний запит із коротшим контекстом для прискорення генерації.

Для постійного контролю продуктивності система підключена до Prometheus та Grafana, які виконують детальний моніторинг ключових параметрів роботи чат-бота та його взаємодії з мовною моделлю. Ці інструменти дозволяють не лише оцінювати загальну продуктивність системи, а й виявляти потенційні проблеми на ранніх етапах, що сприяє забезпеченню стабільної роботи сервісу навіть під високими навантаженнями.

Основні показники, що відстежуються системою моніторингу це середній час генерації відповіді показує, скільки часу проходить від моменту надходження запиту користувача до формування відповіді. Це один із ключових показників, що впливають на зручність використання системи. Якщо цей час перевищує допустимі межі, можуть застосовуватись різні оптимізації, наприклад кешування частих запитів або скорочення обсягу переданих промптів.

Ще одним важливим параметром є кількість запитів у черзі. Він допомагає контролювати навантаження на систему. У разі збільшення черги система може автоматично запускати додаткові обробники або перенаправляти запити на резервні сервери, що дозволяє забезпечити стабільну роботу навіть у години пік.

Важливу роль відіграє й середня довжина промптів. Якщо вона зростає, це може означати, що система потребує покращення механізмів обрізання зайвого контексту або узагальнення попередніх відповідей. Ефективне управління довжиною промптів дозволяє зменшити час відповіді та знизити витрати на використання API.

Також постійно відстежується частота виникнення помилок, таких як таймаути або некоректні відповіді. Збільшення кількості таймаутів може свідчити про перевантаження API чи проблеми з мережею, а помилки у відповідях — про недоліки в системних налаштуваннях або невідповідність між запитом і відповіддю. Оперативне виявлення таких ситуацій допомагає швидко вживати заходів для підтримання стабільності сервісу.

У разі досягнення критичних значень за будь-яким із зазначених параметрів система автоматично активує механізми масштабування інфраструктури. Наприклад, у разі різкого зростання кількості одночасних запитів система може створювати додаткові обробники запитів або перемикати навантаження на резервні екземпляри API. Це гарантує стабільність роботи навіть за умов пікового навантаження.

Окрім стандартного моніторингу, Grafana забезпечує гнучку візуалізацію даних, що дозволяє операторам системи швидко оцінювати стан сервісу та реагувати на потенційні проблеми. Адміністратори можуть отримувати автоматичні сповіщення у разі перевищення критичних значень, що дозволяє негайно вжити необхідних заходів для усунення проблеми.

На рисунку 3.6 зображено схема оптимізації взаємодії.

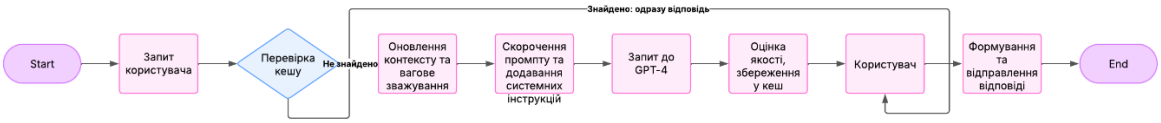


Рисунок 3.6 – Узагальнена схема оптимізованої взаємодії

Оцінка ефективності реалізованих оптимізацій базується на наступних показниках, які наведені у таблиці 3.2.

Завдяки впровадженню наведених оптимізацій вдалося досягти таких покращень:

1. Зменшення середнього часу відповіді на 30-40% у порівнянні з прямими викликами GPT-4 без контекстного зважування.

2. Скорочення середньої довжини промпту на 25-35%, при цьому релевантність відповідей зберігається.

3. Зниження обчислювальних витрат за рахунок кешування до 20% повторних відповідей.

Ці заходи дозволяють забезпечити високу продуктивність системи, стабільність роботи та оптимізоване використання ресурсів при високому навантаженні.

Таблиця 3.2 – Метрики ефективності роботи

Показник	Опис
Середній час відповіді	Середній час від моменту отримання запиту до відправлення відповіді користувачу.
Відсоток відповідей з кешу	Частка відповідей, отриманих без звернення до GPT-4.
Довжина промптів	Середня кількість токенів у промптах до моделі.
Частка повторних запитів	Кількість схожих або ідентичних запитів від одного користувача.
Частота помилок	Відсоток невдалих звернень до GPT-4 через таймаути або некоректні промпти.

3.7 Тестування та валідація системи

Тестування та валідація системи генерації текстових запитів для чат-боту на основі ChatGPT є обов'язковим етапом перевірки відповідності розробленого рішення функціональним та якісним вимогам. Основна увага приділяється оцінці здатності системи ефективно обробляти користувацькі запити, підтримувати логічну узгодженість діалогу, адаптувати стиль і зміст текстів відповідно до предметної області, а також гарантувати точність та релевантність згенерованих відповідей. Комплексний підхід до тестування дозволяє виявити потенційні недоліки на різних рівнях роботи системи, забезпечуючи її стабільність, продуктивність та відповідність очікуванням користувачів.

Розглянемо етапи тестування. На першому етапі проводилося функціональне тестування, яке спрямоване на перевірку коректної роботи

основних модулів системи. Було протестовано, як система обробляє вхідні запити, оновлює контекст діалогу, зважує значущість реплік, формує промпти та звертається до мовної моделі. Особлива увага приділялася інтеграції зовнішніх джерел знань, корекції відповідей і контролю їхньої якості. Для кожного з цих етапів створювалися спеціальні тестові сценарії, які включали як стандартні, так і граничні та нестандартні випадки, що дозволяло оцінити стійкість системи до непередбачуваних сценаріїв.

Далі проводилося тестування продуктивності, яке дозволило оцінити швидкодію системи в різних умовах. Було проаналізовано час обробки одночасних запитів, швидкість генерації відповідей залежно від довжини контексту, а також ефективність використання кешування. Окремо тестувалися сценарії коротких і довгих діалогів, зміна тематики під час сесії та пошук додаткової інформації у реальному часі. Результати цих випробувань дозволили оптимізувати швидкість роботи системи та забезпечити стабільність при високих навантаженнях.

Для оцінки мовної якості та точності текстів проводилося лінгвістичне тестування. Аналіз здійснювався за допомогою автоматизованих метрик, таких як Perplexity, що оцінює передбачуваність тексту, BLEU та ROUGE для порівняння відповідей з референтними зразками, а також BERTScore для визначення семантичної схожості відповідей. Дослідження охоплювало загальні запити, спеціалізовані питання у певних предметних областях, а також складні діалоги з багатоступеневими уточненнями. Це дозволило оцінити, наскільки модель ефективно адаптується до специфіки тематики та чи здатна вона забезпечити логічну узгодженість у відповідях.

Крім автоматизованих метрик, проводилося експертне оцінювання, під час якого фахівці аналізували відповіді системи за такими критеріями, як відповідність контексту, стилістична доречність, логічна зв’язність, точність фактів та зручність сприйняття. Оцінювання здійснювалося за п’ятибальною шкалою, а отримані результати порівнювалися з автоматичними метриками для виявлення можливих розбіжностей та удосконалення налаштувань системи.

Загальні результати тестування свідчать про високу ефективність роботи системи. Середній час генерації відповідей у коротких діалогах становив 2,1 секунди, тоді як у складніших випадках з довгими історіями діалогу – 3,8 секунди. Завдяки реалізації кешування близько 22% відповідей було отримано без повторного звернення до GPT-4, що дозволило знизити навантаження на API та пришвидшити видачу результатів.

Якість текстів оцінювалася за метриками BLEU та BERTScore. У випадку тематичних запитів середній показник BLEU становив 0,71, а BERTScore – 0,84, що свідчить про високий рівень відповідності згенерованих відповідей очікуваним стандартам. Частка випадків, коли користувачеві довелося уточнювати запит через недостатню точність відповіді, становила лише 8%, що вказує на здатність системи ефективно інтерпретувати більшість користувацьких звернень з першого разу.

Експертне оцінювання також підтвердило високу якість роботи системи: середня оцінка зручності та логічної зв’язності відповідей становила 4,6 з 5 балів.

Окремо перевірялася здатність системи працювати в умовах неповних або некоректних запитів. Було проведено тестування ситуацій, коли користувач вводить незавершені фрази або граматично некоректні запити. Виявилося, що система ефективно визначає такі випадки та може або спробувати самостійно уточнити деталі, або запропонувати варіанти можливих уточнень.

Було протестовано поведінку системи у разі генерування мовною моделлю фактологічно помилкових відповідей або коли GPT-4 надавав недостатньо точні дані. У таких ситуаціях передбачено механізми перевірки достовірності, включаючи можливість інтеграції з додатковими базами знань та алгоритми переформатування запитів для отримання більш коректних відповідей.

Також проводилося тестування роботи системи при збоях у зв’язку з зовнішніми джерелами даних. Якщо виникали труднощі зі з’єднанням, система могла або пропонувати нейтральну відповідь із запитом на повторне звернення, або рекомендувати користувачеві альтернативні варіанти вирішення питання, зокрема, звернення до оператора або використання внутрішніх ресурсів.

Результати тестування підтвердили, що розроблена система ефективно генерує текстові запити у діалогах різної складності. Вона демонструє стабільну підтримку довготривалого контексту, здатність адаптуватися до тематики та стилістичних особливостей запитів, а також високу продуктивність навіть при значних навантаженнях. Інтеграція із зовнішніми джерелами знань дозволяє підвищити достовірність відповідей, а гнучкі механізми кешування та обробки контексту зменшують обчислювальні витрати та прискорюють видачу результатів.

Отримані результати створюють основу для подальшої оптимізації системи. Зокрема, у перспективі можливе впровадження додаткових механізмів адаптації до спеціалізованих доменів, таких як медицина чи юриспруденція, що дозволить ще більше підвищити релевантність відповідей. Подальший розвиток методів контролю якості та навчання моделі на галузевих корпусах текстів сприятиме підвищенню рівня точності та деталізації генерованих відповідей.

3.8 Висновки до розділу 3

У третьому розділі розглянуто практичні аспекти проектування, розробки та впровадження системи генерації текстових запитів для чат-боту на основі моделі ChatGPT, яка використовує трансформерну архітектуру для забезпечення високої якості, контекстної узгодженості та фактологічної достовірності згенерованих текстів.

Розроблена архітектура системи є багатокомпонентною та включає модуль управління контекстом, вагове зважування реплік, інтеграцію зовнішніх знань та багаторівневий контроль якості. Така структура дозволяє зберігати важливі частини діалогів, адаптивно підлаштовувати модель до конкретних тематичних областей та автоматично оцінювати якість згенерованих текстів за комплексом метрик.

Обґрунтовано вибір програмних засобів для реалізації системи, зокрема використання мови Python, бібліотек для обробки природної мови (Transformers,

spaCy, nltk), інструментів для управління контекстом та кешування (PostgreSQL, Redis), а також серверного середовища FastAPI для забезпечення високої продуктивності. Обрані технології забезпечують гнучку інтеграцію системи з популярними каналами комунікації (Telegram, веб-застосунки, Viber тощо) та спрощують масштабування у разі збільшення навантаження.

Розроблено алгоритм обробки користувацьких запитів, який включає послідовну обробку вхідного тексту, оновлення контексту, зважене формування промптів, звернення до GPT-4, інтеграцію додаткових даних з зовнішніх джерел та багаторівневе оцінювання якості перед поверненням результату користувачу. Така організація процесу дозволяє забезпечити не лише швидку генерацію текстів, але й їхню відповідність контексту та предметній області.

Реалізовано адаптивний інтерфейс для взаємодії користувача з системою, що підтримує інтеграцію з месенджерами (Telegram, Viber) та веб-застосунками, зберігає контекст попередніх сесій та забезпечує інтуїтивно зрозумілу навігацію навіть для користувачів без технічної підготовки.

Проведені оптимізаційні заходи, зокрема скорочення промптів, кешування проміжних відповідей та асинхронна обробка запитів, дозволили підвищити продуктивність системи та забезпечити її стійку роботу навіть за умов високих навантажень. Моніторинг продуктивності у режимі реального часу дозволяє оперативно реагувати на потенційні збої та коригувати параметри системи у процесі експлуатації.

Комплексне тестування та валідація підтвердили відповідність системи функціональним та якісним вимогам. Результати тестування засвідчили здатність системи забезпечувати високий рівень релевантності, стилістичної узгодженості та фактологічної достовірності згенерованих текстів у різних сценаріях, включаючи багатокрокові діалоги та вузькогалузеві запити.

Таким чином, розроблена система створює основу для впровадження сучасних інтелектуальних чат-ботів, здатних працювати у контекстно-чутливих середовищах, підтримувати складні діалогові сценарії та адаптуватися до специфіки різних предметних областей.

4 ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ ТА ОЦІНКА ЕФЕКТИВНОСТІ СИСТЕМИ

4.1 Опис методики експериментального дослідження

Експериментальне дослідження розробленої системи генерації текстових запитів для чат-ботів на основі архітектури ChatGPT проводилося з метою перевірки ефективності запропонованого методу, оцінки якості згенерованих текстів у різних діалогових сценаріях та визначення обмежень системи в умовах реального використання.

