DOI 10.31891/2307-5732-2024-345-6-5
УДК 004.8

МЕДВЕДЧУК ВІТАЛІЙ
Хмельницький національний університет
https://orcid.org/0009-0005-9661-3251
e-mail: medvedchuk.vitalii@gmail.com
БАГРІЙ РУСЛАН
Хмельницький національний університет
https://orcid.org/0000-0001-5219-1185
e-mail: bahriiro@khmnu.edu.ua
СКРИПНИК ТЕТЯНА
Хмельницький національний університет
https://orcid.org/0000-0002-8531-5348
e-mail: tkskripnik1970@gmail.com
МАЗУРЕЦЬ ОЛЕКСАНДР
Хмельницький національний університет
https://orcid.org/0000-0002-8900-0650
e-mail: exe.chong@gmail.com
МОНАСТИРСЬКА ДАР'Я
Хмельницький національний університет
e-mail: monkadasha@gmail.com

МЕТОД ГЕНЕРАЦІЇ ВІДПОВІДЕЙ З ДОПОВНЮЮЧИМ ІНФОРМАЦІЙНИМ ПОШУКОМ ДЛЯ ДОПОМІЖНОЇ КОМУНІКАЦІЇ

Проблема покращення комунікації для людей з обмеженими можливостями мовлення є надзвичайно важливою у сучасному суспільстві, де технології можуть значно полегшити взаємодію та інтеграцію таких осіб у соціальні процеси. Традиційні методи спілкування, такі як жестова мова чи письмове повідомлення, часто не забезпечують необхідної точності та швидкості комунікації, що призводить до бар'єрів у повсякденному житті. Сучасні технології, такі як великі мовні моделі (LLM) і технологія доповнюючого інформаційного пошуку (RAG), можуть значно покращити ці процеси. Вони дозволяють автоматично генерувати індивідуалізовані текстові відповіді, враховуючи не тільки запит користувача, а й контекст, що робить комунікацію більш точною і швидкою.

У статті пропонується метод генерації відповідей з доповнюючим інформаційним пошуком для допоміжної комунікації, що інтегрує релевантну інформацію з різних джерел, зокрема, історії чату та профілю користувача. Це дозволяє генерувати відповіді, які краще відповідають індивідуальним потребам кожного користувача. Метод передбачає три етапи: завантаження контексту, пошук релевантної інформації та генерацію кількох варіантів відповіді. Особливістю запропонованого методу є здатність враховувати широкий контекст і потреби користувача, що забезпечує високу точність і персоналізацію відповіді.

Проведені експерименти показали високу ефективність методу, зокрема в оцінці точності відповідей, що варіюється в межах 85-95%. Це дозволяє значно покращити комунікацію для людей з обмеженими можливостями мовлення, знижуючи соціальні бар'єри та покращуючи якість їхнього повсякденного життя.

Ключові слова: допоміжна комунікація, генерація відповідей, доповнюючий інформаційний пошук, великі мовні моделі, Retrieval-Augmented Generation, пошукові алгоритми, текстові відповіді, контекст.

MEDVEDCHUK VITALIY, BAHRII RUSLAN, SKRYPNYK TETIANA,
MAZURETS OLEKSANDR, MONASTERY DAR'YA
Khmelnytskyi National University

METHOD OF RESPONSE GENERATION WITH RETRIEVAL-AUGMENTED INFORMATION SEARCH FOR ASSISTIVE COMMUNICATION

The issue of improving communication for people with speech impairments is critically important in modern society, where technology can greatly facilitate interaction and integration of such individuals into social processes. Traditional communication methods, such as sign language or written messages, often fail to provide the necessary accuracy and speed of communication, creating barriers in everyday life. Modern technologies, such as large language models (LLMs) and retrieval-augmented generation (RAG) technology, can significantly enhance these processes. They enable the automatic generation of individualized text responses, considering not only the user's query but also the context, making communication more accurate and faster.

The article proposes a method for response generation with retrieval-augmented information search for assistive communication, which integrates relevant information from various sources, including chat history and user profiles. This allows for generating responses that better address the individual needs of each user. The method consists of three stages: context loading, retrieval of relevant information, and generation of multiple response options. A distinctive feature of the proposed method is its ability to consider extensive context and user needs, ensuring high accuracy and personalization of the response.

Experiments conducted demonstrated the method's high efficiency, particularly in the accuracy of responses, which ranged between 85% and 95%. This significantly improves communication for people with speech impairments, reducing social barriers and enhancing the quality of their daily lives.

Keywords: assistive communication, response generation, retrieval-augmented information search, large language models, Retrieval-Augmented Generation, search algorithms, text responses, context.

Постановка проблеми

Люди з порушеннями мовлення стикаються зі значними бар'єрами у повсякденній комунікації, що

часто призводить до соціальної ізоляції та обмеження їхньої участі у громадському житті [1]. Augmentative and Alternative Communication (AAC) - це поширений спосіб комунікації, який використовується для поповнення або заміни мовлення або письма для тих, хто має обмеження у виробництві або розумінні мовлення [2]. Традиційні методи допоміжної комунікації включають жести, малюнки, текстові таблиці та пристрої для створення письмових або голосових повідомлень. Однак ці методи не завжди забезпечують достатню швидкість і точність спілкування, особливо у складних соціальних контекстах або під час динамічних діалогів [3]. Низькотехнологічні засоби, такі як жести чи символи, обмежені в можливості адаптації до індивідуальних потреб користувача та контексту взаємодії. Високотехнологічні AAC-рішення, що включають програмні додатки для смартфонів і планшетів зі синтезованою мовою, покращують ситуацію, але потребують подальшого вдосконалення для підвищення ефективності комунікації [4].

