Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«Дослідження якості україно-англійського автоматичного перекладу мовлення: семантична точність та міжфразова когерентність»**

Виконав: студент 4-го року навчання,
Спеціальності
113 Прикладна математика

Нанаров Ілля Олександрович

Керівник: Іванюк.А.О,
Рецензент _____
(прізвище та ініціали)

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК _____
(підпис)

«_____» _____ 20___р.

Київ – 2026

Міністерство освіти і науки України
Національний університет «Києво-Могилянська академія»
Факультет інформатики
Кафедра математики

ЗАТВЕРДЖУЮ
Зав. кафедри математики,
доцент, кандидат фіз.-мат. наук
_____ *Чорней Р.К.*
(підпис)
“____”_____ 2025

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на кваліфікаційну роботу
студенту 4-го курсу факультету інформатики
Нанарову Іллі Олександровичу

Тема: «Дослідження якості україно-англійського автоматичного перекладу мовлення: семантична точність та міжфразова когерентність»

Зміст кваліфікаційної роботи:

Анотація
1. Вступ
2. Теоретичні основи: моделі автоматичного розпізнавання мовлення (англ. Automatic Speech Recognition, ASR), машинного перекладу (англ. Machine Translation, MT), методи оцінювання якості
3. Корпус дослідження: формування та характеристики матеріалу
4. Архітектура та компоненти системи: ASR (Whisper), MT (OPUS-MT), система метрик, дизайн експерименту
5. Результати: оцінка якості ASR та MT, каскадний ефект, порівняння з великими мовними моделями
Висновки
Список літератури

Дата видачі “____”_____ 2026 Керівник _____
(підпис)

Завдання отримав _____
(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2025р.

№ з/п	Перелік робіт	Термін виконання етапу	Підпис наукового керівника	Дата ознайомлення наукового керівника	Примітка
1.	Отримання теми кваліфікаційної роботи.	17.10.2025			
2.	Ознайомлення з темою кваліфікаційної роботи.	25.10.2025			
3.	Розробка плану та структури роботи.	15.11.2025			
4.	Робота з науковою літературою, опис основних означень.	17.01.2026			
5.	Дослідження поставленої задачі та подальша оптимізація.	15.02.2026			
6.	Робота над текстовим оформленням отриманих результатів.	18.04.2026			
7.	Попередній аналіз кваліфікаційної роботи. Виправлення помилок.	04.05.2026			
8.	Попередній захист кваліфікаційної роботи.	23.05.2026			
9.	Захист кваліфікаційної роботи.	05.06.2026			

Науковий керівник _____
(ПІБ)

Виконавець кваліфікаційної роботи _____
(ПІБ)

Зміст

Анотація

Кваліфікаційна робота присвячена дослідженню та кількісному оцінюванню якості україно-англійського автоматичного перекладу усного мовлення. Основну увагу зосереджено на аналізі семантичної точності, типових помилок перекладу та міжфразової когерентності.

Як матеріал дослідження використано корпус із 12 сегментів аудіозаписів політичних звернень українською мовою загальною тривалістю близько 45 хвилин. Обробка виконана за допомогою каскадного конвеєра, що включає автоматичне розпізнавання мовлення (ASR, модель Whisper medium) та машинний переклад (MT, модель OPUS-MT). Додатково проведено порівняльний аналіз із чотирма великими мовними моделями (LLM) — GPT-4o, Claude Sonnet 4, Gemini 2.0 Flash та Claude Haiku 4.5. Результати оцінено за системою кількісних метрик: WER, CER для ASR; BLEU, BERTScore-F1, METEOR для MT. Додатково проведено ручну анотацію помилок (пропуски, спотворення, додавання) та розроблено автоматичну композитну метрику міжфразової когерентності C_{auto}, що поєднує флюентність (perplexity), точність іменованих сутностей, збереження дискурсивних маркерів та семантичну зв'язність сусідніх речень.

