

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

**РОЗРОБКА УКРАЇНСЬКОМОВНОГО ПСИХОЛОГІЧНОГО ЧАТ-БОТУ
ДЛЯ ПЕРСОНАЛІЗОВАНОЇ ПІДТРИМКИ З ВИКОРИСТАННЯМ NLP**

Текстова частина до курсової роботи

За спеціальністю «Інженерія програмного забезпечення» 121

Керівник курсової роботи
к.т.н., доц. Дейнеко А.О.

(підпис)

“ ____ ” _____ 2025 р.

Виконала студентка

Речкалова А.В.

“ ____ ” _____ 2025 р.

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра інформатики факультету інформатики

ЗАТВЕРДЖУЮ

Зав. кафедри інформатики,

доцент, к.ф.-м.н

_____ Гороховський С.С.

(підпис)

“ _____ ” _____ 2025 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на курсову роботу

студентці факультету інформатики 3-го курсу Речкаловій Анна Владиславівні

ТЕМА: Розробка українськомовного психологічного чат-боту для
персоналізованої підтримки з використанням NLP

Зміст ТЧ до курсової роботи:

Індивідуальне завдання

Вступ

Аналіз предметної галузі дослідження та постановка задачі

Моделі генерації тексту

Моделі машинного перекладу

Архітектура системи

Висновки

Використані джерела

“ _____ ” _____ 2025 р.

Керівник _____ (підпис)

Завдання отримав _____ (підпис)

Тема: Розробка українськомовного психологічного чат-боту для персоналізованої підтримки з використанням NLP

Календарний план виконання роботи:

№ п/п	Назва етапу	Термін виконання	Примітка
1.	Отримання завдання на курсову роботу	02.10.2024	
2.	Обговорення деталей теми та розробка плану виконання	18.10.2024	
3.	Виконання аналізу наявних рішень.	18.10.2024 – 24.10.2024	
4.	Огляд технічної літератури за темою роботи та опрацювання матеріалів.	18.10.2024 – 01.12.2024	
5.	Написання практичної частини роботи.	Грудень 2024 – Квітень 2025	
6.	Написання текстової частини роботи.	Квітень 2024 – 02.05.2024	
7.	Створення слайдів для доповіді та написання доповіді.	02.05.2024 – 04.05.2024	
8.	Здача курсової роботи на перевірку на плагіат.	05.05.2024	

Студент: Речкалова А.В.

Керівник: Дейнеко А.О.

“ ____ ” _____ р.

ЗМІСТ

ВСТУП.....	6
ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ	8
АНОТАЦІЯ	9
АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ДОСЛІДЖЕННЯ ТА ПОСТАНОВКА ЗАДАЧІ.....	10
1.1. Використання AI в психологічній підтримці	10
1.2. Основи штучного інтелекту та обробки природньої мови.....	11
1.3. Аналіз наявних чат-ботів психологічної підтримки з використанням технологій NLP.....	12
1.4. Постановка задачі	14
МОДЕЛІ ГЕНЕРАЦІЇ ТЕКСТУ	15
2.1. Короткий огляд роботи моделей генерації тексту	15
2.2. Підхід до вибору мовних моделей	16
2.3. Короткий огляд обраної моделі.....	18
2.4. Донавчання моделі Llama-3.2-1B-Instruct	19
2.5. Оцінка моделі після донавчання	20
МОДЕЛІ МАШИННОГО ПЕРЕКЛАДУ	23
3.1. Огляд підходів до машинного перекладу в NLP	23
3.2. Метрики оцінювання моделей перекладу	25
3.3. Вибір моделей для українсько-англійського перекладу.....	26
3.4. Порівняння моделей англійсько-українського перекладу	27
3.5. Вибір та аналіз моделей для англійсько-українського перекладу.....	29
3.6. Донавчання моделі Helsinki-NLP/opus-mt-en-uk	31
3.7. Аналіз fine-tuned-Helsinki та mbart-large-50-many-to-many-mmt	33
АРХІТЕКТУРА СИСТЕМИ	35
4.1. Загальний опис архітектури чат-бота	35
4.2. База даних для зберігання діалогів	36

4.3. Імітаційний сервер генерації відповідей	37
4.4. Аналіз роботи розробленої системи	37
ВИСНОВОК.....	40
ВИКОРИСТАНІ ДЖЕРЕЛА	41

ВСТУП

Ментальне здоров'я є однією з ключових складових загального благополуччя людини, що визначає її емоційний стан, когнітивні процеси та соціальну взаємодію. У сучасному світі усвідомлення важливості психічного здоров'я стрімко зростає. Цей факт відображається у збільшенні освітніх ініціатив, соціальних кампаній та кількості досліджень, спрямованих на популяризацію турботи про ментальне здоров'я та зниження стигматизації психологічних розладів.

Дослідження «Психічне здоров'я та ставлення українців до психологічної допомоги» [1] проведене у 2024 році вказує, що повномасштабна війна залишається найбільшим джерелом стресу для українців. 40% українців підтвердили, що відчували необхідність у психологічній допомозі за останні півроку, але лише 8% звернулися за допомогою до спеціаліста. Майже 45% опитаних вважають свої проблеми недостатніми для звернення або впевнені, що зможуть розв'язати їх самостійно. Частина рецензентів зазначає, що не можуть дозволити собі допомогу спеціаліста через фінансову складову. Це свідчить про попит на альтернативні та доступні методи психологічної підтримки.

Одним із шляхів вирішення проблеми доступності психологічної допомоги є використання сучасних технологій, зокрема штучного інтелекту. Сьогодні все більше сервісів позиціонують себе як платформи для психологічної підтримки на основі обробки природної мови. Вони використовують штучний інтелект для аналізу тексту, розпізнавання емоційного стану користувача та формування відповідних відповідей, що сприяють психологічному комфорту співрозмовника. Такі рішення стають дедалі популярнішими у світі, однак серед них немає повноцінних україномовних альтернатив.

Метою даної роботи є розробка україномовного психологічного чат-бота для персоналізованої підтримки, що використовуватиме методи обробки природної мови для аналізу тексту та забезпечення ефективної взаємодії з користувачем.

Робота складається з кількох розділів.

У першому розділі розглянуто загальне використання штучного інтелекту в психологічній підтримці, переваги, недоліки та виклики. Також проведено аналіз наявних рішень, їхніх позитивних та негативних сторін. Другий розділ містить огляд сучасних моделей генерації тексту, їх застосування в діалогових системах. У третьому розділі розглянуто моделі машинного перекладу, їх особливості для української мови. Також у цих двох розділах зазначено донавчання моделей з використанням алгоритмів машинного навчання. Четвертий розділ присвячено архітектурі та інфраструктурі системи, а п'ятий – оцінці її ефективності та визначенні перспективи розвитку проєкту.

Розроблено програмний продукт, призначений для надання персоналізованої психологічної підтримки українською мовою з використанням методів обробки природної мови.

Постановка задачі:

1. Виконати аналіз застосування штучного інтелекту в психологічній підтримці, визначити його переваги, недоліки та виклики.
2. Дослідити сучасні рішення у сфері психологічних чат-ботів та виявити їхні обмеження, зокрема відсутність якісних україномовних альтернатив.
3. Коротко дослідити генеративні моделі штучного інтелекту, протестувати різні варіанти та обґрунтувати вибір однієї з моделей. Провести донавчання моделі та налаштування її згідно з використанням у сфері психологічної підтримки, оцінити її ефективність опісля.
4. Проаналізувати підходи до машинного перекладу та провести тестування моделей для перекладу англо-українського та україно-англійського тексту, обґрунтувати вибір моделі для донавчання. Провести аналіз за допомогою метрик оцінки машинного перекладу.
5. Розробити та налаштувати архітектуру чат-бота. Обґрунтувати вибір сервісу для розгортання чат-боту, описати технології тунелювання, збереження повідомлень користувачів у базі даних та серверної взаємодії. Провести аналіз розробленої системи, щоб протестувати ефективність роботи чат-боту.

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

AI – Artificial Intelligence – Штучний інтелект

ML – Machine Learning – Машинне навчання

NLP – Natural Language Processing – Обробка природньої мови

NLU – Natural Language Understanding – Розуміння природньої мови

NLG – Natural Language Generation - Створення Природної Мови

MT – Machine Translation – Машинний переклад

АНОТАЦІЯ

У роботі представлено розробку україномовного психологічного чат-бота, орієнтованого на надання персоналізованої підтримки користувачам. Мета полягає у створенні діалогової системи, здатної ефективно взаємодіяти з людиною за допомогою методів NLP. Для досягнення цієї мети здійснено аналіз існуючих AI-рішень у сфері психологічної допомоги, обґрунтовано вибір генеративної моделі, реалізовано донавчання моделі генерації тексту на тематичних датасетах для покращення емпатії в діалозі. Також досліджено сучасні моделі МТ, виконано їх порівняльну оцінку та обрано оптимальну. Розроблена архітектура чат-бота включає обробку запитів, генерацію відповідей і збереження історії взаємодії. Система протестована на прикладі кількох сценаріїв і продемонструвала стабільну роботу. Результати підтверджують можливість практичного застосування AI у сфері ментального здоров'я.

Ключові слова: AI – Artificial Intelligence, NLP – Natural Language Processing, encoder-decoder transformers, МТ – Machine Translation, донавчання.

АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ДОСЛІДЖЕННЯ ТА ПОСТАНОВКА ЗАДАЧІ

1.1. Використання AI в психологічній підтримці

Штучний інтелект (англ. Artificial Intelligence, AI) – це міждисциплінарна область, що прагне розробити системи, здатні виконувати завдання, які зазвичай потребують людського інтелекту, зокрема навчання, планування, прийняття рішень та інше.

Останні роки різноманітні інституції почали активно впроваджувати AI алгоритми у свої сервіси. Наприклад, у фінансовій сфері для оцінки ризиків або відслідковування підозрілої активності. У медичній галузі популярне використання розпізнавання зображень (Image Recognition) для виявлення початкових стадій захворювання. Сфера логістики впроваджує AI, щоб оптимізувати маршрути, які будуть ефективніші та безпечніші [2].