Сучасні технології, такі як великі мовні моделі і доповнюючий інформаційний пошук, відкривають нові можливості для AAC-систем. Великі мовні моделі, здатні генерувати текстові відповіді на основі глибокого аналізу контексту, що дозволяє адаптувати відповіді до потреб користувача та історії взаємодії. Інтеграція технології RAG дозволяє залучати актуальну інформацію з зовнішніх джерел, що підвищує точність і змістовність відповідей у реальному часі [5]. Дослідження показують, що застосування LLM у AAC-системах покращує якість комунікації, знижує соціальні бар'єри та забезпечує гнучкість у різних соціальних контекстах.

Персоналізовані текстові відповіді – це відповіді, які адаптуються до індивідуальних особливостей користувача, зокрема до історії попередніх діалогів, потреб та контексту запиту. У процесі взаємодії людини з особливими потребами з системою комунікації користувач ініціює запит через інтерфейс системи, використовуючи текст, символи або жести. Система завантажує контекст, аналізуючи попередні діалоги та профіль користувача, а потім здійснює пошук релевантної інформації для уточнення відповіді. Після цього генерується персоналізована текстова відповідь, яка враховує як знайдені дані, так і індивідуальні особливості користувача. Саме на етапі генерації відповіді відбувається персоналізація з урахуванням усієї доступної інформації.

Ефективність таких рішень підтверджена науковими дослідженнями, які вказують на зростання автономності користувачів завдяки персоналізованим і швидким відповідям, що генеруються мовними моделями [6]. Таким чином, необхідно розробити метод, який використовує великі мовні моделі та технології доповнюючого пошуку для створення точних і персоналізованих відповідей. Це дозволить значно покращити комунікацію для людей з обмеженими можливостями мовлення та сприяти їхній соціальній інтеграції.

Аналіз останніх джерел

Великі мовні моделі (Large Language Models) є сучасними системами штучного інтелекту, здатними генерувати текстові відповіді завдяки навчанню на величезних обсягах даних. Розвиток LLM почався зі створення моделей на основі трансформерів, таких як BERT, GPT та інших подібних архітектур. Модель GPT (Generative Pre-trained Transformer), розроблена OpenAI, демонструє здатність генерувати тексти з глибоким контекстним розумінням, що робить її ефективною для застосування у допоміжних комунікаційних системах [7].

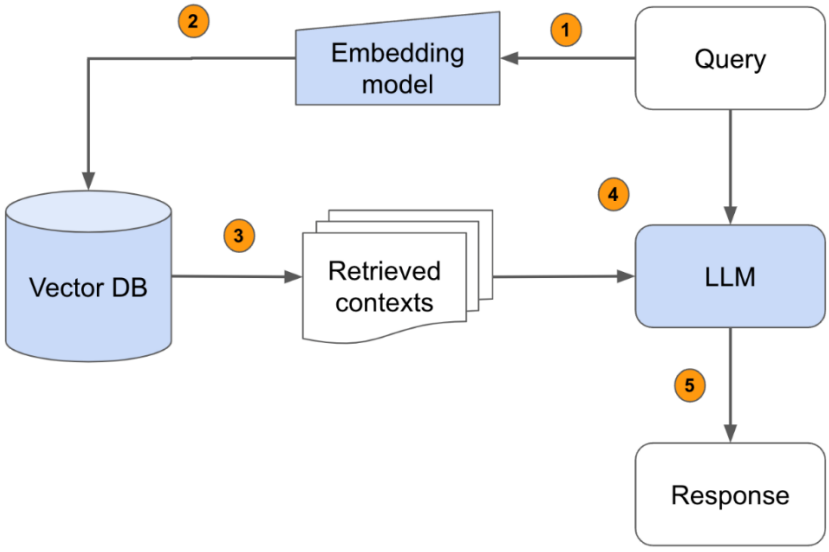

Рис.1. Схема інтеграції LLM з системою RAG

Обробка запиту починається з введення запиту, який може бути питанням, підказкою або будь-яким іншим введенням, на яке мовна модель повинна відповісти. Потім запит передається до моделі вбудовування, яка перетворює його на вектор - числове представлення, яке може бути зрозумілим і обробленим системою. Цей вектор запиту використовується для пошуку у векторній базі даних, яка містить попередньо обчислені вектори потенційних контекстів, що можуть бути використані моделлю для генерації відповіді. Система отримує найбільш релевантні контексти на основі того, наскільки близько їхні вектори відповідають вектору запиту. Отримані контексти передаються до великої мовної моделі, яка використовує цю інформацію для

генерації обґрунтованої та точної відповіді. LLM враховує як оригінальний запит, так і отримані контексти для створення всебічної та релевантної відповіді, синтезуючи інформацію з контекстів, щоб забезпечити, що відповідь не тільки базується на попередніх знаннях, але й доповнена конкретними деталями з отриманих даних. Нарешті, LLM видає відповідь, яка тепер інформована зовнішніми даними, отриманими в процесі, що робить її більш точною та детальною.