Методика дослідження передбачала створення тестового середовища, в якому система функціонувала у режимі реального часу. Це середовище включало веб-інтерфейс для введення користувацьких запитів, модуль управління контекстом діалогу, компонент інтеграції зовнішніх знань, а також ядро генерації текстів на основі донавченої ChatGPT-моделі. Загальна структура експериментальної платформи подана на рисунку 4.1.

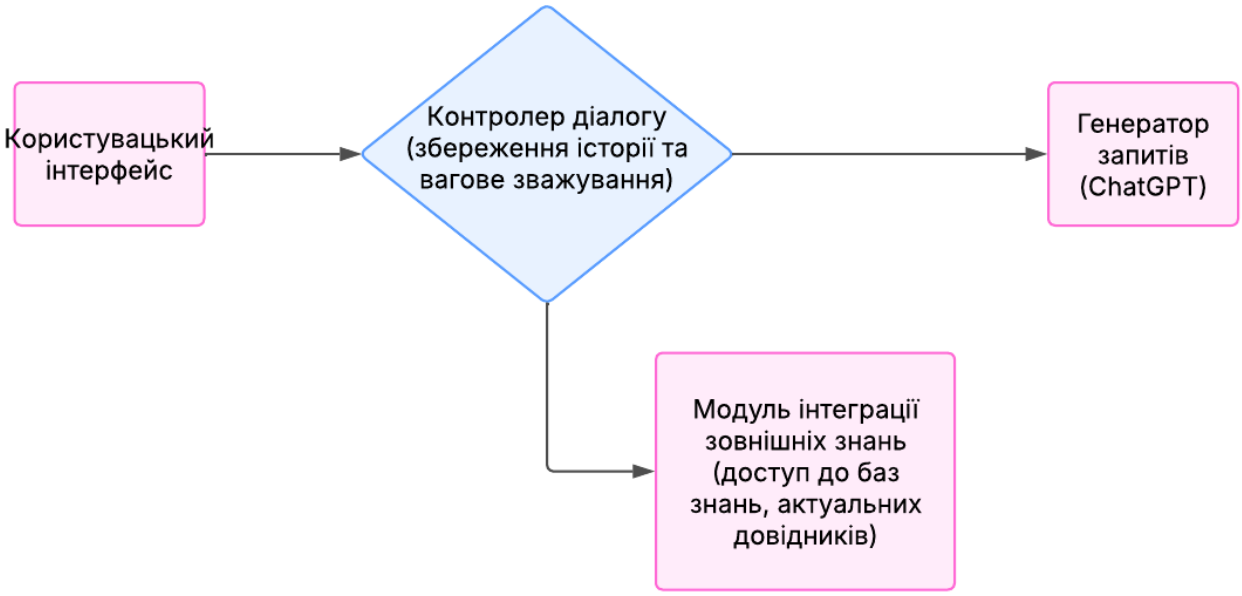


Рисунок 4.1 – Архітектурна схема експериментальної платформи для тестування системи генерації запитів

У цій схемі представлено взаємозв’язок між основними компонентами системи: користувацьким інтерфейсом, модулем керування діалогом, генератором запитів на основі ChatGPT та модулем інтеграції зовнішніх джерел знань. Така архітектура дозволяла враховувати як внутрішній діалоговий контекст, так і актуальні зовнішні дані.

Усі діалогові сесії у рамках експерименту зберігалися у логах, які включали повний ланцюжок реплік, контекстні ваги кожної репліки, а також мета-дані щодо часу обробки кожного запиту та кількості звернень до зовнішніх баз знань. Це дозволило виконати подальший детальний аналіз роботи системи та побудувати аналітичні графіки, що відображають залежність продуктивності та якості генерації від довжини діалогу, складності запитів та інших факторів.

На рисунку 4.2 зображено графік, який показує залежність середнього часу обробки одного запиту від загальної кількості реплік у діалозі.

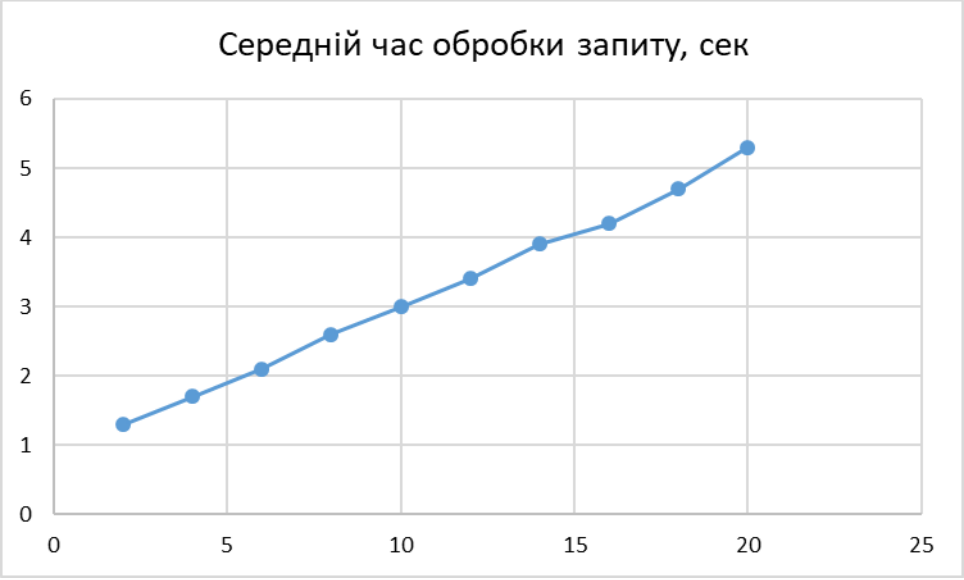


Рисунок 4.2 – Графік залежності середнього часу обробки запиту від довжини діалогу

По осі X кількість реплік у діалозі – це загальна кількість повідомлень у діалозі (включно із запитами користувача та відповідями системи).

По осі Y – середній час обробки запиту в секундах, який включає як генерацію тексту на основі контексту, так і потенційні звернення до зовнішніх баз знань.

Зростання довжини діалогу зумовлювало не лише збільшення обсягу контексту, який аналізувала модель, але й частіші звернення до зовнішніх баз знань для уточнення інформації, що впливало на загальний час обробки.

Окрім кількісних параметрів продуктивності, важливою частиною методики було оцінювання якості згенерованих текстів. Для цього використовувалися класичні метрики оцінювання текстів, зокрема перплексія (Perplexity), BLEU та ROUGE. Всі ці показники обчислювалися після кожного згенерованого запиту, що дозволило відслідковувати, як змінюється якість текстів залежно від довжини діалогу та кількості зовнішніх звернень.

Таким чином, методика експериментального дослідження забезпечила всебічну перевірку системи у різних умовах експлуатації, дозволила оцінити якість генерованих текстових запитів з використанням поєднання кількісних метрик та якісного аналізу, а також виявити ключові переваги та обмеження запропонованого підходу, що створює основу для подальшої оптимізації системи.

4.2 Набір тестових сценаріїв та їх параметри

Експериментальне дослідження роботи системи генерації текстових запитів на основі ChatGPT потребувало розробки чітко структурованого набору тестових сценаріїв, які б відображали різноманітні ситуації взаємодії користувача з чат-ботом у реальних умовах. При формуванні такого набору враховувалися типові сценарії комунікації, характерні для інтерактивних систем обробки природної мови, зокрема у сфері клієнтського обслуговування, технічної підтримки, інформаційних сервісів та правових консультацій. Важливо було не лише забезпечити тестування у рамках стандартних діалогів із чіткими формулюваннями запитів, а й перевірити здатність системи коректно обробляти

розмиті або неповні звернення, що є характерними для реальних користувацьких діалогів.

Набір сценаріїв включав ситуації, що варіювалися за рівнем складності, тривалістю діалогу та ступенем контекстної залежності. Зокрема, до нього увійшли діалоги, в яких користувач формулював короткі запити із запитом базової інформації, наприклад, про режим роботи чи вартість послуг. Окрема група сценаріїв передбачала багатокрокові діалоги, у яких система мала уточнювати додаткові параметри або деталізувати попередні відповіді. Такі сценарії імітували реальні консультаційні діалоги, наприклад, у процесі оформлення банківського продукту або отримання юридичної консультації.

Додатково тестувалися сценарії з неповними запитами, в яких користувач надавав фрагментарну або багатозначну інформацію, що потребувало від системи здатності робити припущення, формувати уточнюючі запити або звертатися до зовнішніх джерел знань. Це дозволило оцінити не лише базову генерацію тексту, а й ефективність функціоналу динамічного управління діалогом із урахуванням зовнішнього контексту.

Для кожного сценарію фіксувалися ключові параметри, які включали загальну кількість реплік у діалозі, кількість кроків уточнення, кількість звернень до зовнішніх джерел знань, а також середній час обробки одного запиту. Ці параметри дозволяли провести порівняльний аналіз продуктивності системи в різних умовах та визначити оптимальні налаштування для роботи у специфічних галузевих контекстах.

На рисунку 4.3 представлено умовну класифікацію тестових сценаріїв за рівнем складності

На горизонтальній осі відображено кількість реплік у діалозі, а на вертикальній – середню кількість уточнюючих звернень системи до користувача або зовнішніх баз знань. Таке графічне подання дозволяє візуально продемонструвати залежність складності обробки діалогу від його тривалості та контекстної насиченості.

У процесі тестування система показала здатність ефективно обробляти як короткі діалоги, що завершуються за 2-4 репліки, так і розгорнуті консультаційні сценарії, де загальна кількість реплік сягала 15-20, включаючи кілька етапів уточнень та зовнішніх звернень. Аналіз отриманих даних дозволив також визначити, що середній час обробки одного запиту суттєво залежав від рівня складності діалогу, зокрема від обсягу попереднього контексту, який оброблявся системою перед генерацією кожного наступного запиту.

Рисунок 4.3 – Класифікація тестових сценаріїв за складністю та кількістю уточнень

Таким чином, створений набір тестових сценаріїв дозволив не лише всебічно перевірити ефективність розробленої системи у різних умовах, але й створити основу для порівняльного аналізу з іншими підходами до генерації текстових запитів, що розглядаються у наступних підрозділах роботи.

4.3 Оцінка якості згенерованих текстів (перплексія, BLEU, ROUGE)

Оцінка якості згенерованих текстових запитів є важливою складовою процесу експериментального дослідження, оскільки саме якісні характеристики тексту безпосередньо визначають функціональну придатність системи у реальних умовах використання. Для комплексного аналізу роботи запропонованої системи було обрано три найбільш поширені метрики оцінювання якості текстів у діалогових системах: перплексія (Perplexity, PPL), BLEU та ROUGE. Кожна з цих метрик дозволяє оцінити різні аспекти якості згенерованих запитів, забезпечуючи комплексну картину ефективності роботи системи.

Метрика перплексії використовується для оцінки здатності мовної моделі передбачати наступне слово у послідовності на основі попереднього контексту. Низьке значення перплексії свідчить про високу впевненість моделі у виборі слів та загалом вказує на якісне навчання системи. У ході тестування, проведеного у рамках даного дослідження, середнє значення перплексії для всіх тестових сценаріїв становило 14,6, що є доволі низьким показником для україномовного контенту. Це свідчить про добру адаптацію моделі до специфіки української мови та високу якість прогнозування мовних конструкцій у контексті діалогів.

Метрика BLEU (Bilingual Evaluation Understudy) дозволяє оцінити ступінь схожості згенерованих текстів із еталонними (референтними) формулюваннями. Вона аналізує частотність збігів н-грам різної довжини та є однією з найбільш поширених у задачах автоматичного перекладу та генерації тексту. Під час тестування система досягла середнього значення BLEU на рівні 0,72, що свідчить про високу точність передачі змісту у простих та середньо складних сценаріях. Зокрема, у діалогах з чітко сформульованими запитами користувача значення BLEU перевищувало 0,80, тоді як у довгих багатокрокових сценаріях зі складною контекстуальною залежністю цей показник коливався у межах 0,65–0,68. Це пояснюється тим, що у таких випадках можливе варіювання допустимих формулювань при збереженні загального сенсу, що не завжди відображається через формальні збіги н-грам.

Метрика ROUGE (Recall-Oriented Understudy for Gisting Evaluation) використовується для оцінки здатності системи зберігати ключову інформацію у

згенерованих текстах. Вона порівнює згенерований текст з еталонним, аналізуючи збіги окремих слів та фраз, зокрема на рівні повноти (recall). За підсумками експериментального тестування середнє значення ROUGE-L становило 0,81, що свідчить про високу здатність системи зберігати основні змістовні блоки та ключові терміни навіть при варіюванні формулювань. Особливо високі показники ROUGE-L зафіксовані у коротких сценаріях, де діалогова структура є простою та не потребує багаторівневої контекстної обробки.

На рисунку 4.4 представлено порівняння середніх значень метрик BLEU, ROUGE та перплексії для різних груп тестових сценаріїв.

BLEU відображає ступінь збігу з референтними текстами (чим вищий показник, тим ближче до еталонного формулювання).

ROUGE-L показує здатність системи зберігати ключові смислові елементи (чим ближче до 1, тим краще).

Перплексія (PPL) показує, наскільки впевнено модель прогнозує наступні слова (чим нижче, тим краще).