Як було зазначено у попередньому розділі, попит на психологічну підтримку зростає, тому інструменти AI не залишаються непоміченими та активно впроваджуються у чат-боти, що не є дивним, враховуючи наступні переваги:

1. Доступність 24/7. Неважливо, коли і де користувачу треба підтримка, він може звернутися до AI асистента у будь-яку годину з будь-якого місця.
2. Зниження стресу. У вступі розглядалося дослідження, яке показало, що схильні знецінювати свої ментальні проблеми, тому для деяких людей звертання до чат-бота може бути менш стресовим, ніж розмова з реальним психологом.
3. Персоналізація. Завдяки алгоритмам, які будуть розглянуті у наступному розділі, чат-боти можуть адаптувати свої відповіді до індивідуальних потреб користувача, розпізнаючи емоційний стан і наміри. Це дозволяє створювати персоналізований досвід, що наближається до реальної психологічної підтримки.
4. Раннє розпізнавання ментальних проблем. AI інструменти допомагають виявити ранні ознаки проблем з психічним здоров'ям і

втрутитися до того, як вони стануть більш серйозними. Наприклад, зміна мовленнєвого патерну, що може свідчити про депресію або тривогу [3].

Проте варто зазначити і недоліки AI технологій. Найбільшим викликом залишаються проблеми з конфіденційністю та безпекою, оскільки чат-боти працюють з особистими даними користувачів. Так само величезною проблемою є здатність вирішувати складні психологічні проблеми. AI системи залишаються обмеженими у розумінні глибоких емоційних контекстів, тому можуть недооцінити складність ситуації та надати неправильні поради. Очевидно, що подібні системи можуть бути корисними для надання базової підтримки, але не здатні замінити повноцінну консультацію з психологом або психотерапевтом, особливо у випадках серйозних психологічних розладів.

1.2. Основи штучного інтелекту та обробки природньої мови

У 1950 році Аланом Тюрингом було запропоновано Turing Test – це механізм для визначення інтелекту комп'ютера. Тестування проходить наступним чином: людина задає кілька питань і визначає хто дав відповідь на них – людина чи машина. Очевидно, що комп'ютер проходить цей тест, якщо його відповідь була позначена як людська. Щоб машина змогла конкурувати, вона повинна вміти зберігати фактичні дані та інформацію, яку вона отримує, щоб створювати базу для подальшого аналізу та прийняття рішень (Knowledge Representation); влучно використовувати свою базу із зібраною інформацією, щоб давати релевантні відповіді та формувати нові висновки (Automated reasoning). Щоб це вміти, машина навчається завдяки алгоритмам Машинного навчання (англ. Machine Learning, ML). Система адаптується до нових обставин, виявляє закономірності та екстраполює їх для покращення якості обслуговування [4]. Слід зазначити, що в даній роботі не передбачено заглиблення в деталі ML, проте короткий аналіз даної технології та її застосування буде представлено у наступних розділах, де розглянуто генеративні моделі та їх донавчання.

Щоб машина могла обробляти, розуміти та генерувати людську мову, використовується технологія Обробки природної мови (англ. Natural Language Processing, NLP) – ще одна підгалузь AI [4]. NLP можна поділити на дві підгалузі: розуміння природної мови та створення природної мови. Розуміння природної мови (англ. Natural Language Understanding, NLU) зосереджує увагу на передбаченні значення тексту [5] та розумінні сенсу та почуттів, які виражені природною мовою. Тоді як створення природної мови (англ. Natural Language Generation, NLG), відповідно, зосереджується вже на генерації тексту машиною.

Завдяки технологіям NLP реалізуються численні прикладні завдання, серед яких: переклад, класифікація емоцій, аналіз тональності, виявлення інтенцій користувача, а також побудова інтерактивних систем, таких як віртуальні асистенти та чат-боти, які були згадані у попередньому підрозділі. Зокрема, останні поєднують компоненти NLU та NLG для забезпечення осмисленої взаємодії з користувачем у текстовому або голосовому форматі [4]. Генерація тексту у цьому контексті виконує ключову функцію — формування релевантної, логічно узгодженої та стилістично доречної відповіді на основі введеного запиту.

Машинний переклад та генерацію тексту буде розглянуто у наступний розділах цієї роботи.

1.3. Аналіз наявних чат-ботів психологічної підтримки з використанням технологій NLP

Перед реалізацією власної розробки важливим кроком став аналіз сервісів, які позиціонують себе виключно як чат-боти психологічної допомоги. Оцінювання відбувалося за критеріями функціональності, використання NLP, мовної доступності та взаємодії з користувачем.

1. Wusa [6]. Чат-бот Wusa пропонує користувачеві на початковому етапі вибрати категорії проблем, з якими він звертається, що дозволяє персоналізувати подальшу взаємодію. Частина діалогових відповідей є фіксованою, хоча деякі запити допускають вільне введення. Попри це,

навіть за умови довільного текстового введення, система схильна спрямовувати діалог у заздалегідь заданому руслі. Це свідчить про обмежене використання NLU та залежність від шаблонних структур відповіді. До недоліків належать: відсутність україномовного інтерфейсу, обмеження в безкоштовній версії та значна частина корисних функцій, доступних лише за підпискою.

2. Online Psychologist AI Therapy (Approvatia, Inc) [7]. Даний сервіс характеризується розширеним функціоналом, зокрема доступом до психологічних тестів, методик і бібліотеки зі звуками для релаксації. Інтерфейс є зручним, а початкова взаємодія справляє позитивне враження завдяки структурованій подачі матеріалів. Чат-бот реалізує базові принципи NLP. Разом з тим, сервіс підтримує виключно англійську мову, що обмежує його доступність для неангломовної аудиторії. Надмірна кількість рекламних повідомлень та обмеження на кількість щоденних реплік значно знижують якість користувацького досвіду. Крім того, при виявленні психологічної проблеми система демонструє низький рівень контекстного розуміння: повторює вже поставлені питання та не враховує попередні відповіді.

3. Elomia [8]. Згідно з відкритими джерелами, Elomia є чат-ботом на основі штучного інтелекту, спеціалізованим на психологічній підтримці. Користувач може обрати тематику звернення, аналогічно до функціоналу Wusa, що дозволяє часткову персоналізацію взаємодії. Однак повноцінний доступ до функцій обмежено триденним безкоштовним періодом. Відомо, що чат-бот використовує генеративні моделі для формування відповідей, але відсутність доступу до повної версії унеможлиблює верифікацію їх ефективності, бо додаток недоступний для користувачів Android, що значно обмежує охоплення аудиторії.

Спільними недоліками проаналізованих сервісів є обмежене використання контексту у діалозі, наявність лише англійської версії та залежність від платного доступу до основних функцій. Окрім того, всі розглянуті чат-боти

потребують встановлення окремого застосунку, що знижує зручність їх використання.

1.4. Постановка задачі

Підсумовуючи недоліки вже наявних рішень, завданням даного дослідження є розробка чат-бота, який буде надавати психологічну допомогу без зазначених обмежень. Система має бути здатною до розпізнавання емоційного стану користувача, коректної генерації відповідей з урахуванням контексту та без необхідності додаткового завантаження програмного забезпечення. Особлива увага буде приділена реалізації україномовного сервісу та доступності основних функцій без платних обмежень, що підвищить зручність і доступність сервісу для широкої аудиторії.

МОДЕЛІ ГЕНЕРАЦІЇ ТЕКСТУ

2.1. Короткий огляд роботи моделей генерації тексту

У попередньому розділі було зазначено основну функцію чат-ботів – генерація зв'язного та змістовного тексту. У сучасній практиці функції створення нових об'єктів (текстів, зображень тощо) штучним інтелектом має назву Generative AI [9]. Для виконання цього завдання використовують Large Language Models – нейронні мережі, які навчені на величезних мовних корпусах, завдяки чому здатні розуміти та генерувати людську мову.

Генерація тексту у контексті LLM реалізується як задача авторегресивного прогнозування наступного токена [10] з урахуванням усієї попередньої послідовності, де умовна ймовірність наступного елемента моделюється згідно з рівнянням (формула 2.1.1).

$$P(x) = \prod_{i=1}^n (x_i | x_1, x_2, \dots, x_{i-1}), \quad (2.1.1)$$

де x_i — i -й токен у послідовності, а x_1, x_2, \dots, x_{i-1} — всі попередні токени, що становлять контекст. Цей процес лежить в основі більшості сучасних LLM, побудованих на архітектурі трансформерів (transformers), запропонована у праці «Attention is All You Need» [11].

Оригінальна архітектура трансформерів складається з двох основних елементів: encoder та decoder. Вхідний текст спочатку розбивається на токени — мінімальні одиниці інформації (слова, частини слів або символи), які модель обробляє. У цьому процесі кожен токен перетворюється на числове представлення (вектор), яке зберігає інформацію про семантичне значення цього токена в контексті всіх інших tokenів у послідовності. Encoder приймає вхідну послідовність, розбиває її на токени та формує зважені векторні представлення кожного токена, враховуючи їхній контекст у всій послідовності. Ці представлення передаються до decoder, який поетапно формує вихідну послідовність.

Трансформери працюють паралельно, що істотно підвищує ефективність обробки довгих текстів. Ключовим елементом трансформера є механізм самоуваги (self-attention), який дозволяє кожному токenu «звертати увагу» на інші токени в послідовності. Завдяки цьому модель здатна враховувати контекст усієї фрази — розуміти співвіднесення займенників та логічні зв'язки в довгих реченнях.

Для обчислення самоуваги кожен токен вхідної послідовності перетворюється на три вектори: ключ (key), запит (query) та значення (value). Вага, з якою один токен впливає на інший, визначається як подібність між його запитом і ключами інших токенів. Ці ваги нормалізуються за допомогою функції softmax, після чого використовуються для взваженого сумування відповідних векторів значень:

$$Attention(Q, K, V) = softmax\left(\frac{QK^V}{\sqrt{d_k}}\right)V, \quad (2.1.2)$$

Де Q (query) — матриця запитів, K (key) — матриця ключів, V (value) — матриця значень, d_k — розмірність простору ключів. Таким чином формується нове представлення кожного токена, яке враховує повний контекст [11].

У результаті, після проходження всіх етапів — від розбиття на токени до самоуваги та генерації векторних представлень — модель обирає наступний токен, який найкраще відповідає контексту. Кожен новий токен додається до послідовності, і процес повторюється доти, доки не буде сформовано повне речення або абзац. Таким чином, машинна генерація тексту поступово, токен за токеном, перетворюється на зв'язне і логічне висловлювання, яке користувач сприймає як цілісний людський текст.