У роботі розглядається метод застосування Retrieval-Augmented Generation (RAG) разом із великими мовними моделями (LLM) для створення налаштованих мовних рішень. Основна увага приділена забезпеченню точності та індивідуалізації відповідей на основі конкретних запитів користувачів. Використовується підхід, де інформація спочатку отримується із зовнішніх джерел даних, а потім інтегрується LLM для генерування відповідей, релевантних до контексту. Цей підхід дозволяє користувачам ААС отримувати відповіді, які адаптовані до їхніх унікальних потреб і контекстів. Наприклад, люди з комунікативними порушеннями можуть швидко отримати підказки або тексти, які відповідають їхній конкретній ситуації. Це покращує якість спілкування та робить його більш природним [6].

Стаття описує інтеграцію методів глибокого навчання та нейронних мереж для текстового генерування, яке базується на специфічному контексті. Підхід RAG дозволяє системі виконувати пошук інформації у великих наборах даних перед генерацією відповідей LLM. Ця комбінація підвищує точність та відповідність створених текстів до потреб користувача. Такий підхід може бути використаний у системах допоміжної комунікації для автоматичного підбору фраз, які підходять до конкретної ситуації. Це особливо корисно для створення комунікативних підказок або фраз на основі специфічного контексту, наприклад, під час спілкування на певні теми чи у конкретних соціальних сценаріях [8].

У цій статті представлено підхід до адаптивної комунікації, де Retrieval-Augmented Generation використовується для покращення продуктивності мовних моделей. RAG забезпечує пошук актуальної інформації, що потім інтегрується LLM для створення адаптованих текстів. Автори досліджують, як цей метод підвищує здатність моделей адаптувати відповіді до конкретних потреб користувачів. Цей підхід може допомогти у створенні персоналізованих систем комунікації для користувачів із порушеннями мовлення. Наприклад, системи можуть швидко підбирати найбільш відповідні фрази або тексти для конкретної ситуації, підвищуючи швидкість та точність комунікації. Це дозволяє користувачам легше та ефективніше висловлювати свої думки та потреби [9].

Таким чином, застосування великих мовних моделей та технології RAG є перспективним напрямком для вдосконалення систем допоміжної комунікації. Це дозволяє створювати адаптивні, персоналізовані та високоточні відповіді, що значно покращує якість життя людей з обмеженими можливостями мовлення.

Метою роботи є: покращення комунікації для людей з обмеженими можливостями мовлення за допомогою генерації відповідей з доповнюючим інформаційним пошуком. Метод повинен підвищити швидкість, точність та гнучкість спілкування для людей з обмеженими можливостями мовлення. Потрібно здійснити аналіз ефективності роботи запропонованого методу.

Виклад основного матеріалу

Метод дозволяє ефективно поєднувати генерацію тексту за допомогою великої мовної моделі та пошук відповідної інформації в документах або базах даних, що забезпечує високу точність і адаптивність відповідей до конкретних запитів. Метод базується на трьох основних етапах: індексація даних, пошук інформації та генерація відповіді на основі знайденої інформації. Ці етапи взаємодіють між собою, що забезпечує високу точність і релевантність відповідей. На рисунку 2 зображено процес роботи методу.