Основні результати: якість ASR — WER = 8,62%, CER = 2,10%; якість MT (OPUS-MT) — BLEU = 23,12, BERTScore-F1 = 0,397, METEOR = 0,452; якість MT (LLM) — BLEU = 55–57, BERTScore-F1 = 0,71–0,74, METEOR = 0,72–0,75, що на 60–147% перевищує показники OPUS-MT ($p < 0,001$). Виявлено статистично значущу кореляцію між помилками ASR та якістю MT ($r = -0,59$ для CER–BLEU, $p < 0,05$), що підтверджує каскадний ефект поширення помилок. Ручна анотація ідентифікувала 114 помилок перекладу OPUS-MT, з яких 79,8% — спотворення (переважно власних назв та географічних назв). За автоматичною оцінкою когерентності LLM статистично значущо переважають OPUS-MT (C_{auto} = 3,90–4,19 vs 3,34, Краскел-Уолліс $p < 0,001$); локалізовано джерело переваги — вища флюентність та точність передачі іменованих сутностей. Запропоновано та інтервенційно перевірено метод перемикання коду з урахуванням іменованих

сутностей (entity-aware code-switching): виявлення іменованих сутностей у ASR-тексті з подальшою підстановкою канонічних англійських форм у вхідний текст перед OPUS-MT та післяперекладна корекція (post-MT repair) через fuzzy-зіставлення. Метод з доменним словником на 176 записів дає BLEU = 20,33 (+6,9% над baseline, Вілкоксон $p = 0{,}046$) та Entity F1 = 0,610 (+36%), що закриває 40% розриву до GPT-4o за точністю передачі власних назв.

Ключові слова: автоматичний переклад мовлення, якість перекладу, WER, BLEU, BERTScore, каскадний ефект, Whisper, OPUS-MT, великі мовні моделі, перемикання коду з урахуванням сутностей, іменовані сутності, українська мова.

1 Вступ

1.1 Актуальність

Якість автоматичного перекладу усного мовлення залишається відкритим питанням, особливо для мовних пар з морфологічно багатою та менш представленою у навчальних даних мовою [26]. Пара українська–англійська має складну морфологію, вільний порядок слів та обмежений обсяг паралельних корпусів; в умовах активної міжнародної комунікації України потреба в якісному перекладі політичного дискурсу зростає.

Більшість досліджень оцінюють окремі компоненти системи — автоматичне розпізнавання мовлення (англ. Automatic Speech Recognition, ASR) або машинний переклад (англ. Machine Translation, MT) — без аналізу накопичення помилок уздовж конвеєра [28, 21] та їх впливу на зв'язність перекладу для слухача. Комплексне дослідження семантичної точності, каскадного ефекту та міжфразової когерентності для пари українська–англійська визначає актуальність роботи.

1.2 Об'єкт, предмет, мета і завдання дослідження

Об'єктом дослідження є якість україно-англійського автоматичного перекладу усного мовлення.

Предметом дослідження є семантична точність, міжфразова когерентність та каскадний ефект поширення помилок в україно-англійському автоматичному перекладі усного мовлення, зокрема вплив помилок у передачі іменованих сутностей на якість перекладу та можливості їх зменшення засобами локальної конвеєрної обробки.

Метою кваліфікаційної роботи є кількісне оцінювання якості україно-англійського автоматичного перекладу усного мовлення з акцентом на семантичну точність, каскадний ефект поширення помилок та міжфразову когерентність, а також розроблення й експериментальна валідація локального методу зменшення каскадного ефекту для точнішої передачі

іменованих сутностей на основі перемикання коду з урахуванням іменованих сутностей (entity-aware code-switching).

Досягнення поставленої мети вимагає розв'язання таких задач:

1. Розглянути сучасні підходи до автоматичного перекладу мовлення та методи оцінювання його якості.
2. Визначити особливості мовної пари українська–англійська, що впливають на якість автоматичного перекладу.
3. Сформувати корпус дослідження на основі записів політичних звернень українською мовою та відповідних референсних перекладів.
4. Обрати та обґрунтувати систему метрик для оцінювання семантичної точності, когерентності та інших характеристик перекладу.
5. Провести експериментальне оцінювання якості перекладу обраних систем (спеціалізованої моделі нейронного машинного перекладу (NMT) та великих мовних моделей (LLM)) та виконати статистичний аналіз результатів.
6. Дослідити каскадний ефект — вплив помилок розпізнавання мовлення на якість машинного перекладу.
7. Порівняти якість перекладу спеціалізованої NMT-моделі та великих мовних моделей для цільової мовної пари та домену.
8. Запропонувати та експериментально валідувати локальний конвеєрний метод зменшення каскадного ефекту для точної передачі іменованих сутностей у каскадних uk→en системах.