2.2. Підхід до вибору мовних моделей

Рішення про вибір саме англomовної моделі також зумовлене тим, що більшість високоякісних інструкційно налаштованих моделей орієнтовані

насамперед на англійську мову. Українська мова, як правило, або відсутня в мовному покритті, або присутня частково, без специфічного тренування на діалогах. Це створює суттєві обмеження щодо якості безпосередньої україномовної генерації, зокрема в сфері психологічної підтримки, де особливо важливі нюанси мови, тон, емпатія та контекст. У зв'язку з цим було прийнято рішення використовувати англомовну модель з подальшим перекладом тексту з/на українську.

У процесі відбору були протестовані моделі із сімейств LLaMA та Qwen. Основну увагу зосереджено на відносно компактних інструкційних (instruction-tuned) моделях, оптимізованих для діалогових задач. Було протестовано таких представників Qwen2.5-1.5B-Instruct-GGUF [12], meta-llama/Llama-2-7b-chat-hf [13], meta-llama/Llama-3.2-1B-Instruct [14]. Головними критеріями під час вибору моделі були якість відповіді (особливо емоційна складова), стабільність роботи, час та ресурси, що потрібні для генерації відповіді, а також відкритість ліцензійного використання. Тестування моделей здійснювалося в середовищі Google Colab із використанням безкоштовного графічного процесора (GPU T4). Було частково взято репліки пацієнта на прийомі у а за основу використано психологічний діалог «Emotional Reasoning: A Dialogue Between A Therapist and Patient» [15]. Для забезпечення відповідного стилю відповідей додатково задано інструкцію (промпт) моделі відповідати в теплій, природній манері, з акцентом на співчутливість, емоційну залученість і підтримку користувача у формі невимушеного діалогу.

Порівняння показало, що всі протестовані моделі демонструють загалом прийнятний рівень емпатії в тексті, однак мають свої особливості:

1. Qwen2.5-1.5B-Instruct-GGUF: Модель добре розрізняє емоції та відповідає у потрібній манері, але відповіді надто довгі та містять повторення, що виглядає неприродно. Середній час генерації становить 24.07 секунд, пост-процесинг обов'язковий через неочікувані елементи (хештеги, надмірна кількість тексту).

2. Llama-2-7b-chat-hf: Відповіді емоційно збалансовані, без зайвих деталей чи повторів. Середній час генерації становить 7.045 секунд. Модель стандартно налаштована на розмовний стиль (chat model). Однак, є схильність додавати непотрібні позначення емоцій (наприклад, empathetic nod), які можна усунути налаштуванням параметрів або пост-процесингом.

3. Llama-3.2-1B-Instruct: Модель менш креативна ніж попередні, але більше фокусується на підтримці діалогу і залученні користувача до розмови. Цей підхід є зручним для тривалих сесій із користувачем. Середній час генерації становить 2.88 секунди.

У підсумку було обрано модель Llama-3.2-1B-Instruct як таку, що найкраще задовольняє вимоги проєкту завдяки швидкості, лаконічності та здатності підтримувати тривалу взаємодію з користувачем.

2.3. Короткий огляд обраної моделі

Модель Llama-3.2-1B-Instruct [14], розроблена компанією Meta, належить до нового покоління багатомовних мовних моделей, оптимізованих для генерації тексту в діалогових сценаріях. Вона містить приблизно 1 мільярда параметрів і побудована на основі авторегресивної трансформерної архітектури, яку було описано у розділі 2.1, точніше - архітектура типу decoder-only без окремого encoder. У структурі передбачено 22 шари трансформера, 16 голів самоуваги.

Довжина контексту може сягати до 128 тисяч токенів, що дозволяє працювати з великими обсягами тексту. Модель навчене на великій багатомовній суміші відкритих даних, з офіційною підтримкою англійської, німецької, французької, іспанської, португальської, італійської, хінді та тайської мов. Для покращення якості відповіді вона була додатково налаштована за допомогою керованого навчання (Supervised Fine-Tuning, SFT) і підкріплення з використанням зворотного зв'язку від людини (Reinforcement Learning with Human Feedback, RLHF), що дозволяє узгоджувати відповіді з людськими очікуваннями щодо корисності, безпеки та емпатійності. Перевага цієї моделі

також полягає у відкритій ліцензії (Llama 3.2 Community License), що дозволяє гнучке використання в дослідницьких цілях.

2.4. Донавчання моделі Llama-3.2-1B-Instruct

Перед початком донавчання модель тестувалася із заздалегідь сформованим промптом, що задавав очікувану поведінку моделі як «психолога-консультанта». У попередніх експериментах із використанням лише промпт-інжинірингу спостерігалось, що модель, хоча й відтворювала структуру емоційної підтримки, схильна була до повторення загальних фраз і не демонструвала належної поведінки, що була схожа на психолога. Однією з причин цього була необхідність формувати надмірно деталізований промпт, через що модель також генерувала довгі і перевантажені відповіді.

Для подолання цих обмежень було застосовано донавчання (fine-tuning)[16] моделі на спеціалізованому корпусі діалогів між клієнтами та консультантами. Fine-tuning, як одна з технік ML, дозволяє адаптувати вже попередньо натреновану мовну модель до специфіки конкретного завдання, зокрема — відтворення стратегії психологічної взаємодії.

Донавчання проводилося у Google Colab на NVIDIA A100 на датасеті LangAGI-Lab/cactus[17], який містить реалістичні діалоги у форматі «клієнт–консультант». Було проведено препроцесинг: виділено окремо запити користувача та реакції консультанта, що подаються у форматі інструкційного навчання. Після обробки було сформовано підмножину з 100 000 прикладів, із яких 80 % призначено для навчання, а 20 % — для валідації. Використовувалася інструкція «You are a professional therapist. You provide psychologically sound answers» (укр. «Ви професійний психотерапевт. Ви даєте психологічно обґрунтовані відповіді»). Модель завантажувалася у 4-бітній квантованій версії, що значно зменшило обсяг пам'яті, необхідний для навчання. Для донавчання було використано бібліотеку PEFT (Parameter-Efficient Fine-Tuning) [18] із конфігурацією LoRA (Low-Rank Adaptation) [19], яка дозволяє адаптувати велику мовну модель за допомогою лише часткової модифікації параметрів. У

рамках цієї конфігурації були обрані всі лінійні шари моделі, окрім вихідного (lm_head). Навчання відбувалося із використанням класу SFTTrainer, що орієнтований саме на завдання інструкційного налаштування (Supervised Fine-Tuning). Донавчання було проведено у 6 епох. Модель збережено на платформі Hugging Face у репозиторію `avlare/llama-therapist-full` для подальшого використання.

2.5. Оцінка моделі після донавчання

Щоб оцінити покращення моделі після проведення донавчання, було обрано кілька реплік зі справжнього сеансу психологічної консультації [33]. Було використано кілька психологічно орієнтованих запитів у форматі повідомлень із ролями `system` (інструкція, що задавалася під час навчання) та `user`. Генерація здійснювалася через застосування шаблону `tokenizer.apply_chat_template`, після чого модель відповідала за допомогою `generate` з обмеженням на довжину відповіді (100 токенів). Відповідь моделі аналізувалася за допомогою регулярних виразів, щоб виділити лише змістовну частину. Під час кожного запуску фіксувався час, необхідний для отримання відповіді, що дозволило оцінити продуктивність. Також вимірювався час, щоб порівняти швидкодію моделей. Один з результатів тестування наведено далі.

Одна з реплік користувача, яка була використана під час тестування: «I've actually been feeling quite exhausted and stressed out about my studies in the past few days and it's affecting how I communicate with others and causing me to doubt myself» (укр. «Останніми днями я відчуваю себе дуже виснаженою і напруженою через навчання, і це впливає на те, як я спілкуюся з іншими, і змушує мене сумніватися в собі»). Донавчена модель `avlare/llama-therapist-full` дала наступну відповідь: «It sounds like this has been really challenging for you. When you say "doubt myself," what kind of thoughts are going through your mind during these moments? What do you tell yourself?» (укр. «Звучить так, ніби це було справді складним випробуванням для вас. Коли ви говорите «сумніватися в собі», які думки проскакують у вашій голові в ці моменти? Що ти собі кажеш?»). Час відповіді

склав 1.763 секунди. Базова модель Llama-3.2-1B-Instruct згенерувала таку відповідь: «It sounds like you're going through a tough time, and it's completely normal to feel overwhelmed and uncertain about your academic performance. I'm here to offer some support and guidance. Firstly, please know that you're not alone in feeling this way. Many students face similar challenges, and it's essential to remember that it's okay to ask for help. It's great that you're reaching out to me, and I'm here to support you. To start, let's explore some possible reasons» (укр. «Схоже, що ти переживаєш важкі часи, і цілком нормально відчувати себе пригніченим і невпевненим у своїх академічних успіхах. Я тут, щоб запропонувати тобі підтримку та поради. По-перше, знайте, що ви не самотні у своїх відчуттях. Багато студентів стикаються з подібними проблемами, і важливо пам'ятати, що просити про допомогу - це нормально. Це чудово, що ти звернувся до мене, і я тут, щоб підтримати тебе. Для початку давайте розглянемо деякі можливі причини»). Відповідь моделі було обірвано через обмеження у токенах. Час генерації відповіді склав 2.187 секунди.

У порівнянні з базовою версією моделі, донавчена модель демонструє покращення як у психологічній релевантності відповіді, так і в ефективності генерації. З психологічної точки зору відповідь донавченої моделі є більш таргетованою, сфокусованою на емоційному досвіді користувача та побудована за принципами активного слухання. Вона не лише виявляє емпатію, але й формулює відкриті питання, спрямовані на розгортання рефлексії користувача, що відповідає базовим технікам когнітивно-поведінкової терапії. Натомість відповідь базової моделі хоч і є співчутливою, залишається загальною, більш декларативною і містить надмірну кількість шаблонних фраз підтримки, що знижує ефективність комунікації у контексті терапевтичної взаємодії. З технічної точки зору модель після донавчання демонструє зменшення часу генерації відповіді при тих самих параметрах запуску (максимум 100 токенів), що свідчить про покращення навченого розподілу ймовірностей наступних токенів у контексті психологічних запитів. Це, ймовірно, пов'язано з кращою адаптацією моделі до тематичного домену за рахунок спеціалізованого корпусу

під час донавчання. Крім того, зниження ентропії вихідного простору дозволяє моделі швидше знаходити релевантні послідовності відповідей, не жертвуючи змістовною якістю. Показник часу відповіді зменшився на $\sim 19.4\%$, що є статистично помітним при фіксованих апаратних і програмних параметрах.