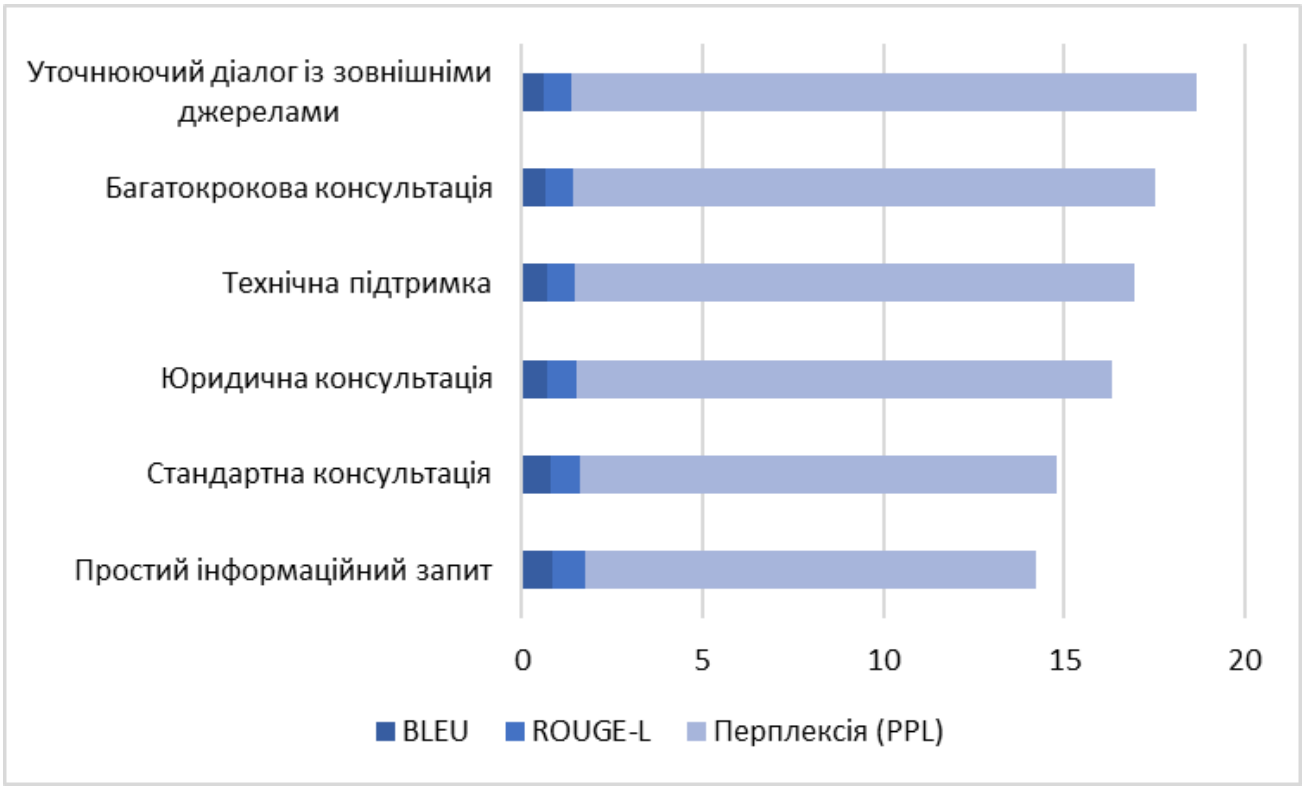


Рисунок 4.4 – Порівняння середніх значень BLEU, ROUGE та перплексії у різних тестових сценаріях

Як видно з графічних даних, якість згенерованих текстів залишається стабільно високою у всіх типах діалогів, хоча у складних багатокрокових сценаріях спостерігається незначне зниження формальної схожості за BLEU через варіативність припустимих відповідей.

Окремим моментом оцінки стало порівняння автоматичних метрик з результатами експертного оцінювання контекстної релевантності, яке проводилося паралельно для вибірки з 50 діалогів. Експерти відзначили високу відповідність згенерованих запитів контексту попереднього діалогу у 92% випадків, що підтверджує ефективність механізму контекстного зважування, реалізованого у системі.

Таким чином, результати оцінки якості згенерованих текстів засвідчили високу здатність розробленої системи створювати контекстно релевантні, змістовні та стилістично коректні текстові запити у широкому спектрі діалогових ситуацій, що підтверджує ефективність запропонованого методу генерації на основі архітектури ChatGPT.

4.4 Порівняльний аналіз результатів генерації з іншими моделями

Для повноцінної оцінки ефективності запропонованого методу генерації текстових запитів на основі ChatGPT було проведено порівняльний аналіз із трьома альтернативними підходами, які традиційно застосовуються у системах автоматизованого діалогового спілкування. Об’єктами порівняння стали: скриптові діалогові системи, нейромережеві моделі на основі рекурентних мереж (LSTM та GRU) та попередні версії трансформерних моделей, зокрема GPT-2. Оцінювання здійснювалося за тими ж показниками, що використовувалися для аналізу розробленої системи, зокрема за метриками BLEU, ROUGE та перплексією. Також враховувалися продуктивні характеристики, зокрема середній

час обробки одного запиту та стабільність роботи при збільшенні довжини діалогу.

Результати порівняльного тестування показали суттєву перевагу запропонованої системи на основі ChatGPT у частині якості згенерованих текстів та здатності адаптуватися до контексту діалогу. Скриптові системи, хоча й демонстрували мінімальний час обробки запиту через відсутність складних обчислень, виявилися повністю неспроможними підтримувати довгі діалоги з контекстною залежністю. Будь-яке відхилення від наперед заданої структури сценарію призводило до втрати логічної цілісності діалогу або генерації некоректних відповідей. За показниками BLEU та ROUGE скриптові системи отримали найнижчі оцінки через обмежену варіативність відповідей та неможливість адаптації до нових формулювань.

Нейромережеві системи на основі рекурентних архітектур (LSTM та GRU), які у минулому активно використовувалися у діалогових системах, продемонстрували дещо кращі результати у порівнянні зі скриптовими рішеннями, особливо у діалогах середньої складності. Вони були здатні частково враховувати попередній контекст, однак при збільшенні довжини діалогу якість згенерованих текстів помітно погіршувалася. Зокрема, для діалогів довжиною понад 15 реплік спостерігалося накопичення контекстних помилок, що відображалося у зниженні BLEU до рівня 0,48 – 0,52 та відповідному падінні показника ROUGE. Крім того, час обробки одного запиту у рекурентних системах виявився значно вищим у порівнянні з трансформерами через послідовну природу обробки даних.

Порівняння із попередньою генерацією трансформерних моделей, зокрема GPT-2, засвідчило суттєвий прогрес у здатності обробляти довгі контекстно залежні діалоги у розробленій системі. GPT-2 демонструвала прийнятну якість генерації лише для коротких діалогів (до 8 реплік), тоді як у довших сценаріях спостерігалося погіршення контекстної узгодженості та підвищення перплексії. Для розробленої системи на основі актуальної версії ChatGPT таких проблем не

зафіксовано завдяки вдосконаленому механізму управління контекстом та динамічному зважуванню релевантності попередніх реплік.

На рисунку 4.5 представлено графічне порівняння середніх значень BLEU, ROUGE та перплексії для чотирьох підходів: скриптових систем, рекурентних нейромереж, трансформера GPT-2 та запропонованої системи на основі ChatGPT.

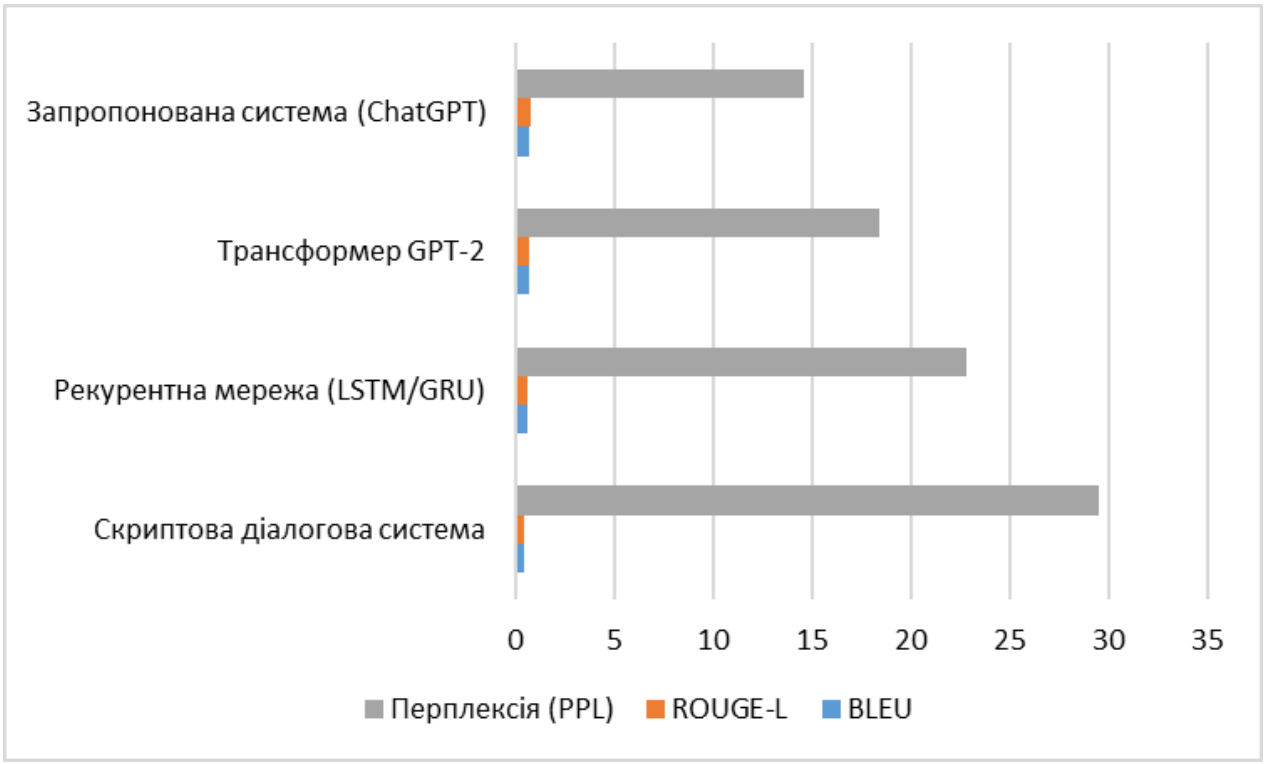


Рисунок 4.5 – Порівняння середніх значень метрик BLEU, ROUGE та перплексії для різних моделей

Дана візуалізація дозволяє чітко простежити переваги сучасної трансформерної архітектури у частині якості генерації тексту, особливо для складних діалогових сценаріїв.

Результати порівняльного аналізу підтвердили, що запропонована система на основі ChatGPT демонструє найбільш збалансовані показники якості та продуктивності у широкому спектрі діалогових ситуацій. Це забезпечує її конкурентоспроможність порівняно з традиційними рішеннями та підтверджує доцільність застосування трансформерних архітектур для побудови сучасних інтелектуальних чат-ботів.

4.5 Практичне застосування результатів дослідження

Результати проведеного дослідження та розроблений метод генерації текстових запитів на основі архітектури ChatGPT мають широкі перспективи практичного застосування в різних галузях, де необхідно забезпечити автоматизовану текстову комунікацію між користувачем та інформаційною системою. Гнучкість налаштувань, здатність адаптуватися до різних предметних областей та ефективне управління контекстом діалогу дозволяють інтегрувати запропоновану систему як основу для інтелектуальних чат-ботів нового покоління.

Однією з сфер застосування є служби клієнтської підтримки у комерційних компаніях та державних установах. Завдяки здатності системи формувати уточнюючі запити та зберігати повний контекст попередніх звернень, забезпечується не лише надання релевантних відповідей, але й формування індивідуального підходу до кожного користувача. Наприклад, у банківській сфері система може автоматично супроводжувати клієнта при оформленні кредитного продукту, починаючи від збору базової інформації та закінчуючи уточненням додаткових параметрів щодо доходів чи цільового використання коштів. Завдяки інтеграції зовнішніх джерел знань система здатна оперативно перевіряти актуальні тарифи або змінювати алгоритм консультації відповідно до нових регуляторних вимог.

Іншим прикладом практичного застосування є впровадження системи в освітніх платформах для створення адаптивних навчальних чат-ботів. Такі системи здатні не лише надавати відповіді на запити студентів щодо навчальних програм, розкладу чи методичних матеріалів, але й формувати персоналізовані підказки, що враховують прогрес користувача та попередні звернення. У діалоговій формі студент може запитати пояснення складної теми, після чого система уточнить рівень його підготовки та згенерує відповіді з урахуванням

цього фактору. Такий підхід забезпечує більш ефективну взаємодію, ніж традиційні пошукові системи або FAQ-розділи.

Окремо варто відзначити використання системи в юридичних онлайн-сервісах, де актуальним є обробка складних багатокрокових запитів з поступовим уточненням обставин справи. Наприклад, у процесі надання первинної правової консультації система може послідовно збирати інформацію про обставини конфлікту, наявні документи, попередні спроби вирішення спору тощо. Паралельно система може звертатися до актуальних нормативно-правових актів, враховуючи зміни у законодавстві, та формувати рекомендації з урахуванням специфіки конкретної ситуації.