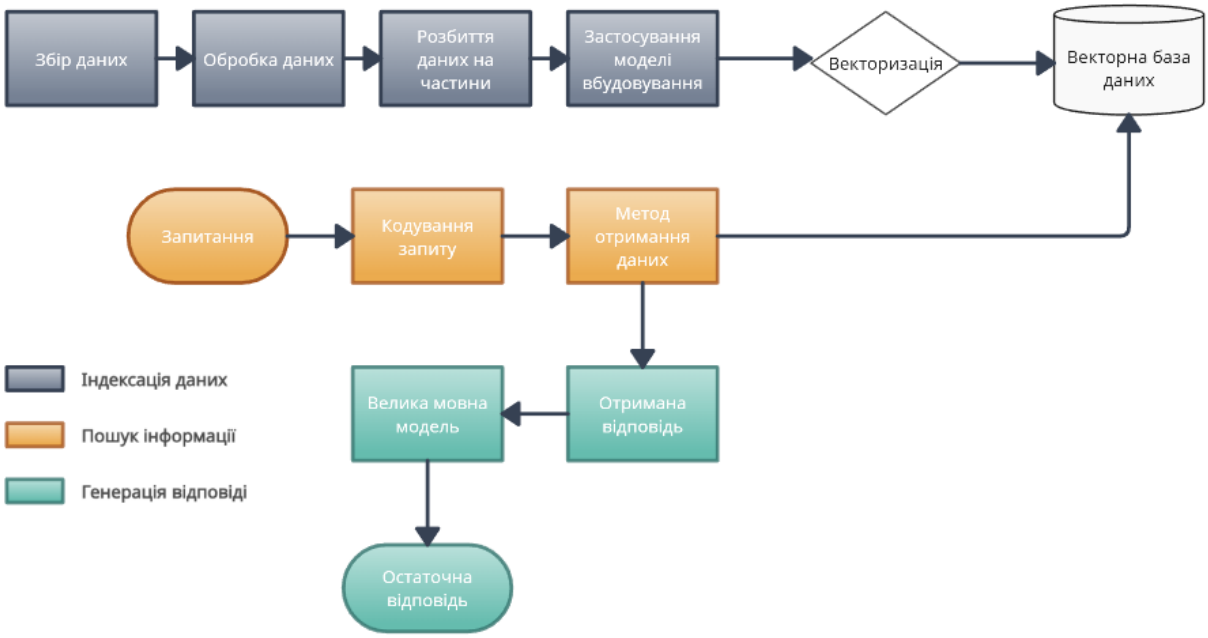

Рис.2. Схема роботи методу генерації відповідей

Індексація даних забезпечує обробку даних для зберігання в базі даних у вигляді векторних представлень. Це дозволяє системі знаходити та зіставляти інформацію, використовуючи семантичну подібність, а не простий текстовий збіг.

Збір даних може охоплювати не лише документи, а й різноманітні джерела, такі як веб-сайти, бази даних, API або навіть зображення. Вибір інструментів для збору інформації залежить від джерела та потреби в забезпеченні цілісності та відповідності запитам користувачів.

На етапі обробки даних важливо очистити інформацію від зайвих елементів та неструктурованих даних. Застосовуються різні техніки, зокрема нормалізація тексту, видалення стоп-слів та вирівнювання формату. Крім того, необхідно виявити та видалити дублікати, які можуть вплинути на ефективність пошуку.

Щоб полегшити пошук та покращити результати, великі текстові блоки або документи розбиваються на менші частини – наприклад, речення чи абзаци. Ці частини можуть бути автономними або взаємопов'язаними залежно від контексту.

Для трансформації тексту в багатовимірні вектори, які відображають зміст, застосовуються моделі вбудовування. Вони можуть бути натреновані на конкретних задачах або використовуватися універсальні моделі, які надають високоякісні ембедінги.

Векторизація дозволяє створити семантичне відображення тексту, що дає змогу системам порівнювати фрагменти тексту через вимірювання відстані між векторами, наприклад, за допомогою косинусної схожості. Чим ближчі два вектори, тим подібніші їх змісти.

Після генерування векторів вони зберігаються в спеціалізованих векторних базах даних. Ці бази даних оптимізовані для швидкого пошуку та масштабування, забезпечуючи ефективний семантичний пошук і надання релевантних результатів у реальному часі.

Пошук інформації відповідає за пошук релевантної інформації, використовуючи векторні представлення. Він дозволяє отримати потрібні документи на основі запиту користувача.

Користувач вводить запит, який може бути як простим питанням, так і складним пошуковим запитом. Часто в запитах є ключові слова або контекст, які допомагають сформулювати точні критерії для подальшого пошуку векторів.

Запит перетворюється на вектор за допомогою спеціалізованих кодувальних моделей, які обробляють текст і відображають його зміст у вигляді багатовимірних векторів. Це дозволяє виявляти схожість між запитом і іншими текстами в базі даних на глибшому рівні, а не лише через точний збіг слів.

Для пошуку найбільш релевантних результатів використовуються методи семантичного пошуку, що порівнюють вектори за допомогою відстані або подібності між ними. Алгоритми, як-от максимальна маржинальна релевантність, не лише підвищують точність, але й покращують різноманітність відповідей, мінімізуючи дублювання інформації у результатах пошуку.

Векторні бази даних, оптимізовані для швидкого пошуку, дозволяють працювати з великими обсягами даних. Вони використовують різні індексаційні структури для ефективного знаходження найбільш релевантних документів, що знижує затримки та покращує час відгуку системи.

Генерація відповіді обробляє знайдену інформацію та створює кінцеву відповідь за допомогою генеративних моделей. Його мета – зробити результат зручним для сприйняття користувачем.

На цьому етапі система отримує фрагменти інформації з бази даних, знайдені під час пошуку. Ці фрагменти можуть містити текст, уривки з документів чи інші дані. Важливо, щоб ці частини були достатньо детальними та містили необхідний контекст для формування змістовної відповіді.

Велика мовна модель (LLM) відповідає за генерацію відповіді, інтегруючи знайдену інформацію в контекст запиту користувача. Модель не лише надає прямі відповіді, але й може додавати пояснення, логічні висновки або структурувати відповідь так, щоб вона ставала зрозумілою та корисною. Для цього використовуються різні стратегії, зокрема переформулювання, підсумки або аналітичні висновки.

Prompt Engineering є важливою складовою методу генерації відповідей із доповнюючим інформаційним пошуком. Оптимізація інструкцій для великої мовної моделі (LLM) дозволяє досягти високої точності, адаптивності та релевантності відповідей. Інструкція для LLM складається з трьох ключових елементів: основної інструкції, контексту та формування відповіді.

Основна інструкція визначає завдання моделі, що полягає в наступному. Необхідно надати відповідь на запитання співрозмовника так, як це могла б зробити людина, інформація про яку міститься у контексті. У відповідях слід використовувати лише ключові слова без додаткових уточнень, описів чи пояснень. Якщо відповідь міститься у контексті, потрібно надати її. Якщо такої відповіді немає, бажано сформулювати кілька варіантів відповідей, що базуються на контексті, історії чату та самому запитанні.