1.3 Методи дослідження

Семантична точність оцінюється метриками BLEU [5], BERTScore [6] та METEOR [7]; міжфразова когерентність — розробленою композитною метрикою C_{auto} (perplexity, F_1 іменованих сутностей, дискурсивні маркери, семантична зв'язність). Помилки перекладу класифікуються за типами (пропуски, спотворення, додавання). Для статистичного аналізу застосовано кореляції Пірсона та Спірмена, критерії Манна-Уітні (попарно) та Краскела-Уолліса (множинно).

1.4 Наукова новизна

Наукова новизна роботи полягає у:

- комплексному кількісному оцінюванні якості україно-англійського автоматичного перекладу усного мовлення з використанням системи автоматичних метрик (BLEU, BERTScore, METEOR) та ручної анотації помилок;
- кількісному дослідженні каскадного ефекту поширення помилок ASR на якість MT з виявленням статистично значущих кореляцій;
- порівняльному аналізі спеціалізованої NMT-моделі та великих мовних моделей для перекладу політичного дискурсу з української на англійську;
- розробці та валідації автоматичної композитної метрики міжфразової когерентності C_{auto} (флюентність, точність іменованих сутностей, збереження дискурсивних маркерів, семантична зв'язність) та її застосуванні для порівняльного аналізу чотирьох MT-систем для пари українська–англійська;
- розробці та інтервенційній перевірці конвеєрного методу перемикання коду з урахуванням іменованих сутностей (entity-aware code-switching) зі стадією післяперекладної корекції (post-MT repair) для зменшення каскадного ефекту в низькоресурсному перекладі; експериментально показано, що запропонований метод забезпечує єдиний серед локальних методів пост-корекції ASR статистично значущий приріст BLEU, а обмеження якості каскадного конвеєра зумовлені переважно MT-моделлю, а не якістю ASR-входу.

1.5 Практичне значення

Запропонована методологія оцінювання застосовна для порівняльного аналізу систем перекладу мовлення в дипломатично-політичній сфері, а порівняння OPUS-MT з великими мовними моделями надає практичні рекомендації щодо вибору MT-компонента каскадного конвеєра залежно від вимог до латентності та якості. Запропонований конвеєр перемикання коду з ура-

хуванням іменованих сутностей (entity-aware code-switching) є готовим до застосування локальним рішенням для доменів з обмеженим набором іменованих сутностей, де критичними є конфіденційність, вартість та відтворюваність. Окремі результати представлено на студентських семінарах та внутрішніх обговореннях на кафедрі.

2 Основні означення

У цьому розділі наведено основні поняття [1]: формальні постановки задач ASR та MT, каскадна архітектура й метрики оцінювання якості перекладу.

2.1 Автоматичне розпізнавання мовлення

Автоматичне розпізнавання мовлення (Automatic Speech Recognition, ASR) — це задача перетворення мовного аудіосигналу у текстову послідовність. Формально ASR подають як задачу знаходження найбільш імовірної текстової послідовності y за наявності акустичного сигналу x:

$$\hat{y} = \arg\max_{y} P(y \mid x).$$

Сучасні ASR-системи базуються на трансформерних end-to-end архітектурах [3, 4], які безпосередньо моделюють $P(y \mid x)$ без явного поділу на акустичну та мовну моделі. Для морфологічно багатих мов (зокрема української) застосовуються субслівні одиниці (BPE, unigram LM).

Основною метрикою якості ASR є *Word Error Rate* (WER): $\mathrm{WER} = (S+D+I)/N$, де S, D, I — кількість замін, вилучень, вставок, N — кількість слів у референсі. Для морфологічно багатих мов додатково використовується *Character Error Rate* (CER) — аналогічно на рівні символів.