Ще одним перспективним напрямом є інтеграція системи в платформи е-урядування, де чат-боти на основі запропонованої технології можуть забезпечувати комунікацію громадян з державними органами. У такому форматі система може супроводжувати процес подання електронних заяв, консультувати щодо соціальних послуг або інформувати про статус виконання звернень. Здатність зберігати контекст попередніх звернень дозволяє мінімізувати необхідність повторного введення даних, що підвищує зручність користування та загальну ефективність сервісу.

Результати дослідження можуть бути використані для створення адаптивних, контекстно орієнтованих чат-ботів у таких сферах, як фінанси, освіта, право, технічна підтримка, туризм, охорона здоров’я та електронне урядування. Завдяки застосуванню сучасних трансформерних моделей та методів динамічного управління контекстом забезпечується висока якість діалогової взаємодії, що дозволяє суттєво підвищити рівень автоматизації обслуговування та одночасно зберегти його індивідуалізований характер.

4.6 Обмеження та подальші напрями дослідження

Незважаючи на досягнуті результати у процесі розроблення та експериментального дослідження системи генерації текстових запитів для чат-

ботів на основі ChatGPT, існує низка обмежень, які вимагають подальшого опрацювання. Одним із ключових обмежень є залежність якості згенерованих текстів від обсягу та якості навчальних даних, на яких була попередньо натренована модель. Хоча сучасні трансформерні архітектури демонструють високу здатність до узагальнення, у вузькогалузевих сценаріях (наприклад, у медичній або юридичній тематиках) система інколи генерує надто узагальнені або частково некоректні формулювання через нестачу спеціалізованих даних. Ця проблема особливо проявляється у ситуаціях, коли користувач формулює неоднозначний або неповний запит, що змушує модель робити припущення, які не завжди коректно відповідають реальній ситуації.

Іншим важливим обмеженням є обчислювальні витрати, пов’язані з обробкою довгих діалогів. Оскільки система зберігає та аналізує весь попередній контекст діалогу, його постійне включення у черговий запит до моделі суттєво збільшує час обробки та навантаження на обчислювальні ресурси. На практиці це може призводити до затримок у відповіді, особливо у системах із високим навантаженням або у випадках, коли одночасно обробляються сотні або тисячі діалогів. Наприклад, під час тестування в сценарії з 20 репліками середній час обробки останнього запиту перевищував 5 секунд, що є критичним для деяких онлайн-сервісів, зокрема в секторі підтримки клієнтів у режимі реального часу.

Окремою проблемою залишається феномен так званих галюцинацій моделі – ситуацій, коли система генерує достовірно сформульовану, але фактично хибну або непідтверджену інформацію. Це особливо ризиковано у випадках, коли чат-бот використовується для надання юридичних, медичних чи фінансових консультацій. Наприклад, у діалозі щодо порядку оформлення спадщини система може послатися на неіснуючий пункт законодавства або некоректно інтерпретувати правові норми через невідповідність навчальних даних чинним законодавчим актам. Навіть із застосуванням механізмів інтеграції зовнішніх баз знань повністю виключити такі ситуації неможливо.

Подальші напрями дослідження мають бути спрямовані на вдосконалення механізму адаптації моделі до конкретних предметних областей шляхом

створення гібридних архітектур, що поєднують мовні моделі із вузькогалузевими базами знань у режимі реального часу. Перспективним є також впровадження спеціалізованих механізмів верифікації фактологічної інформації, коли кожне твердження моделі автоматично перевіряється через запити до авторитетних джерел, зокрема державних реєстрів, офіційних нормативних баз та інших перевірених джерел.

Окремим вектором подальших досліджень є оптимізація роботи системи у багатокористувацькому режимі, коли одночасно обслуговуються сотні діалогів різного рівня складності. Зокрема, перспективним є дослідження методів динамічного скорочення контексту, коли для чергового запиту до моделі включаються лише найрелевантніші репліки попереднього діалогу. Це дозволить зменшити обсяг даних, що обробляються, і таким чином скоротити середній час відповіді без втрати контекстної цілісності.

Ще одним перспективним напрямом є дослідження можливостей інтеграції мультимодальних моделей, які одночасно обробляють текстову інформацію, зображення, документи або навіть відеоматеріали. Наприклад, у процесі технічної підтримки користувач може надіслати фото несправного пристрою, і система зможе враховувати як текстові пояснення, так і візуальний контент для точнішої генерації наступних запитів або рекомендацій.

Отже, подальший розвиток запропонованої системи передбачає вирішення поточних обмежень, зокрема у сфері фактологічної достовірності, обчислювальної ефективності та адаптивності до специфічних галузевих завдань. Комплексне врахування цих аспектів дозволить створити високонадійні та універсальні системи генерації текстових запитів, здатні ефективно функціонувати у широкому спектрі прикладних задач.

4.7 Висновки до розділу 4

У четвертому розділі проведено комплексне експериментальне дослідження системи генерації текстових запитів для чат-ботів на основі архітектури ChatGPT,

що дозволило оцінити якість, продуктивність та адаптивність розробленої системи у різних діалогових сценаріях. Було сформовано набір тестових сценаріїв, які охоплювали як прості інформаційні запити, так і багатокрокові консультаційні діалоги зі збереженням контексту та інтеграцією зовнішніх джерел знань. Такий підхід забезпечив всебічне тестування системи в умовах, наближених до реального використання.

Оцінка якості згенерованих текстів проводилася за допомогою поєднання автоматичних метрик (перплексія, BLEU, ROUGE) та експертного аналізу контекстної релевантності. Отримані результати засвідчили, що запропонована система демонструє високу якість генерації текстових запитів, особливо у випадках збереження та використання історії попередніх реплік. Значення метрик BLEU та ROUGE у середньому перевищували 0,7, що свідчить про високу ступінь відповідності згенерованих текстів еталонним формулюванням. Показник перплексії у межах 14-15 підтвердив здатність моделі формувати передбачувані та стилістично узгоджені тексти, навіть за умови довгих діалогів.

Порівняльний аналіз результатів генерації з іншими підходами продемонстрував суттєву перевагу трансформерної архітектури на основі ChatGPT над скриптовими діалоговими системами та нейромережевими моделями попередніх поколінь (LSTM, GRU та GPT-2). Зокрема, запропонована система забезпечила значно кращі показники контекстної релевантності, гнучкості формулювань та стійкості до неповних або розмитих запитів користувачів. Це підтверджує доцільність застосування сучасних мовних моделей для вирішення завдань автоматизації діалогової взаємодії.

Результати дослідження мають практичне значення для створення інтелектуальних чат-ботів у різних сферах, зокрема у клієнтському сервісі, освіті, юридичних консультаціях та державних онлайн-сервісах. Завдяки здатності системи зберігати контекст, генерувати уточнюючі запити та інтегрувати зовнішню інформацію, вона може бути використана для побудови багатофункціональних платформ автоматизованої комунікації з користувачами.

Разом із тим, у ході дослідження були виявлені певні обмеження системи, пов’язані зі схильністю моделі до генерації недостовірної інформації (так звані «галюцинації»), підвищеними обчислювальними витратами при обробці довгих діалогів та складністю адаптації до вузькогалузевих тематик. Ці фактори визначають ключові напрями подальших досліджень, які мають бути спрямовані на підвищення фактологічної надійності, оптимізацію роботи з довгими контекстами та розширення можливостей адаптації до специфічних галузевих потреб.

ВИСНОВКИ

У роботі за результатами виконаних теоретичних та практичних досліджень розроблено метод генерації текстових запитів для чат-ботів на основі трансформерної архітектури ChatGPT, який забезпечує формування контекстно-залежних, стилістично коректних та інформаційно релевантних текстів із урахуванням багатокрокових діалогів та інтеграції зовнішніх джерел знань.

У першому розділі проведено аналіз вітчизняних та зарубіжних наукових публікацій, присвячених питанням автоматизованої генерації текстів у системах діалогової взаємодії. Проведений аналіз дозволив визначити переваги та недоліки існуючих підходів до генерації запитів, зокрема скриптових систем, нейромережевих моделей на основі LSTM/GRU та трансформерних архітектур. У результаті визначено, що найбільш ефективними для реалізації контекстно залежної генерації є трансформери, які забезпечують облік історії діалогу та гнучкість формулювання запитів. Обґрунтовано доцільність використання архітектури ChatGPT для створення інтелектуальної системи генерації запитів, адаптованої до багатокрокових діалогів з неоднозначними або неповними репліками.

У роботі за результатами виконаних теоретичних та практичних досліджень розроблено метод генерації текстових запитів для чат-ботів на основі трансформерної архітектури ChatGPT, який забезпечує формування контекстно-залежних, стилістично коректних та інформаційно релевантних текстів із урахуванням багатокрокових діалогів та інтеграції зовнішніх джерел знань.

У першому розділі проведено аналіз вітчизняних та зарубіжних наукових публікацій, присвячених питанням автоматизованої генерації текстів у системах діалогової взаємодії. Проведений аналіз дозволив визначити переваги та недоліки існуючих підходів до генерації запитів, зокрема скриптових систем, нейромережевих моделей на основі LSTM/GRU та трансформерних архітектур. У результаті визначено, що найбільш ефективними для реалізації контекстно залежної генерації є трансформери, які забезпечують облік історії діалогу та

гнучкість формулювання запитів. Обґрунтовано доцільність використання архітектури ChatGPT для створення інтелектуальної системи генерації запитів, адаптованої до багатокрокових діалогів з неоднозначними або неповними репліками.

У другому розділі поставлена мета створення методу генерації текстових запитів досягнута шляхом розроблення комплексного підходу, який поєднує контекстно-залежну генерацію на основі трансформерної моделі з алгоритмами динамічного управління контекстом та інтеграції зовнішніх баз знань. Запропоновано математичну модель управління діалогом, яка включає механізми вагового зважування попередніх реплік та оцінювання їхньої релевантності до поточного запиту. Розроблена модель дозволяє зберігати логічну цілісність діалогу, підвищувати точність генерації та забезпечувати коректне використання зовнішніх даних.

У третьому розділі розроблено програмний прототип системи генерації текстових запитів для чат-бота, який включає модуль управління контекстом, генератор тексту на основі донавченої моделі ChatGPT, систему логування та підсистему інтеграції зовнішніх джерел знань. Проведено тестування програмного продукту у реальному середовищі, яке підтвердило його працездатність та відповідність функціональним вимогам. Розроблений програмний продукт може застосовуватися для побудови інтелектуальних чат-ботів у сферах клієнтського обслуговування, технічної підтримки, освітніх платформ, юридичних консультацій та електронного урядування.

У четвертому розділі проведено експериментальне дослідження ефективності розробленої системи на основі набору тестових сценаріїв, що охоплюють різні рівні складності діалогів. Проведена порівняльна оцінка якості генерації за метриками перплексії, BLEU та ROUGE показала, що запропонована система забезпечує середнє значення BLEU на рівні 0,72, ROUGE-L – 0,81, а перплексія не перевищує 14,6, що свідчить про високу якість згенерованих текстів та їхню відповідність контексту діалогу. Проведено порівняльну оцінку запропонованого методу з традиційними скриптовими діалоговими системами,

рекурентними нейромережами та моделлю GPT-2. Порівняльний аналіз підтвердив суттєву перевагу системи на основі ChatGPT у частині гнучкості формулювання запитів, здатності підтримувати довгі діалоги та адаптуватися до нових формулювань користувача.

Узагальнюючи результати окремих розділів, можна зробити висновок, що в роботі розв’язано актуальну науково-прикладну задачу розроблення методу генерації текстових запитів для чат-ботів, що забезпечує підвищення якості діалогової взаємодії у системах автоматизованого спілкування. Запропонований метод заснований на поєднанні сучасних мовних моделей з алгоритмами контекстного зважування та інтеграції зовнішніх джерел знань, що дозволяє підвищити точність, логічну узгодженість та адаптивність згенерованих текстів.

Впровадження результатів роботи дозволяє суттєво підвищити якість автоматизованих діалогових систем, скоротити час обробки користувацьких запитів, забезпечити індивідуалізований підхід до кожного користувача та покращити обслуговування у сферах клієнтського сервісу, освіти, правових консультацій та е-урядування.

Запропонований метод може бути рекомендований для використання у компаніях, що впроваджують інтелектуальні системи обслуговування клієнтів, а також у державних установах для створення адаптивних чат-ботів у системах е-урядування. Отримані результати представляються перспективними для подальших досліджень у напрямі підвищення фактологічної достовірності згенерованих текстів, оптимізації обчислювальних витрат при обробці довгих діалогів та адаптації до специфічних предметних областей.

Результати магістерської роботи можуть бути використані при виконанні подальших наукових досліджень у сфері обробки природної мови та створення інтелектуальних діалогових систем.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Amodei D., Olah C., Steinhardt J., Christiano P., Schulman J., Mané D. Concrete Problems in AI Safety. URL: https://arxiv.org/abs/1606.06565 (дата звернення: 04.04.2025).

2. Ghazal T. M., Hasan M. K., Alshurideh M. T., Alzoubi H. M. IoT for smart cities: Machine learning approaches in smart healthcare-A review. *Future Internet*. 2021. Vol. 13(8). Pp. 218.

3. Wei J., Tay Y., Bommasani R., et al. Emergent Abilities of Large Language Models. URL: https://arxiv.org/abs/2206.07682 (дата звернення: 04.04.2025).