Контекст інструкції включає в себе: контекст (інформацію про людину, якщо є), історію чату (попереднє спілкування), запитання. У контексті використовується інформація з документів, якщо вони наявні, історії чату та запитання. Підсумовуючи, необхідно зазначити, що в контексті інструкції застосовується інформація, отримана від користувача та його співрозмовника.

Формування відповіді уточнює, як повинні виглядати кінцеві варіанти відповіді, а саме – необхідно надати 4 короткі варіанти відповіді. Відповіді повинні:

1. Бути чіткими, короткими, простими і зрозумілими.
2. Уникати будь-яких роздумів, пояснень, уточнень чи рекомендацій.
3. Бути максимально лаконічними та зрозумілими, як у природному діалозі.

Адаптація інструкцій до потреб користувача забезпечує зменшення кількості некоректних результатів та підвищує ефективність комунікації, що є критичним для систем допоміжної комунікації.

Остаточна відповідь, що отримує користувач, об'єднує знайдені дані та логічні висновки, сформовані моделлю. Це може бути текстова відповідь, яка надає прямий відгук на запит, або короткий підсумок, що висвітлює основні моменти. Важливо, щоб відповідь залишалась релевантною і легкою для сприйняття користувачем.

Програмна реалізація методу

Для реалізації методу генерації відповідей з доповнюючим інформаційним пошуком для допоміжної комунікації було використано ряд інструментів та технологій, що забезпечують високу ефективність та точність системи.

Основною мовою програмування, що використана для розробки, є Python. Ця мова є однією з найпопулярніших завдяки своїй універсальності, простоті та багатому набору бібліотек, що забезпечує гнучкість у реалізації різноманітних функцій платформи. Вона дозволяє ефективно інтегрувати різні компоненти системи, що сприяє швидкому розвитку та адаптації до нових вимог [10].

Для створення графічного інтерфейсу користувача застосовувалася бібліотека CustomTkinter. Вона дозволяє створювати адаптивні інтерфейси з підтримкою темної теми, що робить взаємодію з платформою зручною та інтуїтивно зрозумілою для користувачів з різними рівнями підготовки [11].

Для роботи з векторними поданнями даних використано Chroma. Цей інструмент дозволяє перетворювати текстові документи у векторні формати, що забезпечує швидкий пошук релевантної інформації та інтеграцію її з мовними моделями. Chroma є основою для створення бази знань, що забезпечує ефективний доступ до необхідної інформації [12].

Бібліотека PyPDF була використана для автоматичної індексації та обробки PDF-документів. Вона дозволяє ефективно працювати з текстами в складних документах, таких як розділи, таблиці або зображення, що забезпечує точність та повноту даних для подальшої обробки [13].

Моделі, використовувані через Ollama, зокрема aya-expanse:8b та nomic-embed-text, відіграють важливу роль у генерації відповідей. Модель aya-expanse:8b відповідає за створення точних та контекстуальних відповідей, аналізуючи поточний стан чату та враховуючи специфічні потреби користувача [14]. Модель nomic-embed-text відповідає за інтеграцію зовнішніх знань у процес генерації, що дозволяє забезпечити відповідь, яка не лише точна, але й актуальна з точки зору зовнішніх джерел [15].

Для реалізації технології Retrieval-Augmented Generation (RAG) використовувалася бібліотека LangChain. Вона дозволяє ефективно поєднувати пошук релевантної інформації з баз даних та генерацію текстових відповідей, що робить відповіді більш точними і відповідними до контексту запиту користувача [16].

Всі ці інструменти інтегровані в єдину систему, що забезпечує високу точність і швидкість взаємодії з користувачем. Це дозволяє генерувати персоналізовані та релевантні відповіді, значно покращуючи комунікацію для людей з обмеженими можливостями мовлення.

Аналіз ефективності запропонованого методу

Ефективність системи комунікації – це міра відповідності між запланованими цілями та фактичними результатами, отриманими під час тестування за встановленими критеріями. Вона відображає наскільки успішно система задовольняє потреби користувачів та досягає поставлених завдань у визначених умовах використання.

Для тестування системи комунікації з використанням генерації відповідей з доповнюючим інформаційним пошуком використаємо критерії оцінки, що враховують наступні аспекти: Точність відповідей – цей критерій оцінює, наскільки правильно система зрозуміла ваше запитання та надала відповідь, що відповідає його змісту та вимогам; Варіативність відповідей – цей критерій відображає кількість запропонованих варіантів відповідей, їх різноманітність, якість та відповідність запиту користувача; Контекстність відповідей – цей критерій оцінює здатність системи враховувати не тільки попередні повідомлення у діалозі, але й додаткові джерела інформації, особливо у разі запитів, що є частиною тривалої розмови або потребують специфічних знань про користувача; Значущість відповідей – цей критерій оцінює наскільки корисною, релевантною та важливою є отримана відповідь в контексті вашого запиту.

Для тестування системи комунікації будуть створені імітації діалогів на теми «Відвідування ресторану», «Покупки в продуктовому магазині» та «Побут». Імітації діалогів включатимуть конкретні запитання адресовані системі, яка надаватиме варіанти відповіді на запитання для відтворення діалогу. Для порівняння система комунікації буде протестована окремо з використанням методу генерації відповідей з інформаційним пошуком, так і без його використання. Для коректного порівняння запитання під час тестування на однакові теми дублюватимуться – спочатку у варіанті з використанням методу генерації з документом, а потім без нього.