2.2 Машинний переклад

Машинний переклад (Machine Translation, MT) — це задача автоматичного перетворення тексту з мови джерела на мову призначення зі збереженням семантичного змісту. У нейронному машинному перекладі задача формалізується як

$$\hat{z} = \arg\max_{z} P(z \mid y),$$

де y — вхідна текстова послідовність мовою джерела, а z — відповідна послідовність мовою призначення.

Найбільш поширеною архітектурою для МТ є *Transformer* [2] з механізмом самоуваги (self-attention).

2.3 Каскадна архітектура перекладу мовлення

Автоматичний переклад мовлення (Speech Translation, ST) — перетворення мовлення однією мовою на текст іншою. У дослідженні розглядається офлайн-режим у каскадній архітектурі:

$$x(t) \xrightarrow{\text{ASR}} \hat{y} \xrightarrow{\text{MT}} \hat{z},$$

де $x(t)$ — вхідний сигнал, $\hat{y}$ — розпізнаний текст джерела, $\hat{z}$ — переклад. Ключовою особливістю є *каскадний ефект поширення помилок* [28]: помилки ASR потрапляють на вхід МТ і можуть суттєво погіршити якість фінального перекладу.

2.4 Семантична точність перекладу

Семантична точність — ступінь збереження змісту оригінального висловлювання у перекладі $\hat{z}$ відносно референсу z^*. Оцінюється автоматичними метриками BLEU [5], BERTScore [6] та METEOR [7]; детальний опис формул — у підрозділі 4.5.

2.5 Міжфразова когерентність

Міжфразова когерентність — це ступінь контекстуальної зв'язності між послідовними сегментами перекладу. У каскадних системах перекладу мовлення кожен сегмент обробляється незалежно, що означає, що система працює з обмеженим контекстом, що може призводити до порушення зв'язності: вибір значення полісемічного слова у попередньому сегменті може не узгоджуватися з контекстом наступного, займенникові посилання можуть втрачати антецедент, а тематичні переходи — залишатися непозначеними.

Порушення когерентності є критичним для слухача, оскільки навіть за умови високої семантичної точності окремих сегментів втрата зв'язності між ними може суттєво ускладнити розуміння загального змісту повідомлення.

2.6 Класифікація помилок перекладу

Для детального аналізу якості перекладу застосовують класифікацію помилок за типами:

- *Пропуски* (omissions) — частина змісту оригіналу відсутня у перекладі;
- *Спотворення* (distortions) — зміст оригіналу передано невірно;
- *Додавання* (additions) — у перекладі присутня інформація, якої немає в оригіналі.

Для кількісного оцінювання обчислюють відповідні коефіцієнти:

$$R_{\text{omit}} = \frac{N_{\text{omit}}}{N_{\text{total}}}, \quad R_{\text{dist}} = \frac{N_{\text{dist}}}{N_{\text{total}}}, \quad R_{\text{add}} = \frac{N_{\text{add}}}{N_{\text{total}}},$$

де $N_{\text{omit}}, N_{\text{dist}}, N_{\text{add}}$ — кількість сегментів із відповідним типом помилки, а N_{total} — загальна кількість сегментів.

2.7 Огляд літератури

Каскадні ST-системи залишаються конкурентоспроможними для більшості мовних пар завдяки модульності та покомпонентному оцінюванню [19, 29]. Каскадний ефект досліджено в [28] (підвищення WER на 10% знижує BLEU на 5–15 пунктів) та в [21, 20], де показано, що ефект найбільш критичний для сегментів із високою щільністю іменованих сутностей — що узгоджується з результатами цього дослідження.

Для української Whisper [8] демонструє WER 10–15%, тоді як wav2vec 2.0 [24] потребує fine-tuning. Пара uk–en належить до низькоресурсних [26]; основним джерелом даних є корпус OPUS [25], на якому натреновано модель OPUS-MT [10] (MarianNMT [16], BLEU 25–30 на FLORES [18]); мультимовна NLLB-200 [11] — 35–38.