Таким чином, покращення моделі є комплексним: донавчання не лише підвищило якість відповіді у змістовному і психологічному вимірі, а й оптимізувало технічні характеристики, що має значення для подальшого використання моделі в реальному часі, наприклад, у чат-ботах психоемоційної підтримки.

МОДЕЛІ МАШИННОГО ПЕРЕКЛАДУ

3.1. Огляд підходів до машинного перекладу в NLP

Machine translation (MT) – це підгалузь комп'ютерної лінгвістики, основною роботою якої є розробка систем перекладу текстів з однієї мови на іншу [20]. Головним завданням є використання правильної граматики зі збереженням контексту оригіналу. Нейронний машинний переклад (Neural Machine Translation, NMT) [21] сьогодні є домінуючим підходом до автоматичного перекладу текстів, що ґрунтується на використанні нейронних мереж для моделювання процесу перекладу як задачі послідовного перетворення. Зазвичай завдання MT працюють так, що здійснюється переклад кожного окремого речення, а не тексту загалом. MT використовує підхід Supervised machine learning, тобто під час тренування системі дано датасет паралельних речень, і вона вчиться розмічувати речення з оригінальної мови на обрану мову перекладу.

Важливо уточнити, що токенізація речення відбувається не звичайним поділом на слова, а виокремленням частин слів або навіть окремих символів. Основними типами токенізації у MT є BPE, wordpiece та unigram (SentencePiece). Алгоритм Byte Pair Encoding (BPE) — метод токенізації, який поєднує часті послідовності символів у нові токени, поступово створюючи словник з найбільш поширених підслів. Починаючи з розбиття слів на окремі символи, BPE ітеративно зливає найчастіші пари символів або підслів у нові одиниці. Wordpiece — алгоритм, який є досить схожим на BPE, але замість злиття найбільш частих пар, WordPiece вибирає злиття, яке максимізує ймовірність даних за мовною моделлю. У підсумку модель отримує токени, які краще відображають морфологічну структуру мови та зменшують кількість незнайомих слів. Unigram — це один із найпоширеніших алгоритмів токенізації, який реалізується через бібліотеку SentencePiece (тому інколи і носить таку назву). На відміну від BPE, цей алгоритм не об'єднує пари токенів, а починає з великого словника, що включає всі окремі символи Unicode та частотні послідовності символів (зокрема всі слова, розділені пробілами). Потім він

ітеративно видаляє менш ймовірні токени, поки не досягне бажаного розміру словника. На кожному кроці оцінюється, наскільки ймовірним є кожен токен, і вибираються ті, які найчастіше входять до ймовірніших розбиттів тексту. Unigram вважається кращим за VPE, бо створює більш семантично осмислені токени й уникає дрібних, позбавлених значення частинок, характерних для VPE [22, розділ 13.2.1].

Стандартною архітектурою для МТ є encoder-decoder transformer, яку представлено на рисунку 3.1.1. Ця структура була коротко розглянута в розділі 2.1, однак тут важливо зосередитись на її ключових етапах у контексті завдання перекладу.

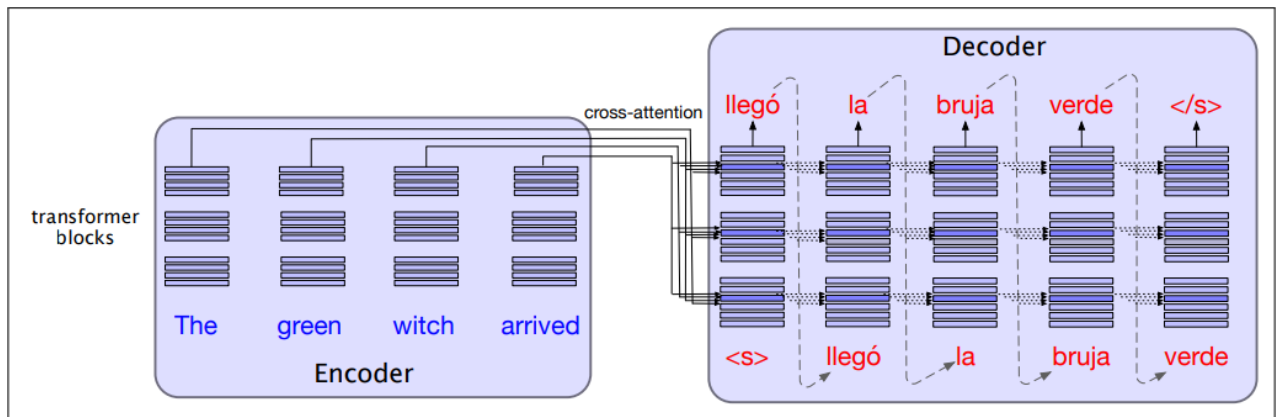


Рисунок 3.1.1 Архітектура encoder-decoder transformer для МТ [22, с. 272]

Encoder приймає речення мовою оригіналу, яке було токенизовано (одним з алгоритмів наведених раніше), який перетворює їх а векторні представлення — контекстуально зважені за допомогою механізму self-attention. Отримані дані далі використовує decoder, який також працює за принципом self-attention, але додатково включає механізм cross-attention. Завдяки cross-attention decoder під час генерації кожного нового токена мовою перекладу порівнює їх з результатами у encoder [22, розділ 13.3]. Таким чином, модель не лише запам'ятовує вже згенеровані токени перекладу, але й на кожному кроці заново зіставляє їх із повною інформацією з вхідного речення. Завдяки такому

механізму система здатна підтримувати граматичну узгодженість та зберігати значення речення при перекладі.

3.2. Метрики оцінювання моделей перекладу

Оцінювання МТ ділиться на дві категорії: Automatic Evaluation (Автоматизоване оцінювання) та Human Evaluation (Людське оцінювання) [22, розділ 13.6.2]. Тестування проводиться на «золотому» датасеті [23] — це набір даних з еталонними, перевіреними вручну відповідями, що використовується для оцінки якості моделей.

Двома найпопулярнішими метриками Automatic Evaluation є BLEU [24] та chrF [25], які використовуються не лише для завдань МТ.

BLEU (BiLingual Evaluation Understudy) — найрозповсюдженіша метрика оцінювання, яка базується на підрахунку збігів n-грам (послідовностей із n токенів) між МТ і «золотим» перекладом. BLEU оцінює лише точність (precision), ігноруючи повноту (recall), тобто перевіряє, наскільки добре токени згенерованого перекладу повторюють токени з оригінального [24]. Підсумовуючи, наведена метрика має серйозні обмеження, зокрема чутливість до токенизації й складність роботи з мовами зі складною морфологією. Дана метрика не використовується у даному дослідженні, але згадується в оригінальних дослідженнях авторів моделей МТ, які буде наведено у розділі 3.3.

ChrF (character F-score) — метрика, яка також базується на збігах між символічними n-грамами, тобто на відміну від BLEU, chrF працює не з токенами слів, а з символами (та їх комбінаціями), що робить її особливо корисною для мов із багатою морфологією, таких як українська. ChrF обчислюється як гармонічне середнє між середньою точністю (chrP) та середньою повнотою (chrR) для всіх символічних n-грам від 1 до k. ChrF розраховується за формулою 3.2.1:

$$chrF = (1 + \beta^2) \cdot \frac{chrP \cdot chrR}{\beta^2 \cdot chrP + chrR}, \quad (3.2.1)$$

де chrP (character Precision) — середня точність, тобто відсоток символічних n-грам у гіпотезі (машинному перекладі), які також є в референсному перекладі; chrR (character Recall) — середня повнота, тобто відсоток символічних n-грам у референсі, що також є в гіпотезі; β — ваговий коефіцієнт, який дозволяє надавати перевагу точності або повноті (стандартне значення $\beta=2$ (інколи використовується значення 6) надає більше ваги повноті, тобто краще відобразити сенс ніж вгадати окремі слова). $\beta=2$ означає, що використовуються лише unigrams та bigrams [25].

Для визначення статистичної значущості різниці між результатами двох систем машинного перекладу (наприклад, за метрикою chrF) використовують методи статистичного тестування, зокрема парний бутстрап-тест (paired bootstrap test). Цей метод дозволяє перевірити, чи є спостережувана різниця між оцінками моделей A і B значущою, а не випадковою. Ідея полягає у формуванні псевдо-наборів даних шляхом вибірки з поверненням із тестового датасету. Для кожного псевдо-набору обчислюють chrF-оцінку для обох систем A і B. Далі аналізують частку випадків, у яких одна система перевищує іншу за якістю перекладу. Цей підхід також дозволяє обчислити довірчий інтервал chrF для окремої системи, якщо розглядати лише її оцінки й виключити верхні та нижні 2.5% значень [22, с. 283]. Важливо зазначити, що якщо довірчий інтервал включає 0, то неможливо стверджувати про перевагу однієї моделі над іншою, тобто вони можуть працювати однаково. У такий спосіб можна зробити статистично обґрунтований висновок щодо переваги однієї моделі над іншою або відсутності значущої різниці.

3.3. Вибір моделей для українсько-англійського перекладу

Для реалізації чат-бота, який взаємодіє з українськомовними користувачами, генеруючи відповіді англійською, ключовою вимогою є наявність якісного та швидкого механізму перекладу. У цьому розділі розглянуто українсько-англійські моделі МТ, які у даній розробці використовуються для перекладу повідомлень користувача.

Було протестовано дві вузькоспеціалізовані моделі для конкретної мовної пари — Helsinki-NLP/opus-mt-uk-en[26] та facebook/mbart-large-50-many-to-many-mmt[27].

Модель Helsinki-NLP/opus-mt-uk-en побудовано на основі фреймворку MarianNMT [28] — це реалізація Transformer-моделі, оптимізована для ефективного inference на CPU. Архітектура моделі відповідає класичному encoder-decoder transformer з 6 шарами як в encoder, так і в decoder. Модель має 75 мільйонів параметрів. Препроцесингом є нормалізація та токенизація SentencePiece. Модель тренувалася на корпусах OPUS [29], такі як OpenSubtitles, GlobalVoices, Tatoeba та інші. Chr-F score вказаний розробниками дорівнює 0.757, BLEU – 64.1 на датасеті Tatoeba.uk.en [26].