Після проведення ряду тестувань, проведено дослідження для аналізу ефективності системи комунікації. На рисунку 3 продемонстрована статистика балів кожного тестування з використанням доповнюючого пошуку.

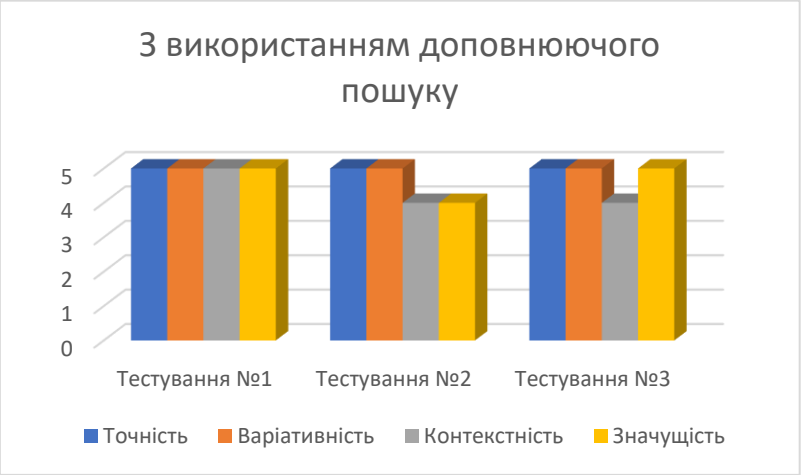

Рис.3. Діаграма тестування з використанням доповнюючого пошуку

На рисунку 4 продемонстрована статистика балів кожного тестування без використання доповнюючого пошуку.

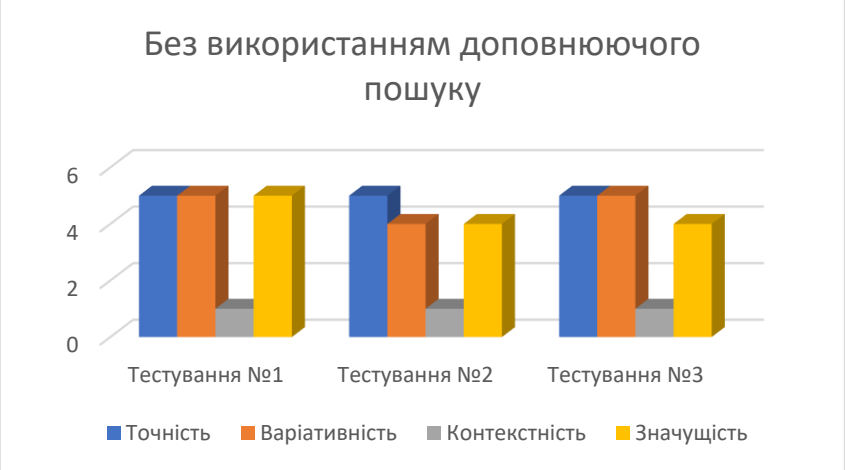

Рис.4. Діаграма тестування без використання доповнюючого пошуку

Після проходження тестування з та без використання доповнюючого пошуку, система комунікації отримала 57 балів із 60 можливих та 45 балів із 60 можливих відповідно. Успішність тестувань можна оцінити за наведеними результатами:

1. Тестування на тему «Відвідування ресторану» з використанням доповнюючого пошуку – 100%;
2. Тестування на тему «Покупки у продуктовому магазині» з використанням доповнюючого пошуку – 90%;
3. Тестування на тему «Побут» з використанням доповнюючого пошуку – 95%;
4. Тестування на тему «Відвідування ресторану» без використання доповнюючого пошуку – 80%;
5. Тестування на тему «Покупки у продуктовому магазині» без використання доповнюючого пошуку – 70%;
6. Тестування на тему «Побут» без використання доповнюючого пошуку – 75%.

Середнє значення ефективності системи комунікації з використанням доповнюючого пошуку сягає 95% із 100% можливих – це прийнятний результат для системи комунікації, коли без використання доповнюючого пошуку лише – 75% із 100% можливих. В основному система комунікації проявила себе досить гарно у всіх тестуваннях, хоча іноді були і недоліки. Можна сказати, що система успішно впоралась із своїми задачами.

На рисунку 5 представлено діаграму, що демонструє статистику по кожному з критеріїв, заданих для оцінки системи комунікації.