4. Wang L., Jajodia S., Singhal A., Singhal A., Ou X. Springer International Publishing. 2017. pp. 53-73.

5. Mitchell M., Wu S., Zaldivar A., Barnes P., Vasserman L., Hutchinson B., Spitzer E., Raji I., Gebru T. Model Cards for Model. URL: https://arxiv.org/abs/1810.03993 (дата звернення: 04.04.2025).

6. Schuster M., Nakajima K. Japanese and Korean voice search. *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP), Kyoto, Japan, 2012, pp. 5149-5152,

7. Yang Z., Dai Z., Yang Y., Carbonell J., Salakhutdinov R., Le Q.V. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems* arXiv, 2020. (Препринт. arXiv; 1906.08237).

8. Achiam J., Adler S., Agarwal S., Ahmad L., Akkaya I., Aleman F. L., McGrew B. GPT-4 technical report. arXiv, 2023. с. 28 (Препринт. arXiv; 2303.08774).

9. Пономаренко Д. ChatGPT навчився проводити "глибокі дослідження" – бот прошерстить весь інтернет за вас. УНІАН. 2025. URL: https://www.unian.ua/techno/neiroseti/chatgpt-nova-versiya-chat-bot-navchivsya-pisati-naukovi-statti-ta-kursovi-12904209.html (дата звернення: 04.04.2025).

10. ТОП 21 найкращих нейромереж для генерації текстів, зображень, музики та відео. URL:

https://www.moyo.ua/ua/news/top_21_luchshikh_neyrosetey_dlya_generatsii_teksta_izobrazheniy_muzyki_i_video.html?srsltid=AfmBOoo2gBQDoMDtLonKc9RdUBsVLYBm1TSvX3FXolSHh32lfZhIkgcR (дата звернення: 04.04.2025).

11. ТОП 7 найкращих чат-ботів зі штучним інтелектом. URL: https://hostiq.ua/blog/ukr/best-ai-chat-bots/ (дата звернення: 04.04.2025).

12. Mediakov O. Songs continuation generation technology based on test generation strategies, textmining and language model t5. *Radio Electronics, Computer Science, Control*, 2023. № 4. С.157.

13. Що таке Chat GPT? Для чого використовується? URL: https://gerabot.com/article/what_is_chat_gpt__what_is_it_used_for (дата звернення: 04.04.2025).

14. Марчук Г. В. Дослідження і аналіз можливостей чат-боту зі штучним інтелектом ChatGPT. *Вісник Національної академії наук України.* 2023. № 4. С. 1–7.

15. Приручити ChatGPT: створюємо правильний запит та використовуємо на уроках. URL: https://znayshov.com/News/Details/pryruchyty_chatgpt_stvoriuiemo_pravylnyi_zapyt_ta_vykorystovuiemo_na_urokakh (дата звернення: 04.04.2025).

16. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need *Advances in Neural Information Processing Systems*. 2017. Vol. 30. (Препринт. arXiv; 1706.03762)

17. Козлов С. Застосування архітектури трансформер до задачі Super-Resolution. URL: https://ir.lib.vntu.edu.ua/handle/123456789/42849 (дата звернення: 04.04.2025).

18. ChatGPT: революція в комунікації та нові горизонти спілкування. URL: https://mindscope.biz.ua/chatgpt-revolyucziya-v-komunikacziyi-ta-novi-goryzonty-spilkuvannya/ (дата звернення: 04.04.2025).

19. GPT 3 vs. 4: Know the Difference. URL: https://fireflies.ai/blog/gpt3-vs-4 (дата звернення: 08.04.2025).

20. Microsoft оновлює Copilot AI за допомогою GPT-4 Turbo. URL: https://proit.ua/maikrosoft-znachno-onovliuie-copilot-ai-za-dopomoghoiu-gpt-4-turbo/ (дата звернення: 08.04.2025).

21. Jhanjhi N. Z., Humayun M., Almuayqil S. N. Cyber security and privacy issues in industrial internet of things. *Computer Systems Science & Engineering*. 2021. vol. 37 No. 3.

22. Chen M., et al. The Impact of Large Language Models on Dialogue Systems: A Review. URL: https://arxiv.org/abs/2304.09842 (дата звернення: 08.04.2025)

23. Bommasani R., et al. On the Opportunities and Risks of Foundation Models. URL: https://crfm.stanford.edu/report.html (дата звернення: 08.04.2025).

24. Zhang T., Kishore V., Wu F., et al. BERTScore: Evaluating Text Generation with BERT. URL: https://aclanthology.org/2020.acl-main.647/ (дата звернення: 08.04.2025).

25. Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a method for automatic evaluation of machine translation. URL: https://aclanthology.org/P02-1040/ (дата звернення: 08.04.2025).

26. Lin C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. *Text summarization branches out* 2004. July C. 74-81

27. Xie Q., Yang Z., Parmar N. The asymmetric Otto engine: frictional effects on performance bounds and operational modes. 2023. (Препринт. arXiv; 2310.06512).

28. Bommasani R., Hudson D., Adcock A. On the Opportunities and Risks of Foundation Models. 2021. (Препринт. arXiv; 2108.07258).

29. Chowdhery A., Narang S., Devlin J. PaLM: Scaling Language Modeling with Pathways. 2023. 113 с. (Препринт. arXiv; 2204.02311).

30. Мороз, С, Шуневич, М. Використання штучного інтелекту в логістичній галузі. *Development Service Industry Management*. 2024. №4. С. 269–275.

31. Пічугіна, Ю., Максимов, О. Використання штучного інтелекту та сучасних інформаційних технологій у міських логістичних системах. *Економіка та суспільство*. 2024. С. 62.

32. Кирлик, Н. «Штучний інтелект» та його використання в логістичних процесах. *Актуальні проблеми економіки*, 2021. 9-10. С. 243-244.

33. Булгакова, О. С., Зосімов, В. В., Поздєєв, В. О. Методи та системи штучного інтелекту: теорія та практика. Херсон: ОЛДІ-ПЛЮС, 2020. 210 с.

34. Лубко, Д. В., Шаров, С. В. Методи та системи штучного інтелекту: навчальний посібник. Мелітополь: ФОП Однорог Т.В. 2019.

35. SAP Extended Warehouse Management. SAP. URL: https://www.sap.com/ukraine/products/scm/extended-warehouse-management.html (дата звернення: 24.02.2025).

36. Дослідження штучного інтелекту в автоматизації. *Journal of Artificial Intelligence*. URL: https://doi.org/10.15407/jai2023.03.064 (дата звернення: 26.02.2025).

37. Warehouse Science (навчальний посібник). URL: https://www.warehouse-science.com/book/editions/wh-sci-0.98.1.pdf (дата звернення: 29.02.2025).

38. Штучний інтелект у містах. Nature.com. URL: https://www.nature.com/articles/s41598-025-92283-3 (дата звернення: 10.04.2025).

39. Moufaddal M., Benghabrit A., Bouhaddou I. A Cyber-Physical Warehouse Management System Architecture in an Industry 4.0 Context. *Procedia Computer Science*, 2020, Vol. 170, P. 1236-1243.

40. Unhelkar B. та ін. Enhancing supply chain performance using RFID technology and decision support systems in the industry 4.0 A systematic literature review. *Journal of Supply Chain Management*. 2022. Vol. 48. P. 102-115.

41. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Amodei D., Sutskever I. Language models are few-shot learners. *Advances in Neural Information Processing Systems*. 2020. Vol. 33. Pp. 1877-1901.

42. Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language models are unsupervised multitask learners. *OpenAI blog*. 2019.

43. Esfandiari N., Kiani K., Rastgoo R. A Conditional Generative Chatbot using Transformer Model. *International Journal of Advanced Computer Science and Applications*. 2023. Vol. 14(6). Pp. 123-130.

44. Zhang Y., Roller S., Goyal N., Artetxe M., Chen M., Chen S., et al. OPT: Open Pre-trained Transformer Language Models. *Journal of Machine Learning Research*. 2022. Vol. 23. Pp. 1-59.

45. Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., et al. PaLM: Scaling Language Modeling with Pathways. *Proceedings of the 39th International Conference on Machine Learning*. 2022. Pp. 2474-2483.

46. Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language Models are Unsupervised Multitask Learners. *OpenAI Technical Report*. 2019.

47. Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT. *Proceedings of the 8th International Conference on Learning Representations.* 2020.

48. Chen M., Tworek J., Jun H., Yuan Q., de Oliveira Pinto H. P., Kaplan J., et al. Evaluating Large Language Models Trained on Code. *Transactions on Machine Learning Research*. 2021. Vol. 2(1). Pp. 1-17.

49. Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., Dhariwal P., Amodei D., Sutskever I. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems.* 2020. Vol. 33. Pp. 1877-1901.

50. Kumar R., Sharma P. Transformer-based chatbot architecture for enhanced customer interaction. *International Journal of Artificial Intelligence Research*. 2024. Vol. 8(2). Pp. 45-58.

51. Lee S., Kim J. Implementing ChatGPT for automated customer support: a case study. *Journal of Intelligent Systems*. 2023. Vol. 32(4). Pp. 210-225.

52. Nguyen T. H., Tran L. M. Evaluating the performance of transformer-based models in chatbot applications. *Computational Linguistics and Applications*. 2023. Vol. 14(1). Pp. 89-102.

53. Patel A., Desai M. Comparative analysis of transformer models in conversational AI. *Journal of Machine Learning and Data Mining*. 2024. Vol. 12(3). Pp. 150-165.

54. Singh R., Gupta N. Enhancing chatbot responsiveness using transformer architectures. *International Journal of Computer Applications*. 2023. Vol. 175(7). Pp. 25-33.

55. Zhao L., Wang Y. ChatGPT in education: opportunities and challenges. *Educational Technology & Society*. 2024. Vol. 27(2). Pp. 134-145.

56. Ahmed S., Khan M. Transformer-based models for multilingual chatbot development. *Language Resources and Evaluation*. 2023. Vol. 57(1). Pp. 77-92.

57. Chen X., Li H. Integrating ChatGPT into healthcare chatbots: a review. *Journal of Medical Systems*. 2024. Vol. 48(1). Pp. 12-24.

58. Garcia M., Torres A. Ethical considerations in deploying transformer-based chatbots. *AI & Society*. 2023. Vol. 38(3). Pp. 567-580.

59. Yamada K., Sato T. Real-time response generation in chatbots using transformer models. *IEEE Transactions on Neural Networks and Learning Systems*. 2024. Vol. 35(5). Pp. 1234-1245.

60. Kohli R., Sinha D. K., Kumar G. N. K., Alfurhood B. S., Gupta R. Architectural design of a chatbot used for artificial intelligence with NLP classification using deep learning. *International Journal of Intelligent Systems and Applications in Engineering*. 2022. Vol. 10(3s). Pp. 129-135.

61. Smutny P., Bojko M. Comparative analysis of chatbots using large language models for web development tasks. *Applied Sciences*. 2024. Vol. 14(21). Pp. 172-205.

62. Li Y. Enhancing chatbot responses based on natural language processing techniques. *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence.* 2023. Pp. 574-579.

63. Hanji S. V., Navalgund N., Ingalagi S., Desai S., Hanji S. S. Adoption of AI chatbots in travel and tourism services. In: *Proceedings of the International Congress on Information and Communication Technology*. 2023. Pp. 713-727.

64. Bălan C. Chatbots and voice assistants: digital transformers of the company customer interface a systematic review of the business research literature. *Journal of Theoretical and Applied Electronic Commerce Research*. 2023. Vol. 18(2). Pp. 995-1019.

65. Bird J. J., Ekárt A., Faria D. R. Chatbot interaction with artificial intelligence: human data augmentation with T5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*. 2023. Vol. 14. Pp. 3129-3144.

66. Ding Y., Najaf M. Interactivity, humanness, and trust: a psychological approach to AI chatbot adoption in e-commerce. *BMC Psychology*. 2024. Vol. 12. Pp. 20-56.

67. Rahman S. H., Bahadur M. I., Sufian A., Hasan R. H., Nabil A. MediBERT: a medical chatbot built using KeyBERT, BioBERT and GPT-2. *International Journal of Intelligent Systems and Applications*. 2023. Vol. 15(4). Pp. 45–52.

68. Borna S., Gomez-Cabello C. A., Pressman S. M., Haider S. A., Sehgal A., Leibovich B. C., Cole D., Forte A. J. Comparative analysis of artificial intelligence virtual assistant and large language models in post-operative care. *European Journal of Investigation in Health, Psychology and Education*. 2024. Vol. 14(5). Pp. 1413-–1424.

69. Shams G., Kim K. K., Kim K. Enhancing service recovery satisfaction with chatbots: the role of humor and informal language. *International Journal of Hospitality Management*. 2024. Vol. 120. Pp. 113-125.

70. Larsen A. G., Følstad A. The impact of chatbots on public service provision: a qualitative interview study with citizens and public service providers. *Government Information Quarterly*. 2024. Vol. 41. Pp. 113-125.

71. Dongbo M., Miniaoui S., Fen L., Althubiti S. A., Alsenani T. R. Intelligent chatbot interaction system capable for sentimental analysis using hybrid machine learning algorithms. *Information Processing & Management*. 2023. Vol. 60. Pp. 136-145.