Модель facebook/mbart-large-50-many-to-many-mmt базується на багатомовному Transformer-моделі mBART50, що є донавченим розширенням mBART (Multilingual BART) [30]. Вона підтримує прямий переклад між будь-якою парою з 50 мов. Архітектура моделі — класичний encoder-decoder Transformer із 12 шарами в кожному блоці, загальною кількістю параметрів близько 611 мільйонів. Для вказівки цільової мови використовується спеціальний BOS-токен. Початкове навчання mBART здійснювалося на 25 мовах з високим ресурсом, після чого модель була розширена на ще 25 мов через ініціалізацію нових embedding-векторів і спільне донавчання на монолінгвальних корпусах. Українська мова була включена до моделі mBART50 у категорії мов із середньою кількістю даних — від 100 тис. до 1 млн паралельних речень. Для продовження навчання використовували монолінгвальні дані. Модель mBART50 була донавчена на 50 мовах, зокрема українській, упродовж 300 тис. ітерацій із великою партією (1700 токенів). Найбільший приріст (до 7 BLEU) спостерігався саме у мов з обсягом даних, подібним до української [31].

3.4. Порівняння моделей англійсько-українського перекладу

Для проведення аналізу було власноруч зібрано «золотий» датасет – 100 записів англійський речень та їхній переклад українською. Оцінювання

проводилося за трьома метриками: медіана часу перекладу у секундах та chrF бутстрап-тест.

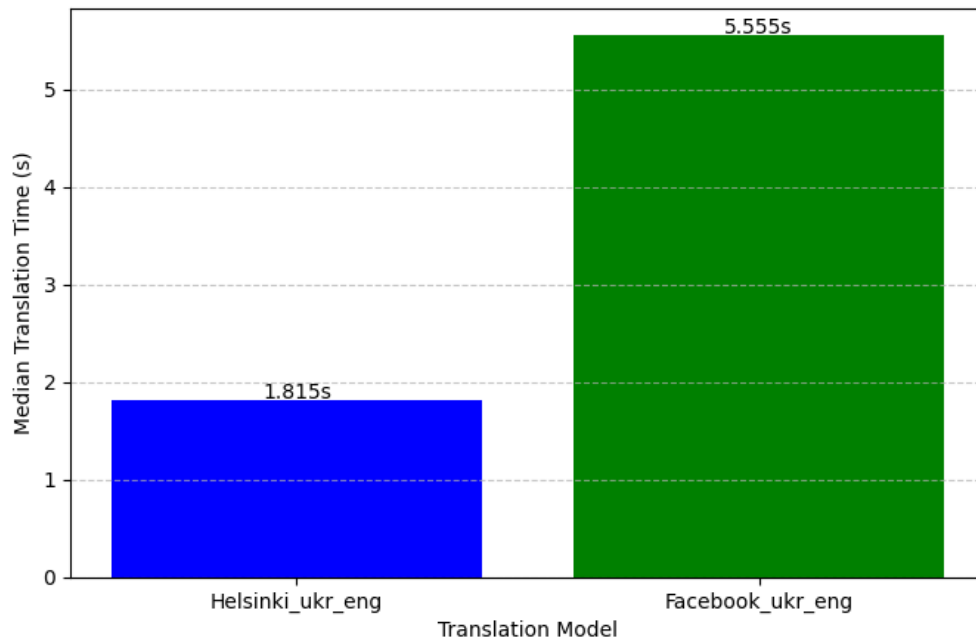


Рисунок 3.4.1 Порівняння медіани часу українсько-англійського перекладу моделями opus-mt-uk-en та mbart-large-50-many-to-many-mmt

На графіку 3.4.1 показано, що медіана часу перекладу «золотого» датасету моделлю Helsinki-NLP/opus-mt-uk-en становить 1.815 секунди, тоді як для facebook/mbart-large-50-many-to-many-mmt значення є 5.555 секунд. Тобто перша модель працює у 3 рази швидше.

Тестування chr-F проводилося бутстрап-тестом, визначеним у розділі 3.2. На графіку 3.4.2 зображено розподіл різниці оцінок якості перекладу між Helsinki-NLP/opus-mt-uk-en та facebook/mbart-large-50-many-to-many-mmt, у результаті довірчий інтервал складає від -3.147 до 1.666, медіана -0.687.

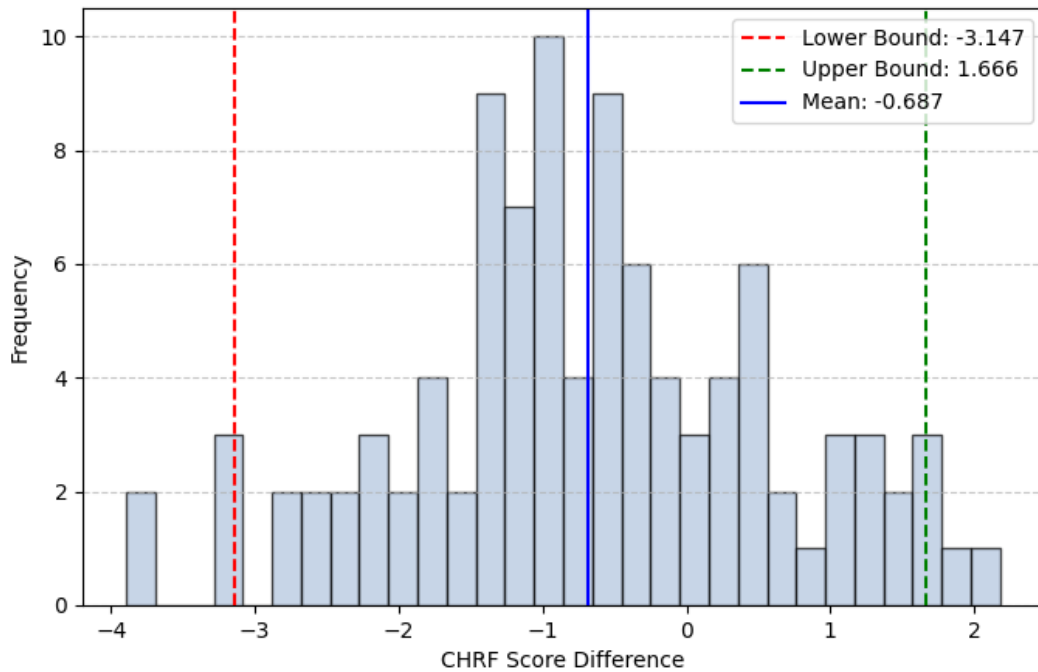


Рисунок 3.4.2 Результати *chrF* бутстреп-тесту українсько-англійського перекладу моделями *opus-mt-uk-en* та *mbart-large-50-many-to-many-mmt*

Від’ємне значення медіани та нижньої границі інтервалу вказує, що модель Helsinki-NLP/opus-mt-uk-en зазвичай перекладає гірше, проте довірчий інтервал включає 0, тобто робота моделей сильно не відрізняється.

За результатами тестувань для подальшої розробки було обрано модель Helsinki-NLP/opus-mt-uk-en, оскільки вона демонструє втричі швидший час перекладу, а різниця в якості перекладу з моделлю Facebook не є статистично значущою, що робить її ефективнішим вибором для реального застосування.

3.5. Вибір та аналіз моделей для англійсько-українського перекладу

Щоб перекласти відповідь від Llama, треба якісна та швидка англійсько-українська модель машинного перекладу. Для аналізу було обрано facebook/mbart-large-50-many-to-many-mmt, архітектуру якої розглянуто у розділі 3.3, та Helsinki-NLP/opus-mt-en-uk, що базується на тій самій архітектурі, що й Helsinki-NLP/opus-mt-uk-en, описана раніше.

Тестування проводилося на тому самому «золотому» датасеті з порівнянням медіани часу перекладу та метрики chrF, яка визначалася через бутстрап-тест.

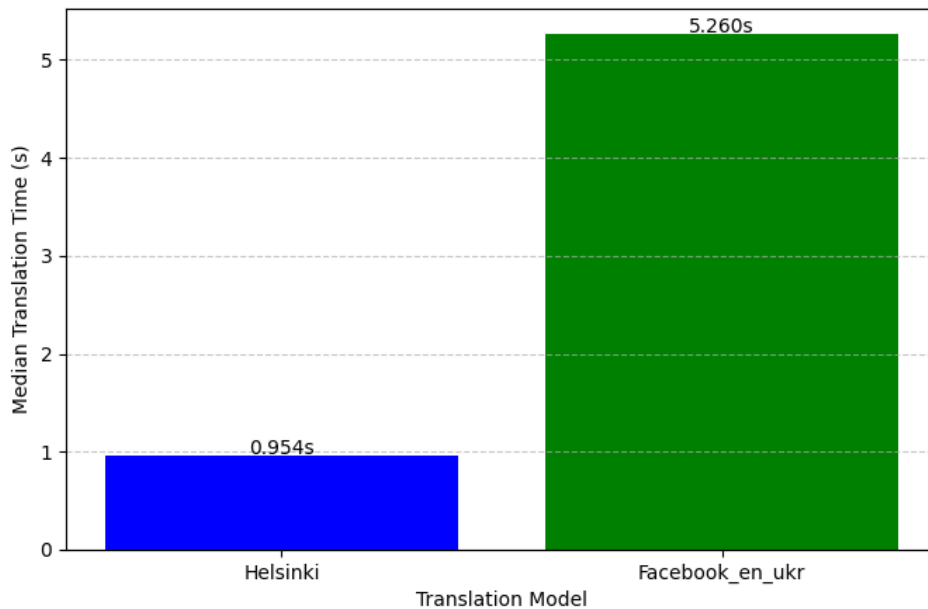


Рисунок 3.5.1 Порівняння медіани часу англійсько-українського перекладу моделями opus-mt-en-uk та mbart-large-50-many-to-many-mmt

На графіку 3.5.1 показано, що медіана часу перекладу моделлю Helsinki-NLP/opus-mt-en-uk має значення 0.954 секунди, facebook/mbart-large-50-many-to-many-mmt – 5.260 секунд. Тобто перша модель працює у п’ять разів швидше. На графіку 3.5.2 зображено розподіл різниці оцінок якості перекладу між Helsinki-NLP/opus-mt-en-uk та facebook/mbart-large-50-many-to-many-mmt, у результаті довірчий інтервал складає від -5.060 до 0.671, медіана -2.223.