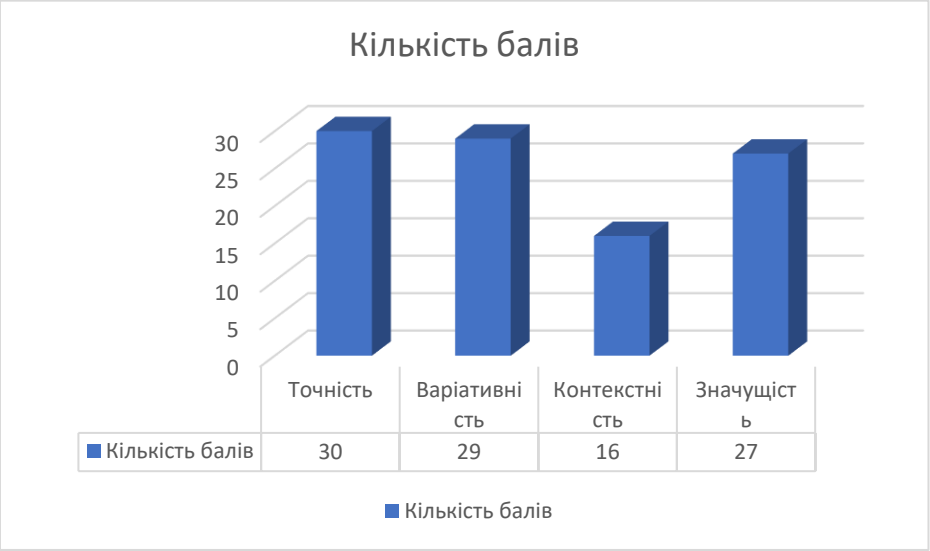

Рис.5. Діаграма статистики кожного з критеріїв оцінки

Після проходження тестування з та без використання доповнюючого пошуку, система комунікації отримувала оцінку по кожному з критеріїв. Успішність критеріїв за результатами тестувань можна оцінити нижче:

1. Точність відповідей – 100%;
2. Варіативність відповідей – 96,67%;
3. Контекстність відповідей – 53,33%;
4. Значущість відповідей – 90%.

Середня оцінка критеріїв системи комунікації сягає 85% із 100% можливих – це прийнятний результат для системи комунікації.

В результаті тестувань система комунікації з використанням генерації відповідей та доповнюючого інформаційного пошуку для допоміжної комунікації продемонструвала високий рівень ефективності. Оцінка за критеріями системи також підтвердила її високу ефективність. Загалом, система успішно виконала поставлені завдання, хоча й потребує подальшого вдосконалення.

Висновки

У результаті проведеної роботи покращно метод генерації відповідей з доповнюючим інформаційним пошуком для допоміжної комунікації, який ефективно підвищує швидкість, точність і гнучкість спілкування для людей з обмеженими можливостями мовлення. Використання таких передових технологій, як великі мовні моделі (LLM) та технологія Retrieval-Augmented Generation (RAG), дозволяє значно покращити комунікацію шляхом інтеграції актуальної інформації з різних джерел, що робить відповіді більш релевантними і змістовними.

Під час тестування системи з використанням методу доповнюючого пошуку було виявлено високий рівень ефективності. Система продемонструвала середній результат 95% ефективності, що свідчить про її здатність забезпечити точні і варіативні відповіді на різні запити користувачів. У порівнянні з традиційними методами без доповнюючого пошуку, ефективність системи без використання цієї технології склала лише 75%. Це підтверджує важливість інтеграції додаткового пошуку для забезпечення більш високої якості комунікації.

Крім того, система продемонструвала відмінні результати за критерієм точності (100%) та варіативності відповідей (96,67%). Хоча критерій контекстності відповіді виявив певні недоліки (53,33%), загальний рівень значущості відповідей (90%) підтверджує, що система може бути ефективно використана в реальних умовах. Результати критерію контекстності притерпіли сильного зниження в тестуваннях без доповнюючого пошуку, через що загальна оцінка також знизилась.

Таким чином, запропонований метод генерації відповідей з доповнюючим інформаційним пошуком має значний потенціал для полегшення комунікації для людей з вадами слуху та мовлення, а також для підвищення їх автономії в повсякденному житті. Подальші дослідження і вдосконалення технології дозволять ще більше покращити якість спілкування та зменшити бар'єри у взаємодії між людьми з різними можливостями.

Література

1. Shining a light on Augmentative and Alternative Communication [Електронний ресурс]. – Режим доступу: https://www.communicationmatters.org.uk/wp-content/uploads/2019/01/2013_Shining_a_Light_on_AAC.pdf
2. Augmentative and Alternative Communication (AAC) [Електронний ресурс] // American Speech-

Language-Hearing Association | ASHA. – Режим доступу: https://www.asha.org/public/speech/disorders/AAc/

3. Chris Klein : Communication and Developing Relationships for People Who Use Augmentative and Alternative Communication, Assistive Technology Outcomes and Benefits, Volume 11, Summer 2017, pp.58-65 URL: https://www.atia.org/wp-content/uploads/2017/11/ATOB_ATOBN1V11_ART-5.pdf

4. Allen, A. A., Schlosser, R. W., Brock, K. L., & Shane, H. C. : The effectiveness of aided augmented input techniques for persons with developmental disabilities: a systematic review. Augmentative and Alternative Communication, 33(3), 149–159. (2017) URL: https://doi.org/10.1080/07434618.2017.1338752

5. Florianne Rademaker, Anke de Boer, Elisa Kupers, Alexander Minnaert: Applying the Contact Theory in Inclusive Education: A Systematic Review on the Impact of Contact and Information on the Social Participation of Students With Disabilities (2020) URL: https://doi.org/10.3389/feduc.2020.602414

6. Sayantan Pal, Souvik Das, Rohini K. Srihari, Jeffery Higginbotham, Jenna Bizovi : Empowering AAC Users: A Systematic Integration of Personal Narratives with Conversational AI (2024) URL: https://doi.org/10.18653/v1/2024.customnlp4u-1.2