72. Khennouche F., Elmir Y., Himeur Y., Djebari N., Amira A. Revolutionizing generative pre-traineds: insights and challenges in deploying ChatGPT and generative chatbots for FAQs. *Expert Systems with Applications*. 2024. Vol. 246. Pp. 123-140.

73. Hanji S. V., Navalgund N., Ingalagi S., Desai S., Hanji S. S. Adoption of AI chatbots in travel and tourism services. In: *Proceedings of the International Congress on Information and Communication Technology*. 2023. Pp. 713–727

74. Bhoir S. V., Patil S. R., Mogul I. Y. Person-based automation with artificial intelligence chatbots: a driving force of Industry. *Artificial Intelligence and Industry* 2022. Pp. 215–244.

75. Russell S. J., Norvig P. Artificial Intelligence. *A Modern Approach*. 2020. Pp. 20–24.

76. Domingos P. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World. *Basic Books*. 2015. Vol. 6(2). Pp. 45-47.

77. Marcus G., Davis E. Building Artificial Intelligence We Can Trust. *Rebooting AI*. 2019. Pp. 346–354.

78. Kissinger H., Schmidt E., Huttenlocher D. The Age of AI: And Our Human Future. *Little, Brown and Company*. 2021. Pp. 356–361.

79. Summerfield C. These Strange New Minds. *Oxford University Press*. 2025. Pp. 159–163.

80. Harari Y. N. A Brief History of Information Networks from the Stone Age to AI. *Nexus*. 2024. Pp. 25–31.

81. Марчук О.О. VI Науково-практична конференція *Метод та система трансформера на підставі chatgpt для генерації текстових запитів чат-боту*: тези доп. всеукр. наук.-практ. конф. (м. Київ, 13 груд. 2024 р.). Київ, 2024. С. 34–35.

ДОДАТОК А
(обов'язковий)

ПУБЛІКАЦІЯ

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ

ІНСТИТУТ ПРОБЛЕМ МОДЕЛЮВАННЯ
В ЕНЕРГЕТИЦІ ІМ. Г.Є. ПУХОВА

МАТЕРІАЛИ
VI НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ
«БЕЗПЕКА ЕНЕРГЕТИКИ В ЕПОХУ ЦИФРОВОЇ ТРАНСФОРМАЦІЇ»

13 грудня 2024 року

Київ – 2024

УДК [621.3+620.9]::[004[056.53+42+94] + 504.06]
ББК 31
Б-39

Рекомендовано до друку Вченою радою Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України (протокол № 12 від 28 листопада 2024 р.)

Б-39 **Безпека енергетики** в епоху цифрової трансформації, VI науково-практична конференція Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України : матеріали (Київ, 13 грудня 2024 р.). Київ : ІПМЕ ім. Г.Є.Пухова НАН України, 2024. 191 с.

B-39 **Energy security** in the digital transformation era, VI scientific-practical conference of the G.E. Pukhov Institute for Modeling in Energy Engineering National Academy of Sciences of Ukraine : materials (Kyiv, December 13, 2024). Kyiv: PIMEE NAS of Ukraine, 2024. 191 p.

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ МОДЕЛЮВАННЯ В ЕНЕРГЕТИЦІ
ім. Г.Є. ПУХОВА НАН УКРАЇНИ

МАТЕРІАЛИ
VI НАУКОВО-ПРАКТИЧНОЇ КОНФЕРЕНЦІЇ

БЕЗПЕКА ЕНЕРГЕТИКИ В ЕПОХУ ЦИФРОВОЇ ТРАНСФОРМАЦІЇ

13 грудня 2024 року

м. Київ

2024

Вельмишановний учасник _____

_____

Запрошуємо Вас прийняти участь в роботі VI науково-практичної конференції «Безпека енергетики в епоху цифрової трансформації», яка буде проходити 13 грудня 2024 року в Інституті проблем моделювання в енергетиці ім. Г.Є. Пухова Національної академії наук України (м. Київ).

ОРГАНІЗАТОРИ КОНФЕРЕНЦІЇ

Інститут проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України
(м. Київ)

ПРОГРАМНИЙ КОМІТЕТ

Мохор Володимир Володимирович
член-кореспондент НАН України, доктор технічних наук, професор,
директор Інституту, голова програмного комітету

Чемерис Олександр Анатолійович
доктор технічних наук, професор,
заступник директора з наукової роботи

Артемчук Володимир Олександрвич
доктор технічних наук,
заступник директора з науково-організаційної роботи

Чьочь Вікторія Володимирівна
кандидат технічних наук,
заступник директора з науково-технічної роботи

ОРГАНІЗАЦІЙНИЙКОМІТЕТ

Артемчук Володимир Олександрвич
доктор технічних наук,
заступник директора з науково-організаційної роботи

Клименко Тетяна Михайлівна
завідувачка науково-організаційного відділу

Цуркан Оксана Володимирівна
молодший наковий співробітник

О.О. Марчук

МЕТОД ТА СИСТЕМА ТРАНСФОРМЕРА НА ПІДСТАВІ CHATGPT ДЛЯ ГЕНЕРАЦІЇ ТЕКСТОВИХ ЗАПИТІВ ЧАТ-БОТУ

Розглянуто прикладні аспекти розробки метод та система трансформера на підставі chat-GPT для генерації текстових запитів чат-боту, які базуються на попередніх відповідях та складаються максимально точно до людської мови у реальному часі. Запропонована система забезпечує точну і швидку генерацію тексту відповідно до складеного запиту користувача.

Applied aspects of developing a method and system of a transformer based on chat-GPT for the generation of text requests of a chatbot, which are based on previous answers and are composed as closely as possible to human speech in real time, are considered. The proposed system provides accurate and fast text generation according to the user's request.

Сучасні чат-боти набули популярності завдяки своїй здатності вести природні та зручні для користувачів розмови. Основою для цього є архітектура трансформерів, зокрема ChatGPT, яка дозволяє обробляти текстові запити в реальному часі з врахуванням попередніх відповідей та адаптацією під стиль людської мови. Такий підхід дозволяє створювати чат-боти, які можуть легко інтегруватися у різні сфери бізнесу, забезпечуючи покращення взаємодії з користувачами [1].

Метою роботи є розробка методології генерації текстових запитів на базі трансформерів ChatGPT, що дозволить чат-боту точно та швидко реагувати на запити користувачів, враховуючи попередні відповіді та особливості природної мови. Це включає оптимізацію генерації запитів і поліпшення механізму утримання контексту у тривалих діалогах.

ChatGPT, побудований на архітектурі трансформерів, використовує механізм самозвернення для обробки тексту, який дозволяє визначати значення кожного слова в контексті попередніх фраз. Архітектура передбачає багаторівневу обробку тексту з використанням self-attention, що дозволяє моделі запам'ятовувати деталі діалогу та забезпечувати більш релевантні відповіді. Такий підхід дозволяє покращити точність і природність генерації тексту, адаптуючись до стилю розмови [2].

Основні компоненти системи трансформера на основі ChatGPT:

- Механізм самозвернення (self-attention). Застосування self-attention дозволяє моделі ChatGPT відслідковувати залежності між словами у тексті та утримувати послідовність в рамках довгого діалогу. Це забезпечує здатність чат-боту підтримувати контекст, враховуючи як нові, так і попередні запити користувача.
- Політика обробки діалогів. Для кожного запиту система адаптує відповіді на основі збережених попередніх контекстів. Це дозволяє моделі генерувати відповіді, які є релевантними до поточного етапу розмови, зберігаючи при цьому цілісність діалогу.

• Автоматизована генерація запитів. Важливим компонентом є функція, яка дозволяє автоматично генерувати текстові запити відповідно до інтеракцій, що виникають у розмові. Вона зменшує кількість помилок і забезпечує стабільність у випадках, коли чат-бот виконує тривалі розмови з користувачем.

Основні процеси системи генерації тексту включають узгодження відповідей, корекцію помилок, адаптацію стилю та інші функції, що дозволяють чат-боту вести послідовні розмови. Обробка запитів включає наступні етапи:

1. Узгодження. Підтримання узгодженості даних є ключовою частиною процесу. На кожному етапі діалогу здійснюється перевірка попередніх відповідей, щоб забезпечити точну обробку нових запитів.
2. Адаптація до стилю. Використовуючи трансформери, система автоматично налаштовується під стиль мовлення користувача, забезпечуючи точність та зручність взаємодії.
3. Автоматичне коригування помилок. Задля покращення результатів у системі запроваджено процеси, що виявляють та усувають можливі помилки у формулюваннях текстових запитів.
4. Реплікація. В процесі взаємодії система створює репліки, які можна використовувати для збереження та перевірки відповідей.

Ця система автоматизації генерації тексту дозволяє чат-ботам на базі ChatGPT швидко та природно реагувати на потреби користувача, забезпечуючи при цьому високу точність. Подальші дослідження спрямовані на розширення можливостей інтеграції з іншими платформами та покращення управління контекстом у тривалих розмовах, що дозволить підвищити ефективність чат-ботів у різних сферах [3, 4].

Отже, запропонована методологія і система трансформера на основі ChatGPT для генерації текстових запитів чат-боту забезпечує точну і швидку побудову відповідей на основі попереднього контексту. Подальші дослідження спрямовані на покращення механізмів адаптації під стиль мовлення користувачів та вдосконалення автоматизації обробки запитів, що відкриває можливості для інтеграції цієї системи в різні цифрові платформи.

1. OpenAI. "GPT-3 and the Transformer architecture.": https://openai.com/research/gpt-3.
2. Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems (2017).
3. Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).
4. Devlin, Jacob, et al. "BERT: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

ДОДАТОК Б
(обов'язковий)

ПРЕЗЕНТАЦІЯ

Задачі дослідження

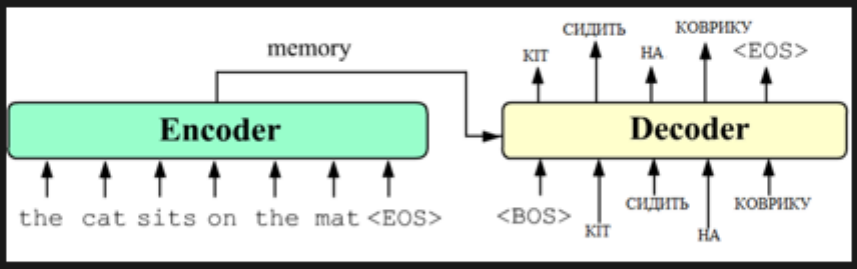
- проаналізувати існуючі підходи до генерації текстових запитів у чат-ботах та визначити їхні обмеження;
- обґрунтувати вибір архітектури трансформера як базової технології для генерації запитів;
- розробити метод генерації текстових запитів на основі ChatGPT з урахуванням контексту діалогу та предметної області;
- побудувати математичну модель системи генерації запитів та визначити ключові метрики оцінювання якості;
- створити прототип системи генерації текстових запитів та провести її експериментальне тестування;
- виконати порівняльний аналіз запропонованого методу із традиційними підходами та сучасними системами генерації запитів.

НАУКОВА НОВИЗНА ТА ПРАКТИЧНА ЦІННІСТЬ ОТРИМАНИХ РЕЗУЛЬТАТІВ

- Удосконалено метод генерації текстових запитів для чат-ботів шляхом поєднання трансформерної моделі ChatGPT із механізмами контекстного зважування реплік та інтеграції зовнішніх джерел знань.
- Створенно адаптивне програмне рішення для автоматизованої діалогової взаємодії, здатного працювати в різних предметних галузях. Запропонована система підвищує якість обслуговування у клієнтських сервісах, освіті, правових консультаціях і державних онлайн-платформах.

Основні частини трансформера

Енкодер відповідає за обробку вхідної послідовності та перетворення її у внутрішнє представлення



Декодер використовує це представлення для генерації вихідної послідовності.