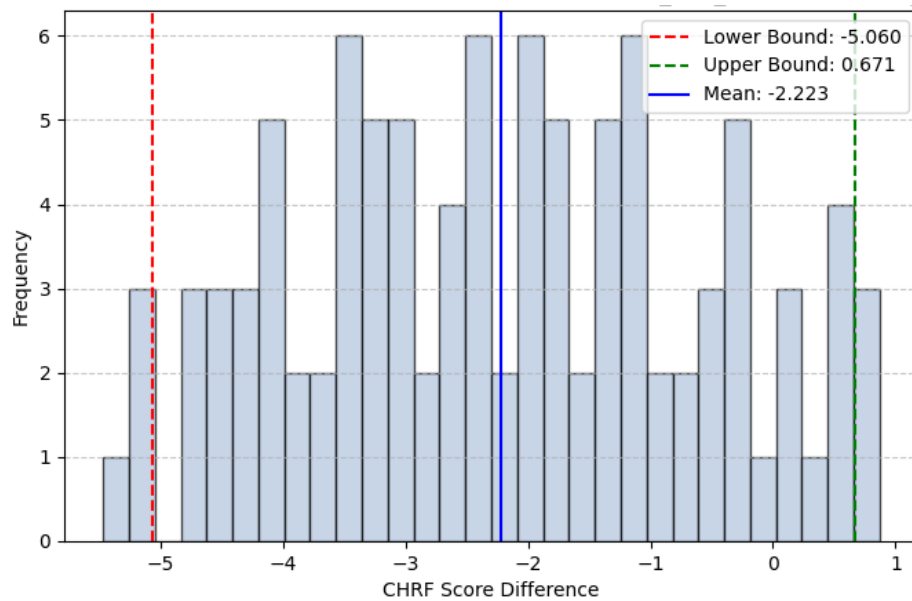


Рисунок 3.5.2 Результати $chrF$ бутстреп-тесту англійсько-українського перекладу моделями *opus-mt-en-uk* та *mbart-large-50-many-to-many-mmt*

Довірчий інтервал знову містить 0, що говорить про майже відсутність статистичної різниці у роботі моделей, але результати для моделі від Helsinki-NLP для англійсько-українсько перекладу гірші ніж для українсько-англійського. Під час додаткового тестування на особистих даних було помічено, що модель Helsinki-NLP/opus-mt-en-uk може інколи перекладати іншими мовами, окрім української. Скоріш за все, корпус, на якому навчалася модель, містив речення або навіть тексти відмінними мовами. Дана модель мала важливу перевагу – швидкість перекладу, тому було вирішено покращити якість перекладу моделі через донавчання та знову порівняти з моделлю від Facebook.

3.6. Донавчання моделі Helsinki-NLP/opus-mt-en-uk

Для донавчання було використано датасет FLoRes [32] від Facebook. Датасет було об'єднано з підмножин dev та devtest (це склало близько 2000 записів). Так як метою було лише зменшити випадки перекладу моделлю іншими мовами, то увага зосереджувалася не на розмірі датасету, а якості. Датасет було поділено у відношенні 80/20, 80% – тренування, 20% – валідація. Перед обробкою було адаптовано токенизатор, додано спеціальний падінг-токена та

реалізовано препроцесинг із чіткою інструкцією перекладу. Для зменшення обчислювальних витрат та пришвидшення процесу донавчання була використана методика LoRA (Low-Rank Adaptation), яка згадувалася у розділі 2.4 під час донавчання Llama-3.2-1B-Instruct. Навчання проходило в Seq2SeqTrainer із використанням оптимізатора AdamW, batch size = 64, learning rate = 1e-5 та 25 епох.

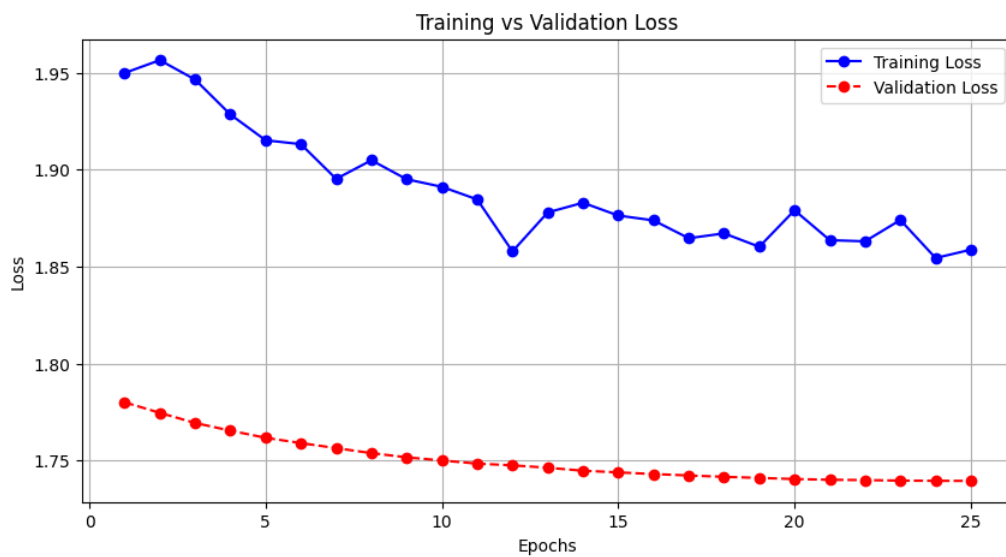


Рисунок 3.6.1 Графік навчання Helsinki-NLP/opus-mt-en-uk

Графік навчання 3.6.1 показує поступове зменшення втрат на тренувальному та валідаційному наборах, що свідчить про стабільне навчання моделі. Хоча тренувальні втрати мають незначні коливання, загальний тренд залишається спадним, що підтверджує коректну адаптацію моделі. Натомість validation loss демонструє чітке і плавне зниження, досягаючи плато ближче до 20-ої епохи. Це свідчить про те, що модель ефективно узагальнює нові знання без втрати якості на невідомих прикладах. Модель збережено з назвою fine-tuned-Helsinki.

3.7. Аналіз fine-tuned-Helsinki та mbart-large-50-many-to-many-mmt

Тестування проводилося у тих самих умовах, що і попереднє. Медіана часу роботи донавченої моделі від Helsinki-NLP зросла на секунду (див. рисунок 3.7.1) і становить 2.100 секунди, що у 2.5 рази менше за facebook/mbart-large-50-many-to-many-mmt.

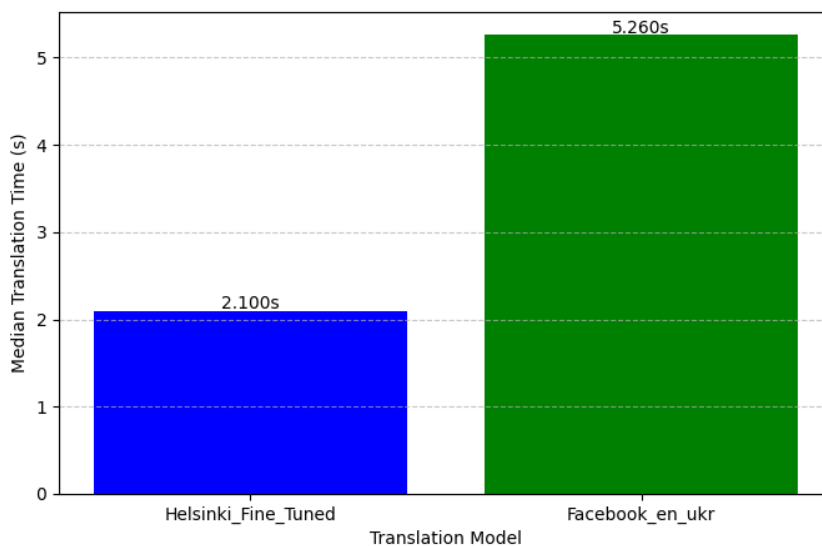


Рисунок 3.7.1 Порівняння медіани часу англійсько-українського перекладу моделями *fine-tuned_Helsinki* та *mbart-large-50-many-to-many-mmt*

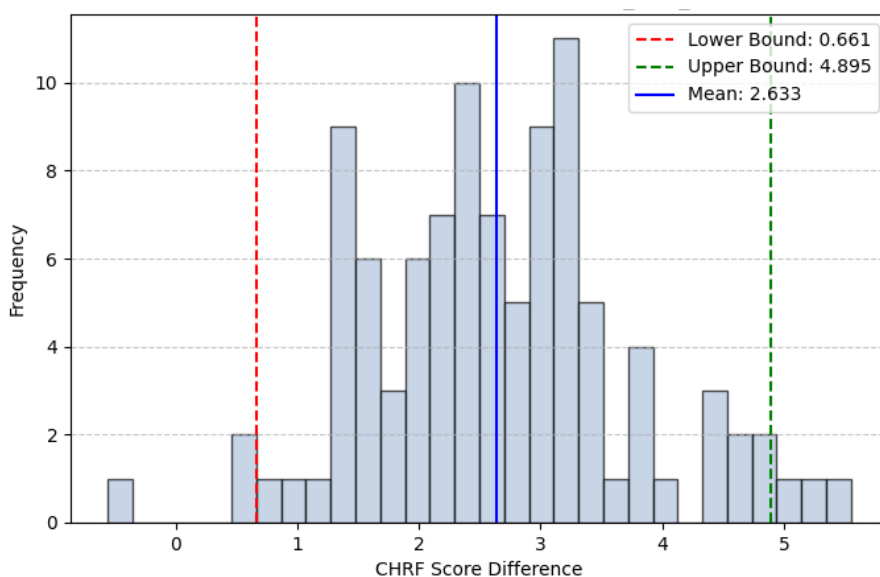


Рисунок 3.7.2 Результати *chrF* бутстрап-тесту англійсько-українського перекладу моделями *fine-tuned-Helsinki* та *mbart-large-50-many-to-many-mmt*

На графіку 3.7.2 зображено розподіл різниці оцінок якості перекладу між *fine-tuned-Helsinki* та моделлю від Facebook, у результаті довірчий інтервал складає від 0.661 до 4.895, медіана 2.633. Тобто донавчена модель працює краще.