7. Zhibo Chu, Zichong Wang, Chengming Li, Ruifeng Xu, Shiwen Ni, Xi Feng, Xiping Hu, Min Yang, Wenbin Zhang : History, Development, and Principles of Large Language Models—An Introductory Survey (2024) URL: https://ar5iv.org/abs/2402.06853

8. Byun, J.; Kim, B.; Cha, K.-A.; Lee, E. : Design and Implementation of an Interactive Question-Answering System with Retrieval-Augmented Generation for Personalized Databases (2024) URL: https://doi.org/10.3390/app14177995

9. Nicholas Thomas Walker, Stefan Ultes, Pierre Lison : Retrieval-Augmented Neural Response Generation Using Logical Reasoning and Relevance Scoring (2023) URL: https://ar5iv.org/html/2310.13566

10. Python [Електронний ресурс]. – Режим доступу: https://uk.wikipedia.org/wiki/Python

11. CustomTkinter: A modern and easy-to-use GUI package for Python. [Електронний ресурс]. – Режим доступу: https://customtkinter.tomschimansky.com

12. Chroma: The simplest way to build search systems. [Електронний ресурс]. – Режим доступу: https://www.trychroma.com/

13. PyPDF Documentation [Електронний ресурс]. – Режим доступу: https://pypdf.readthedocs.io/en/stable/

14. AYA Expanse: Connecting our world. [Електронний ресурс]. – Режим доступу: https://cohere.com/blog/aya-expanse-connecting-our-world

15. Nomic Embed Text v1 [Електронний ресурс]. – Режим доступу: https://www.nomic.ai/blog/posts/nomic-embed-text-v1

16. Langchain: Building and deploying large language models. [Електронний ресурс]. – Режим доступу: https://www.langchain.com

References

1. Shining a light on Augmentative and Alternative Communication [Elektronnyi resurs]. – Rezhym dostupu: https://www.communicationmatters.org.uk/wp-content/uploads/2019/01/2013_Shining_a_Light_on_AAC.pdf

2. Augmentative and Alternative Communication (AAC) [Elektronnyi resurs] // American Speech-Language-Hearing Association | ASHA. – Rezhym dostupu: https://www.asha.org/public/speech/disorders/AAc/

3. Chris Klein : Communication and Developing Relationships for People Who Use Augmentative and Alternative Communication, Assistive Technology Outcomes and Benefits, Volume 11, Summer 2017, pp.58-65 URL: https://www.atia.org/wp-content/uploads/2017/11/ATOB_ATOBN1V11_ART-5.pdf

4. Allen, A. A., Schlosser, R. W., Brock, K. L., & Shane, H. C. : The effectiveness of aided augmented input techniques for persons with developmental disabilities: a systematic review. Augmentative and Alternative Communication, 33(3), 149–159. (2017) URL: https://doi.org/10.1080/07434618.2017.1338752

5. Florianne Rademaker, Anke de Boer, Elisa Kupers, Alexander Minnaert: Applying the Contact Theory in Inclusive Education: A Systematic Review on the Impact of Contact and Information on the Social Participation of Students With Disabilities (2020) URL: https://doi.org/10.3389/feduc.2020.602414

6. Sayantan Pal, Souvik Das, Rohini K. Srihari, Jeffery Higginbotham, Jenna Bizovi : Empowering AAC Users: A Systematic Integration of Personal Narratives with Conversational AI (2024) URL: https://doi.org/10.18653/v1/2024.customnlp4u-1.2

7. Zhibo Chu, Zichong Wang, Chengming Li, Ruifeng Xu, Shiwen Ni, Xi Feng, Xiping Hu, Min Yang, Wenbin Zhang : History, Development, and Principles of Large Language Models—An Introductory Survey (2024) URL: https://ar5iv.org/abs/2402.06853

8. Byun, J.; Kim, B.; Cha, K.-A.; Lee, E. : Design and Implementation of an Interactive Question-Answering System with Retrieval-Augmented Generation for Personalized Databases (2024) URL: https://doi.org/10.3390/app14177995

9. Nicholas Thomas Walker, Stefan Ultes, Pierre Lison : Retrieval-Augmented Neural Response Generation Using Logical Reasoning and Relevance Scoring (2023) URL: https://ar5iv.org/html/2310.13566

10. Python [Elektronnyi resurs]. – Rezhym dostupu: https://uk.wikipedia.org/wiki/Python

11. CustomTkinter: A modern and easy-to-use GUI package for Python. [Elektronnyi resurs]. – Rezhym dostupu: https://customtkinter.tomschimansky.com

12. Chroma: The simplest way to build search systems. [Elektronnyi resurs]. – Rezhym dostupu: https://www.trychroma.com/

13. PyPDF Documentation [Elektronnyi resurs]. – Rezhym dostupu: https://pypdf.readthedocs.io/en/stable/

14. AYA Expanse: Connecting our world. [Elektronnyi resurs]. – Rezhym dostupu: https://cohere.com/blog/aya-expanse-connecting-our-world

15. Nomic Embed Text v1 [Elektronnyi resurs]. – Rezhym dostupu: https://www.nomic.ai/blog/posts/nomic-embed-text-v1

16. Langchain: Building and deploying large language models. [Elektronnyi resurs]. – Rezhym dostupu: https://www.langchain.com