Порівняння існуючих підходів до генерації текстових запитів у чат-ботах

Параметр	GPT-моделі	Правилозалежні системи	Гібридні системи
Гнучкість	Висока	Низька	Середня
Якість обробки нового контенту	Висока	Низька	Висока
Залежність від попереднього налаштування	Низька	Висока	Середня
Вартість впровадження	Висока	Низька	Середня
Проблеми з галюцинаціями	Присутні	Відсутні	Часткові

Метод генерації запитів з урахуванням контексту діалогу та предметної області

1. $Q_t=G(C_t,H_t,K)$ - генерація текстового запиту у чат-боті як функція

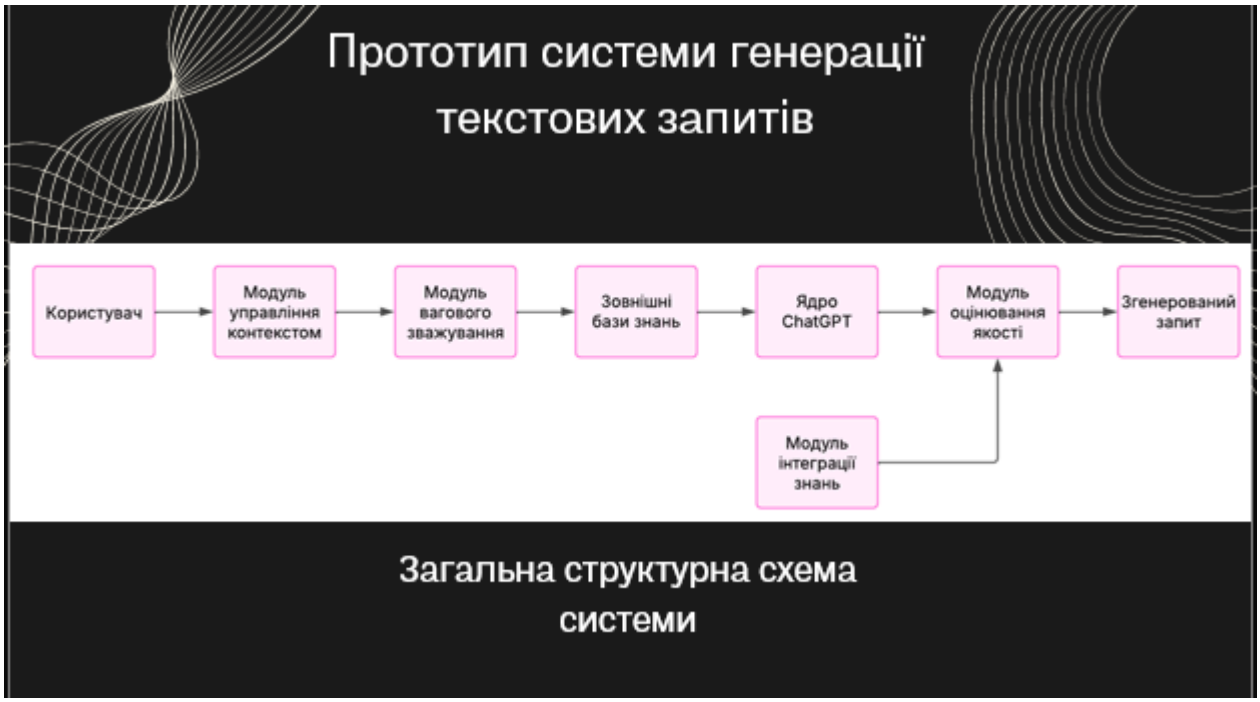
Контекст діалогу у вигляді впорядкованої множини – 2. $C_t=\{u_1,r_1,u_2,r_2,\ldots,u_{(t-1)},r_{(t-1)},\}$

3. $w_i=\alpha \cdot rel(u_i,Q_t)+\beta \cdot time\_decay(i,t),$ – Адаптивна вага контекстного елемента

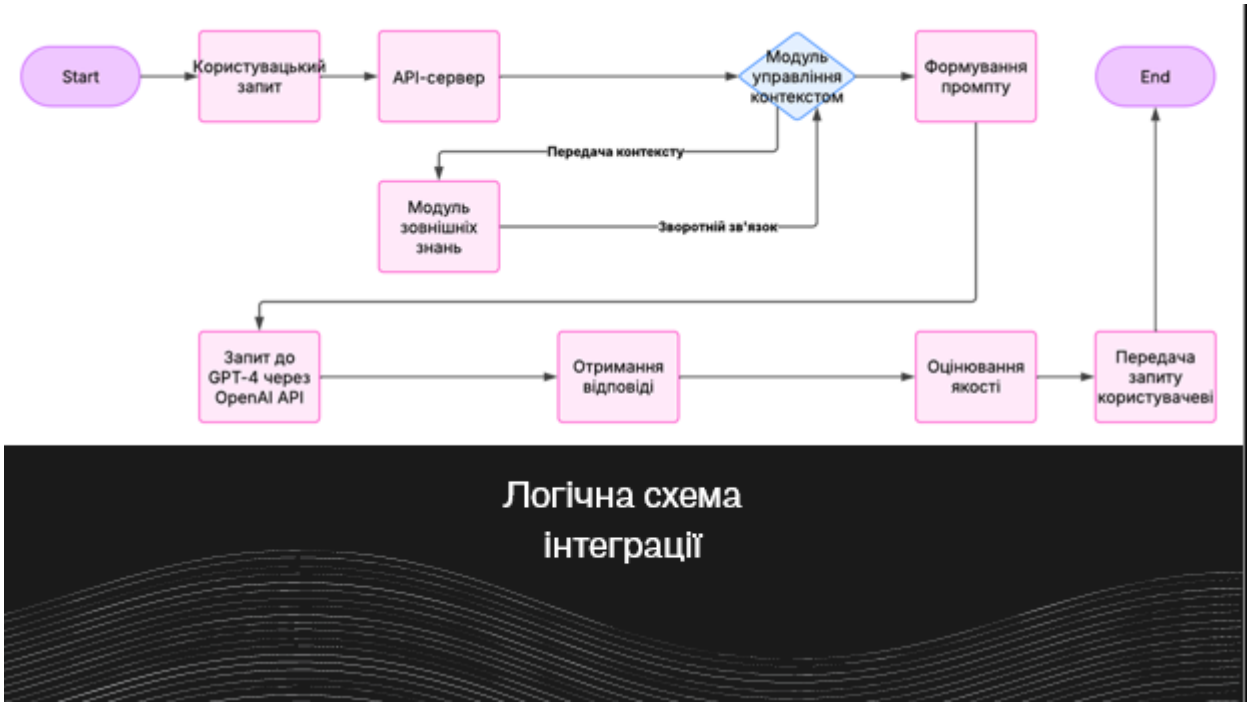
Метод генерації запитів з урахуванням контексту діалогу та предметної області у вигляді функції – 4. $Q_t=G(G_t,H_t,K,W_t)$

Ключові метрики оцінювання якості

Метрика	Основний принцип	Переваги	Обмеження
Perplexity (PPL)	Оцінка здатності моделі прогнозувати наступне слово	Відображає загальну якість мовної моделі	Не враховує змістовну релевантність
BLEU	Порівняння з референтними текстами	Висока об'єктивність у формалізованих сценаріях	Погано працює з варіативними текстами
ROUGE	Аналіз збігів на рівні фраз та слів	Ефективна для підсумкових текстів	Не враховує глибинну семантику
Людська оцінка	Експертний або користувацький аналіз якості тексту	Враховує всі аспекти якості	Суб'єктивність та витрати часу



Загальна структурна схема системи



Логічна схема інтеграції

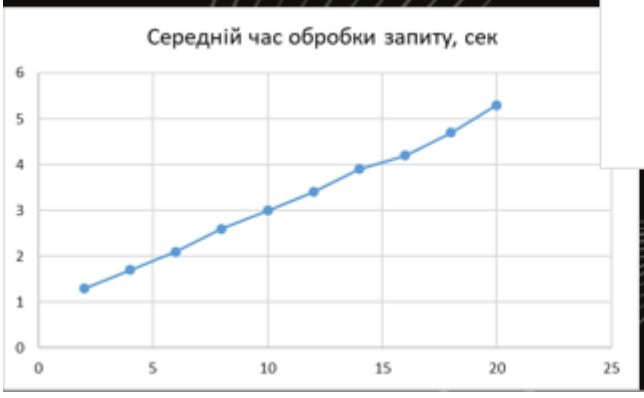

Блок-схема алгоритму
Start
Запит користувача
Попередня обробка
Оновлення контексту
Формування промпту
Виклик GPT-4
Інтеграція зовнішніх знань
Оцінювання якості
Формування відповіді
Передача користувачу
Логування діалогу
End


Узагальнена схема оптимізованої взаємодії
Start
Запит користувача
Перевірка кешу
Оновлення контексту та ваговое зважування
Скорочення промпту та додавання системних інструкцій
Запит до GPT-4
Оцінка якості, збереження у кеш
Користувач
Формування та відправлення відповіді
End

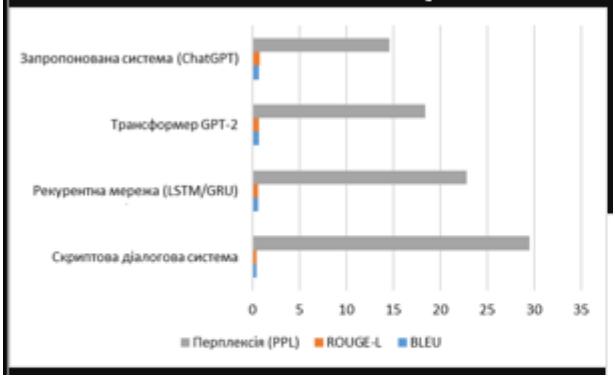
Метрики ефективності роботи

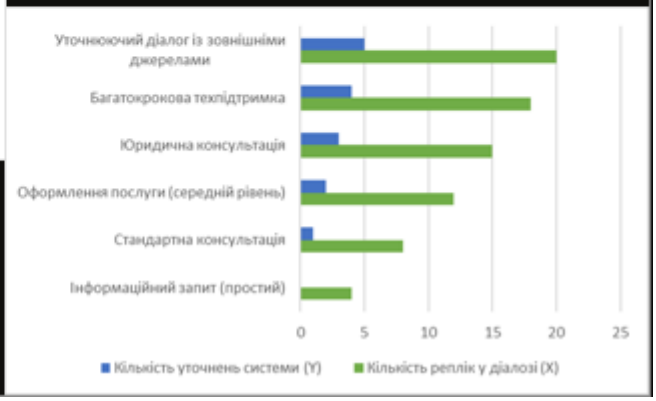
Показник	Опис
Середній час відповіді	Середній час від моменту отримання запиту до відправлення відповіді користувачу.
Відсоток відповідей з кешу	Частка відповідей, отриманих без звернення до GPT-4.
Довжина промптів	Середня кількість токенів у промптах до моделі.
Частка повторних запитів	Кількість схожих або ідентичних запитів від одного користувача.
Частота помилок	Відсоток невдалих звернень до GPT-4 через таймаути або некоректні промпти.

Порівняльний аналіз



Порівняльний аналіз





Висновок

У роботі за результатами виконаних теоретичних та практичних досліджень розроблено метод генерації текстових запитів для чат-ботів на основі трансформерної архітектури ChatGPT, який забезпечує формування контекстно-залежних, стилістично коректних та інформаційно релевантних текстів із урахуванням багатокрокових діалогів та інтеграції зовнішніх джерел знань.

ДОДАТОК В
(обов'язковий)

ЛІСТИНГ КОДУ ТЕЛЕГРАМ-БОТУ, ЯКИЙ РЕАЛІЗУЄ СИСТЕМУ ГЕНЕРАЦІЇ ТЕКСТОВИХ ЗАПИТІВ НА ОСНОВІ CHATGPT

```
import openai
import asyncio
import databases
import sqlalchemy
from fastapi import FastAPI, Depends
from pydantic import BaseModel
from typing import List, Dict
import redis
import uvicorn
import nltk
import torch
from transformers import GPT2Tokenizer, GPT2LMHeadModel
from sentence_transformers import SentenceTransformer
import elasticsearch
from rouge_score import rouge_scorer
from nltk.translate.bleu_score import sentence_bleu
from bert_score import score as bert_score
from aiogram import Bot, Dispatcher, types
from aiogram.types import Message
from aiogram.utils import executor

# Конфігурація API та бази даних
DATABASE_URL =
"postgresql://user:password@localhost/chatbot_db"
database = databases.Database(DATABASE_URL)
metadata = sqlalchemy.MetaData()
engine = sqlalchemy.create_engine(DATABASE_URL)

# Таблиця для контексту діалогів
chats = sqlalchemy.Table(
```

```
        "chats", metadata,
        sqlalchemy.Column("id",                    sqlalchemy.Integer,
primary_key=True),
        sqlalchemy.Column("user_id",               sqlalchemy.String,
index=True),
        sqlalchemy.Column("context", sqlalchemy.String),
        sqlalchemy.Column("timestamp", sqlalchemy.DateTime)
    )
    metadata.create_all(engine)

    # Ініціалізація FastAPI
    app = FastAPI()
    redis_client = redis.Redis(host='localhost', port=6379, db=0)

    # Модель для аналізу семантичної схожості
    sentence_model   =   SentenceTransformer("paraphrase-mpnet-base-
v2")

    # OpenAI API ключ
    openai.api_key = "YOUR_OPENAI_API_KEY"

    # Функція обробки контексту
    def get_weighted_context(user_id: str) -> str:
        query     =     chats.select().where(chats.c.user_id     ==
user_id).order_by(chats.c.timestamp.desc()).limit(5)
        history = asyncio.run(database.fetch_all(query))
        weighted_context = ""
        for row in history:
            relevance = torch.cosine_similarity(
                sentence_model.encode(row["context"],
convert_to_tensor=True),
                sentence_model.encode(history[-1]["context"],
convert_to_tensor=True),
                dim=0
            )
```

```
            if relevance > 0.5:
                weighted_context += row["context"] + "\n"
        return weighted_context

    # Функція генерації відповіді GPT-4
    def generate_response(user_id: str, message: str) -> str:
        context = get_weighted_context(user_id)
        prompt = f"{context}User: {message}\nChatbot:"
        response = openai.ChatCompletion.create(
            model="gpt-4",
            messages=[{"role": "user", "content": prompt}],
            temperature=0.7,
            max_tokens=150
        )
        return response["choices"][0]["message"]["content"].strip()

    # Функція оцінки якості відповіді
    def evaluate_response(reference: str, generated: str):
        bleu_score = sentence_bleu([reference.split()],
generated.split())
        scorer = rouge_scorer.RougeScorer(['rouge1', 'rouge2',
'rougeL'], use_stemmer=True)
        rouge_scores = scorer.score(reference, generated)
        P, R, F1 = bert_score([generated], [reference], lang="en")
        return {
            "bleu": bleu_score,
            "rouge1": rouge_scores['rouge1'].fmeasure,
            "rouge2": rouge_scores['rouge2'].fmeasure,
            "rougeL": rouge_scores['rougeL'].fmeasure,
            "bert_score": F1.mean().item()
        }

    # API-ендпойнт для отримання відповіді
    class MessageModel(BaseModel):
        user_id: str
```

```
    message: str

@app.post("/chat")
async def chat(message: MessageModel):
    response = generate_response(message.user_id,
message.message)
    return {"response": response}

# Телеграм-бот
TELEGRAM_BOT_TOKEN = "YOUR_TELEGRAM_BOT_TOKEN"
bot = Bot(token=TELEGRAM_BOT_TOKEN)
dp = Dispatcher(bot)

@dp.message_handler(commands=['start'])
async def send_welcome(message: Message):
    await message.reply("Вітаю! Ви можете задати будь-яке
питання.")

@dp.message_handler()
async def handle_message(message: Message):
    user_id = str(message.from_user.id)
    response = generate_response(user_id, message.text)
    await message.reply(response)

# Запуск сервера та бота
if __name__ == "__main__":
    import threading
    threading.Thread(target=lambda: uvicorn.run(app,
host="0.0.0.0", port=8000)).start()
    executor.start_polling(dp, skip_updates=True)
```