Підсумовуючи результати тестування, модель *fine-tuned-Helsinki* беззаперечно переважає у якості та швидкості перекладу, тому її було впроваджено у розробку.

АРХІТЕКТУРА СИСТЕМИ

4.1. Загальний опис архітектури чат-бота

Facebook Messenger було обрано як платформу для чат-бота з огляду на баланс між безпекою, доступністю як користувачам, так і у процесі розгортання та підтримки.

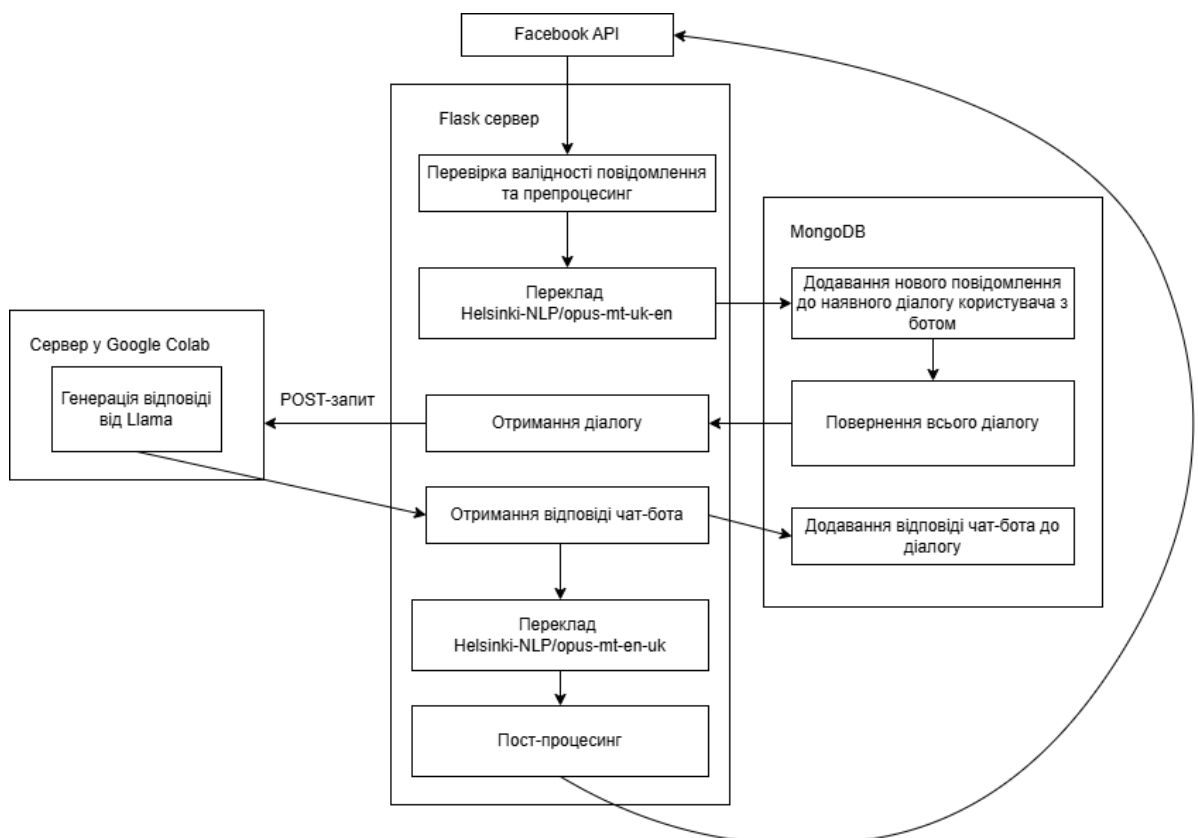


Рисунок 4.1.1 Архітектура роботи чат-бота

На рисунку 4.1.1 представлено загальну архітектуру системи. Основу серверної логіки побудовано за допомогою фреймворку Flask, який обробляє вхідні запити користувачів через окремий endpoint /webhook. Вся бізнес-логіка структурована по модулях згідно з принципами Controller-Service-Repository. Паралельна обробка повідомлень досягається за допомогою черги Queue і окремого потоку Thread, що дозволяє уникнути блокування основного потоку

Flask. Для забезпечення зовнішнього доступу до локального сервера використовується тунелювання через ngrok, яке дозволяє Facebook Messenger надсилати webhook-запити. Першим етапом є валідація повідомлення – сервер парсить запит і опрацьовує ідентифікатор користувача та вміст повідомлення.

Повідомлення перед перекладом проходить препроцесинг – ділиться на окремі речення утворюючи масив. Перекладений вміст додається до бази даних, структура якої буде розглянуто у розділі 4.2. Щоб згенерувати відповідь увесь діалог користувача із чат-ботом відправляється на сервер у Google Colab, де відбувається генерація відповіді попередньо вибраною моделлю Llama. Архітектура та особливості цього сервера буде розглянуто у розділі 4.3.

Отримана відповідь проходить пост-процесинг: вміст ділиться на окремі речення, кожне з якого перекладається та опрацьовується з урахуванням форматування, відновленням нумерованих списків і стилістичних маркерів.

HTTP-запити до Facebook обробляються через стандартну бібліотеку requests, яка забезпечує надсилання згенерованих ботом відповідей (кожного окремого речення) користувачу.

4.2. База даних для зберігання діалогів

Для зберігання історії діалогів між користувачем та чат-ботом використовується нереляційна база даних MongoDB. Кожен користувач зберігається як окремий документ у колекції `user_messages`, ідентифікований за `_id`, який відповідає його унікальному ідентифікатору. Всі повідомлення зберігаються у вигляді масиву об'єктів `messages`, де перше — системне повідомлення з інструкцією до моделі, а наступні — послідовні повідомлення користувача і відповіді моделі. У базі зберігаються усі повідомлення користувача, але для етапу генерації відповіді використовується останні 15 та промпт. Це забезпечує стабільну швидкість роботи навіть при великій кількості користувачів. Окрім того, модель LLaMA найкраще працює з контекстом останніх реплік, фокусуючи увагу на актуальному діалозі без перенавантаження зайвою інформацією.

4.3. Імітаційний сервер генерації відповідей

Модель Llama-therapist-full є досить великою і для швидкої відповіді вимагає GPU-ресурс. На Google Colab розгорнуто імітаційний сервер генерації відповідей, який слугує прототипом inference-сервера для LLM і використовує безкоштовний GPU T4. Було розгорнуто REST API за допомогою FastAPI, який забезпечує обробку HTTP-запитів. Для асинхронної підтримки в середовищі Colab застосовано nest_asyncio, що дозволяє одночасно запускати події подібно до справжнього серверного циклу. Модель Llama-therapist-full було завантажено через HuggingFace Transformers pipeline.

Для розгортання API в публічному доступі використано сервіс ngrok, який створює зовнішній тунель до локального порту FastAPI. Це дає можливість взаємодіяти з моделлю ззовні, імітуючи поведінку повноцінного сервера.

Отримані відповіді моделі надсилаються через HTTP POST-запити з повідомленнями до Flask-серверу, який надалі займається надсиланням обробленого тексту користувачу.

4.4. Аналіз роботи розробленої системи

З огляду на обмеження середовища тестування, повноцінне навантажувальне тестування системи не проводилося. Проте було здійснено базову оцінку роботи чат-бота в умовах імітації двох активних користувачів, які надсилають кілька повідомлень поспіль. Це дозволило перевірити коректність поетапної обробки запитів та оцінити швидкодію системи на практиці.

Система обробляє повідомлення від одного користувача послідовно, ставлячи кожне нове в чергу після попереднього. Це дозволяє зберігати логіку діалогу та уникати перетину відповідей. Кожен цикл включає переклад запиту, генерацію відповіді, зворотній переклад, опрацювання (видалення зайвих символів) і надсилання. Фрагмент результатів тестування наведено у таблиці 4.4.1.

Час перекладу повідомлення користувача (у секундах)	Час генерації (у секундах)	Час переклад відповіді моделі (у секундах)	Час постпроцесингу та надсилання повідомлень (у секундах)	Кількість згенерованих речень
0.942	2.059	2.289	1.715	3
0.449	2.42	2.15	1.907	3
2.543	2.415	3.914	2.181	4
1.454	2.406	4.316	1.753	3
1.271	2.819	4.982	2.551	5
1.681	2.936	4.107	2.218	4

Таблиця 4.4.1 Фрагмент окремих метрик під час тестування системи

Підсумовуючи середні показники для проаналізованих метрик, час перекладу повідомлення користувача (українсько-англійський переклад) становить 1.39 секунди, генерації відповіді моделлю `avlare/llama-therapist-full` – 2.509 секунди, переклад відповіді (згенерованого тексту) (англійсько-український переклад) – 3.62 секунди, постпробробка та надсилання всіх речень складає 2.054 секунд. У середньому модель генерує у відповідь 4 речення (3.66).

Таким чином, середній повний час обробки одного запиту, починаючи від моменту його надсилання до завершення формування відповіді, становить приблизно 9.57 секунди. Водночас варто враховувати, що постпроцесинг та надсилання охоплюють одразу кілька речень (у середньому — 3.66), тому перше речення користувач отримує значно швидше — орієнтовно через 7.5 секунди після надсилання запиту. Це дозволяє зменшити суб'єктивне сприйняття затримки та створює враження швидкої реакції чат-бота.

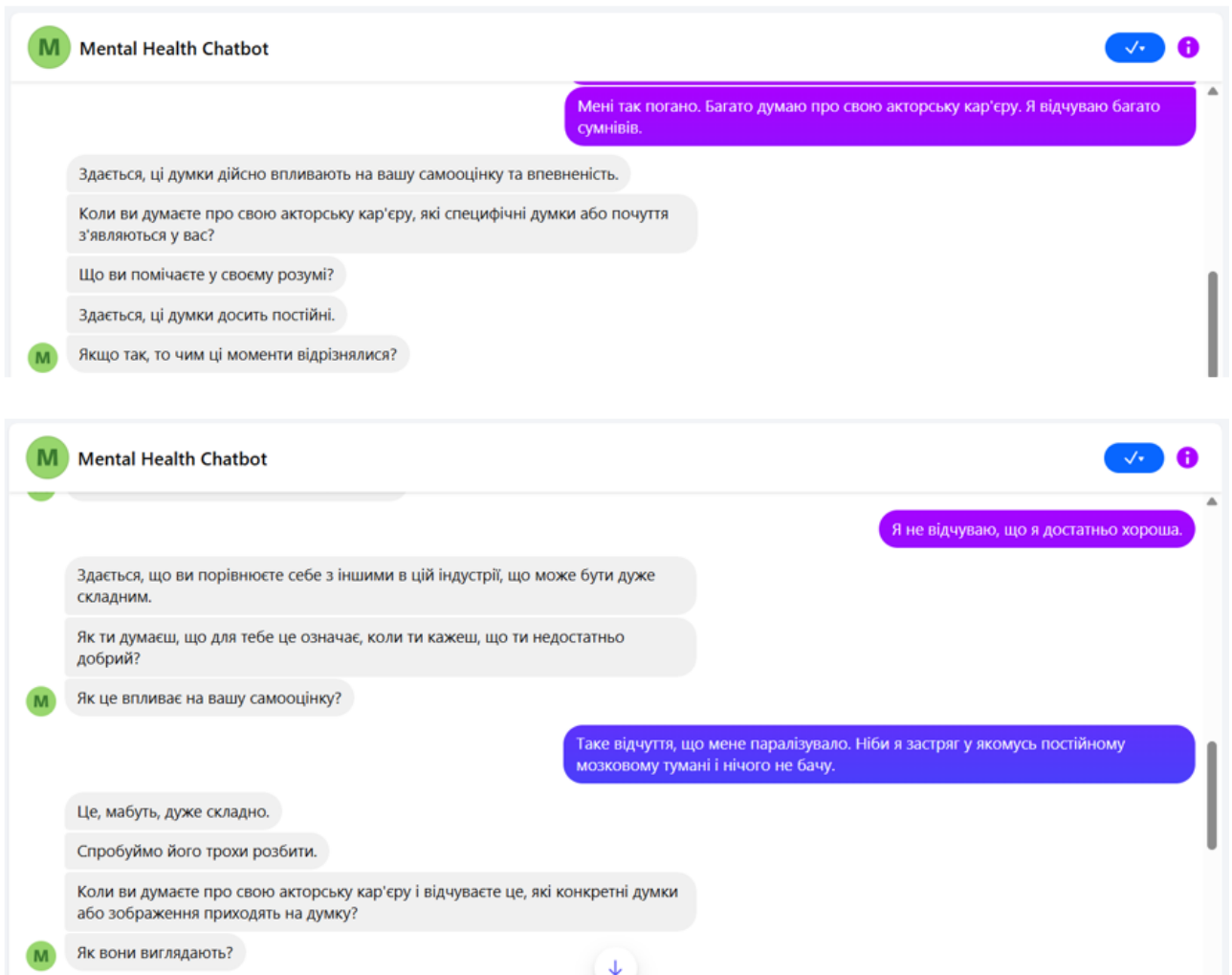


Рисунок 4.4.1 Фрагмент тестування моделі

На рисунку 4.4.1 зображено приклад роботи системи через інтерфейс Facebook Messenger. Загалом отримані результати свідчать про прийнятну швидкодію системи в умовах обмеженого навантаження. Подальше тестування на більших обсягах даних дозволить точніше оцінити її стабільність і продуктивність у реальному середовищі.

ВИСНОВОК

У результаті виконання роботи було реалізовано прототип україномовного психологічного чат-бота, який забезпечує персоналізовану підтримку користувачів. Поставлені завдання були успішно виконані: здійснено аналіз сучасних технологій штучного інтелекту у сфері психологічної підтримки, досліджено переваги та обмеження існуючих рішень. Було обрано та донавчено модель сімейства LLaMA, адаптовану для відповідей у ролі психолога в діалозі. Також було протестовано кілька моделей машинного перекладу, проведено оцінювання якості перекладу та обґрунтовано вибір моделі для подальшого використання. Створено архітектуру системи з підтримкою асинхронної обробки запитів і збереженням історії діалогу. Завдяки цьому чат-бот здатен забезпечити контекстуальну, змістовну відповідь, враховуючи попередні репліки користувача. Система протестована на прикладі двох активних користувачів і продемонструвала стабільну роботу. Було зафіксовано прийнятні значення часу обробки запитів. Робота демонструє потенціал використання великих мовних моделей у сфері ментального здоров'я. Основна цінність розробки полягає в адаптації глобальних технологій до українського контексту. Чат-бот може бути масштабований і вдосконалений за рахунок подальшого донавчання, використання потужніших ресурсів, які дозволять зменшити час відповіді чат-бота. Перспективним напрямом є розширення функціоналу: аналіз емоцій, кризові інтервенції, інтеграція з гарячими лініями.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Психічне здоров'я та ставлення українців до психологічної допомоги [Електронний ресурс]. — Режим доступу: <https://gradus.app/uk/open-reports/mental-health-and-attitudes-ukrainians-towards-psychological-assistance-during-war/>
2. Artificial intelligence in positive mental health: a narrative review [Електронний ресурс]. — Режим доступу: <https://pmc.ncbi.nlm.nih.gov/articles/PMC10982476/#s1>
3. Sarwar Z., Haider M., Alam T. та ін. Conversational AI for Mental Health Support [Електронний ресурс]. — Режим доступу: https://www.researchgate.net/publication/381950638_Conversational_AI_for_Mental_Health_Support
4. Russell S. J., Norvig P. Artificial Intelligence: A Modern Approach. Third Edition. Lecture Slides. Section 1.1 [Електронний ресурс]. — Режим доступу: <https://people.engr.tamu.edu/guni/csce625/slides/AI.pdf>
5. DeepLearning.AI. Natural Language Processing Resources [Електронний ресурс]. — Режим доступу: <https://www.deeplearning.ai/resources/natural-language-processing/>
6. Wysa: AI Chatbot for Mental Health [Електронний ресурс]. — Режим доступу: <https://www.wysa.com/>
7. Online Therapy AI Psychologist [Електронний ресурс]. — Режим доступу: <https://play.google.com/store/apps/details?id=com.SugarSocialNetworks.SecondLife&hl=uk&pli=1>
8. Elomia: AI Chatbot for Mental Health [Електронний ресурс]. — Режим доступу: <https://elomia.com/>
9. Amazon Web Services. What is Generative AI? [Електронний ресурс]. — Режим доступу: <https://aws.amazon.com/what-is/generative-ai/>
10. Medium. Autoregressive Models World. [Електронний ресурс]. — Режим доступу: <https://medium.com/@peechapp/text-to-speech-models-part-2-autoregressive-models-world-636d8aa0932d>

11. Vaswani A., Shazeer N., Parmar N. та ін. Attention is All You Need [Електронний ресурс]. — Режим доступу: <https://arxiv.org/pdf/1706.03762>
12. Hugging Face. Qwen2.5-1.5B-Instruct-GGUF Model [Електронний ресурс]. — Режим доступу: <https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct-GGUF>
13. Hugging Face. Llama-2-7b-chat-hf Model [Електронний ресурс]. — <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>
14. Hugging Face. Llama-3.2-1B-Instruct Model [Електронний ресурс]. — <https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct>
15. Shamam I. Emotional Reasoning: A Dialogue Between a Therapist and Patient [Електронний ресурс]. — Режим доступу: <https://www.isaacmshamam.com/blog/emotional-reasoning-a-dialogue-between-a-therapist-and-patient>
16. IBM. Fine-Tuning in AI Models [Електронний ресурс]. — Режим доступу: <https://www.ibm.com/think/topics/fine-tuning>
17. Hugging Face. LangAGI-Lab/cactus Dataset [Електронний ресурс]. — Режим доступу: <https://huggingface.co/datasets/LangAGI-Lab/cactus>
18. Hugging Face. Parameter-Efficient Fine-Tuning Documentation [Електронний ресурс]. — Режим доступу: <https://huggingface.co/docs/peft/index>
19. Hugging Face. Training with LoRA [Електронний ресурс]. — Режим доступу: <https://huggingface.co/docs/diffusers/main/en/training/lora>
20. GeeksforGeeks. Machine Translation of Languages in Artificial Intelligence [Електронний ресурс]. — Режим доступу: <https://www.geeksforgeeks.org/machine-translation-of-languages-in-artificial-intelligence/>
21. TranslateFX. What is Neural Machine Translation Engine? [Електронний ресурс]. — <https://www.translatefx.com/blog/what-is-neural-machine-translation-engine-how-does-it-work>

22. Jurafsky D., Martin J. H. Speech and Language Processing : [draft of the 3rd ed.] [Электронний ресурс]. — Stanford University, 2023. — Режим доступу: <https://web.stanford.edu/~jurafsky/slp3/>
23. DagsHub. Golden Dataset [Электронний ресурс]. — Режим доступу: <https://dagshub.com/glossary/golden-dataset/>
24. Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: a method for automatic evaluation of machine translation [Электронний ресурс]. — 2002. — С. 311–318. — Режим доступу: <https://dl.acm.org/doi/10.3115/1073083.1073135>
25. Maja Popović. chrF: character n-gram F-score for automatic MT evaluation [Электронний ресурс]. — 2015. — С. 392–395. — Режим доступу: <https://aclanthology.org/W15-3049/>
26. Hugging Face. opus-mt-uk-en Model [Электронний ресурс]. — Режим доступу: <https://huggingface.co/Helsinki-NLP/opus-mt-uk-en>
27. Hugging Face. mBART-large-50-many-to-many-mmt Model [Электронний ресурс]. — Режим доступу: <https://huggingface.co/facebook/mbart-large-50-many-to-many-mmt> - просто модель
28. Marian-NMT [Электронний ресурс]. — Режим доступу: <https://marian-nmt.github.io/>
29. OPUS [Электронний ресурс] — Режим доступу: <https://opus.nlpl.eu/>
30. Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li та ін. Multilingual Denoising Pre-training for Neural Machine Translation [Электронний ресурс]. — 2020. — Режим доступу: <https://arxiv.org/pdf/2001.08210>
31. Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen та ін. Multilingual Translation with Extensible Multilingual Pretraining and Finetuning [Электронний ресурс]. — 2020. — Режим доступу: <https://arxiv.org/abs/2008.00401>
32. Hugging Face. Facebook FLORES Dataset [Электронний ресурс] — Режим доступу: <https://huggingface.co/datasets/facebook/flores>

33. Studocu. Counseling script [Электронный ресурс] – Режим доступа:
<https://www.studocu.com/ph/document/de-la-salle-university-dasmarinas/communication-culture-and-society/counseling-script/55617814>