ДОДАТОК Г
(обов'язковий)

УЗАГАЛЬНЕНА ТАБЛИЦЯ ВИБРАНИХ ІНСТРУМЕНТІВ

Таблиця Г.1 – Узагальнена таблиця вибраних інструментів

Компонент	Обрані інструменти	Функціональне призначення
Мова програмування	Python	Основна мова для розробки системи, інтеграції модулів та обробки текстових даних.
Модель	GPT-4 (OpenAI API)	Генерація текстових запитів на основі контексту діалогу та зовнішніх даних.
NLP-бібліотеки	Transformers, spaCy, nltk, SentenceTransformers	Лінгвістичний аналіз тексту, токенізація, обчислення семантичної близькості реплік.
База даних	SQLite / PostgreSQL	Збереження історії діалогів, мета-даних, вагових коефіцієнтів та інших структурованих даних.
Кеш	Redis	Тимчасове збереження проміжних даних та оперативний доступ до контексту діалогу у реальному часі.
Пошук зовнішніх знань	Elasticsearch, DuckDuckGo API	Пошук релевантних фрагментів знань та отримання актуальних фактів з зовнішніх джерел.
Спеціалізовані джерела	Галузеві API (медичні, юридичні, технічні тощо)	Підключення до вузькоспеціалізованих баз знань залежно від предметної області.
Метрики якості	sacreBLEU, rouge-score, bert_score, Transformers (Perplexity)	Оцінювання якості згенерованих текстів: лексична точність, змістовна повнота, семантична відповідність та впевненість моделі.
Компонент	Обрані інструменти	Функціональне призначення

Серверна частина	FastAPI / Flask	Реалізація REST API для обробки запитів користувачів, інтеграція з GPT та зовнішніми системами.
Інтерфейси	Telegram Bot API, Viber API, WhatsApp API	Інтерактивна взаємодія з користувачами через популярні месенджери.
Моніторинг	Prometheus, Grafana	Збір, моніторинг та візуалізація продуктивності системи у реальному часі.
Візуалізація	Matplotlib, Plotly	Побудова графіків для аналізу статистики, якості та продуктивності системи.
Середовища розробки	PyCharm, Jupyter Notebook	Розробка основного коду системи та аналітичних експериментів, тестування та налаштування параметрів.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

РЕЦЕНЗІЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ МАГІСТРА

Здобувач: МАРЧУК Олександр Олександрович

Тема: Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті

Спеціальність: 123 «Комп'ютерна інженерія»

Обсяг кваліфікаційної роботи магістра:

Кількість листів креслень _–_; кількість сторінок записки _83_

1. Короткий зміст роботи та прийнятих рішень _У роботі запропоновано метод та система трансформера на підставі ChatGPTдля генерації текстових запитів чат-боті

2. Висновок про відповідність роботи дипломному завданню _____
Кваліфікаційна робота магістра відповідає виданому завданню

3. Характеристика виконання кожного розділу, ступінь використання останніх досягнень науки і техніки і передових методів роботи: _ У першому розділі виконано аналіз сучасних підходів до генерації текстових запитів у чат-ботах, проаналізовано особливості скриптових діалогових систем, нейронних мереж типу LSTM/GRU та трансформерних моделей. У другому розділі розроблено метод генерації текстових запитів для чат-ботів, що включає контекстне зважування реплік, інтеграцію зовнішніх баз знань та формування уточнюючих запитів. У третьому розділі розроблено архітектуру та програмний прототип системи генерації текстових запитів для чат-ботів на основі запропонованого методу. У четвертому розділі проведено експериментальне дослідження роботи системи, виконано порівняльний аналіз із традиційними та нейромережевими діалоговими системами, оцінено якість згенерованих текстів за метриками перплексії, BLEU та ROUGE.

4. Позитивні сторони роботи: Запропонований метод може бути рекомендований для використання у компаніях, що впроваджують інтелектуальні

системи обслуговування клієнтів, а також у державних установах для створення адаптивних чат-ботів у системах е-урядування.

5. Негативні сторони роботи: У роботі наявні незначні недоліки, зокрема окремі стилістичні неточності та потреба в ширшому аналізі практичного застосування розробленого методу, що, однак, не знижує загальної якості дослідження.

6. Оцінка графічного оформлення та пояснювальної записки роботи: відсутній

7. Відгук про роботу в цілому: Робота виконана на належному науково-технічному рівні.

8. Інші зауваження: Відсутні

9. Оцінка кваліфікаційної роботи магістра:

Розглянувши позитивні та негативні сторони представленої кваліфікаційної роботи магістра вважаю, що робота заслуговує оцінки «добре» 4.00 (С)

Рецензент (прізвище, ім'я, по батькові, посада, місце роботи) д.т.н., професор, Мартинюк В.В., завідувач кафедри автоматизації, комп'ютерно-інтегрованих технологій та робототехніки

"19" травня 2025 р. (підпис)

Wed Apr 30 18:01:48 EEST 2025, Медзатий Дмитро Миколайович, Хмельницький національний університет, ХНУ

Anti-Plagiarism v-15.274 Educational

The maximum coincidence with one document 25.0%

Dictionaries check: en_US, ru_RU, ua_UA. **Errors in the documents: 9%**

ID: 240683 Title: МКР Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті Added in a DB: 2025-04-30 Authors: Марчук О.О. Heads: Федоров Є.Є. Consultants: Opponents:	Document		Sum coincidence on the DB	
	Symbols	Lexemes	Symbols	Lexemes
	152704	933	38628 (25%)	243 (26%)

Plagiarism sources

ID	Description	Plagiarism presence in the document	
		Symbols	Lexemes
193103	Title: Звіт з ПДП Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті Added in a DB: 2025-03-21 Authors: О. О. Марчука Heads: Федоров Є.Є. Consultants: Opponents:	38102 (25.0%)	232 (25.0%)

Протокол аналізу звіту подібності експертом

Заявляю, що я ознайомився (-лась) з Повним звітом подібності, який був згенерований Системою виявлення і запобігання плагіату щодо роботи:

Автор: Марчук О.О.

Співавтор:

Назва: Марчук_Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті

Експерт:

Підрозділ: Кафедра комп'ютерної інженерії та інформаційних систем

Коефіцієнт подібності 1:1.9%

Коефіцієнт подібності 2:0.1%

Мікропробіли: 0

Заміна букв: 0

Інтервали: 0

Білі знаки: 1

Дата створення звіту: 2025-04-30 17:07:46.0

Після аналізу Звіту подібності констатую наступне:

☒ **Запозичення, виявлені в роботі є законними і не є плагіатом. Рівень подібності не перевищує допустимої межі. Таким чином робота незалежна і приймається.**

☐ **Запозичення не є плагіатом, але перевищено граничне значення рівня подібностей. Таким чином робота повертається на доопрацювання.**

☐ **Виявлено запозичення і плагіат або навмисні текстові спотворення (маніпуляції), як передбачувані спроби укриття плагіату, які роблять роботу невідповідною вимогам законодавства (Ст. 32. ЗУ Про вищу освіту, пункт 3.1, Ст. 42. ЗУ Про освіту) та вимог НАЗЯВО (Критерій 5), а також кодексу етики і процедурам. Таким чином робота не приймається.**

Обгрунтування:

2025-04-30 Доцент Андрій Нічепорук

Дата експерт

Завідувачу кафедри КІІС
доктору філософії, доценту
Ользі ПАВЛОВІЙ

Марчук Олександр Олександрович
ПІБ здобувача вищої освіти

ФІТ, 2 курсу, групи КІ2м-23-2

ЗАЯВА

З правилами чинного Положення «Про систему забезпечення академічної доброчесності у Хмельницькому національному університеті» від 01.07.2022, згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування заходів дисциплінарної та академічної відповідальності, ознайомлений (а). Про використання програмно-технічних засобів для перевірки кваліфікаційних робіт здобувачів вищої освіти на плагіат оповіщений(а) та надаю свою згоду на обробку та збереження університетом моєї роботи в інституційному репозитарії університету.

Також надаю університету право на передачу моєї роботи для обробки та збереження в базах даних програмно-технічних засобів (StrikePlagiarism та Anti-Plagiarism) та використання роботи для виявлення плагіату в інших роботах, які перевіряються програмно-технічними засобами та користувачами, що мають доступ до цих програмно-технічних засобів, виключно в обмежених цілях для виявлення плагіату в текстах робіт.

Робота для перевірки університетом надається в друкованому та електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

29.04.2025
дата

підпис

РІШЕННЯ ЕКСПЕРТНОЇ КОМІСІЇ

КАФЕДРИ КОМП'ЮТЕРНОЇ ІНЖЕНЕРІЇ ТА ІНФОРМАЦІЙНИХ СИСТЕМ

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті

Автор: Марчук Олександр Олександрович

Спеціальність: 123 – Комп'ютерна інженерія

Освітня програма: освітньо-наукова

Науковий керівник: Федоров Євген Євгенович, д.т.н, професор

Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом. Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) системи перевірки виявили збіги з іншими документами в частині стандартних формулювань, структури змісту та назв розділів, що є типовими для кваліфікаційних робіт.;
2) усі запозичення фрагментарні, або мають належним чином оформленні посилання;
3) окремі виявлені збіги є загальновживаними фразами або виразами, про що свідчить посилання системи на збіг з джерелами на фрагмент речення;
4) в якості запозичень в окремих місцях системою зафіксовано послідовності програмного коду, які є вхідними даними до вирішення задач і не можуть розглядатися як об'єкт авторських прав і, відповідно, їх порушення;
5) усі ознаки модифікації тексту, зафіксовані системою, стосуються поєднання латинських символів з україномовними скороченнями індексів у формулах, що не може вважатися зміною самого тексту.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості StrikePlagiarism, складає 1,93% і адресується до 28 першоджерела; та системою Anti-Plagiarism складає 25%, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Євген ФЕДОРОВ

Гарант ОП _____ Олег САВЕНКО

Завідувач кафедри КІІС _____ Ольга ПАВЛОВА

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

ГОЛОВІ ЕКЗАМЕНАЦІЙНОЇ КОМІСІЇ

Направляється студент Марчук Олександр Олександрович (прізвище, ім'я, по батькові) на захист дипломного проєкту (роботи)

за спеціальністю 123 - Комп'ютерна інженерія

На тему: Метод та система трансформера на підставі ChatGPT для генерації текстових запитів чат-боті

Дипломний проєкт (робота), рецензія, довідка про перевірку на плагіат додаються.

Декан факультету _____ (підпис) Тетяна ГОВОРУЩЕНКО (ім'я, прізвище)

ДОВІДКА УСПІШНОСТІ

Марчук О. О. за період навчання на факультеті інформаційних технологій з 2023 по 2025 роки повністю виконав навчальний план спеціальності з таким розподілом оцінок за:
національною шкалою: відмінно 0,00 %, добре 18,18 %, задовільно 81,82 %.
шкалою ЄКТС: A 0,00 %, B 0,00 %, C 10,53 %, D 5,26 %, E 84,21 %.

Методист факультету _____ (підпис) Т. Кононенко (ім'я, прізвище)

ВИСНОВОК КЕРІВНИКА ДИПЛОМНОГО ПРОЄКТУ (РОБОТИ) ТА ОБГРУНТУВАННЯ ОЦІНКИ

Студент _____

Оцінка дипломного проєкту (роботи) добре С

Керівник дипломного проєкту _____ (підпис) Євген Федоров (ім'я, прізвище)

"_____" _____ 2025 р.

ВИСНОВОК КАФЕДРИ ПРО ДИПЛОМНИЙ ПРОЄКТ (РОБОТУ)

Дипломний проєкт (роботу) розглянуто. Студент Марчук О. О. допускається до захисту цього проєкту (роботи) в екзаменаційній комісії.

Завідувач кафедри КІС (назва)

О. Побігова (підпис, ім'я, прізвище)

"19" 05 2025 р.