LLM-системи (GPT, Claude, Gemini) перевищують спеціалізовані NMT для більшості мовних пар [27, 23], особливо у доменах з високою щільністю іменованих сутностей. Стандартні автоматичні метрики (BLEU, BERTScore, METEOR) мають відомі обмеження [9]; для детального аналізу помилок застосовується фреймворк MQM [22]. Комплексних досліджень якості перекладу для пари українська–англійська не проводилось.

2.8 Висновки до розділу 2

У розділі формалізовано задачі ASR та MT, описано каскадну архітектуру, метрики оцінювання та класифікацію помилок перекладу. Огляд літератури показав, що комплексне дослідження якості перекладу для пари uk–en не проводилось.

3 Корпус дослідження

3.1 Джерела та критерії відбору матеріалу

Матеріалом дослідження є відеозаписи звернень з офіційного вебпорталу Офісу Президента (`president.gov.ua`). Джерело обрано через однорідність аудіо-якості (студійний запис, один мовець, відсутність шуму) та належність до політико-дипломатичного дискурсу — домену з високими вимогами до точності перекладу. Критерії відбору: тривалість 5–15 хв, тематика міжнародних відносин/безпекової політики, відсутність технічних перешкод. Відібрано 7 аудіозаписів за січень–лютий 2025 року.

3.2 Підготовка корпусу

Аудіозаписи завантажено через `yt-dlp` і конвертовано у WAV (PCM 16-bit, 44.1 kHz, моно) за допомогою FFmpeg. Для отримання оптимальної тривалості вхідних фрагментів для ASR файли розділено на сегменти 3–5 хв скриптом на базі `pydub`: точки розрізки обираються на природних паузах (RMS нижче -40 дБFS протягом ≥ 500 мс) у межах цільового часового вікна. Загалом отримано 12 аудіосегментів (таблиця 1).

Табл. 1: Склад корпусу дослідження

№	Назва запису	К-сть сегм.	Прим. трив.
1	В Еміратах сьогодні	2	≈8 хв
2	Діалог з Америкою	2	≈8 хв
3	Знаємо, що росіяни	2	≈7 хв
4	Командна ланка	1	≈4 хв
5	Наша стратегія	2	≈7 хв
6	Отримуємо хороші	1	≈4 хв
7	Поведінка	2	≈7 хв
	Разом	**12**	**≈45 хв**

3.2.1 Транскрипція та ручна корекція

Для автоматичної транскрипції використано Faster-Whisper medium [8] (мова: `uk`, beam size = 5, температура = 0, VAD-фільтр та `condition_on_previous_text` увімкнено). Кожен текстовий файл вручну перевірено та відкориговано. Типові помилки ASR — спотворення власних назв («Рустему Міров» замість «Рустем Умєров», «Богдановим» замість «Будановим»), географічних назв («Кривирі» замість «Кривий Ріг») та лексичні підміни («вбойнів» замість «воїнів»). У 12 сегментах виявлено та виправлено близько 60–80 помилок.

3.2.2 Створення референсних перекладів

Для кожного українського сегмента створено референсний переклад англійською (gold standard) із збереженням політико-дипломатичного регістру, точності власних назв та структури висловлювань для коректного посегментного зіставлення. Переклади виконано вручну з подальшою верифікацією фактичної точності та термінології.

3.3 Висновки до розділу 3

Сформовано корпус із 12 аудіосегментів (≈45 хв, 7 звернень) з відкоригованими транскрипціями та паралельними референсними перекладами.

4 Архітектура та компоненти системи

Для проведення експериментального оцінювання якості перекладу мовлення було обрано каскадну архітектуру (див. підрозділ 2.3), у якій етапи розпізнавання мовлення та машинного перекладу працюють послідовно: вхідний аудіосигнал українською $x(t)$ перетворюється на текст $\hat{y}$ (ASR), який потім перекладається на англійську $\hat{z}$ (MT).

Каскадний підхід обрано з огляду на такі переваги:

- *модульність* — можливість замінювати окремі компоненти та порівнювати різні MT-системи при одному й тому ж ASR-виході;
- *інтерпретованість* — якість кожного етапу оцінюється окремо (WER/CER для ASR, BLEU/BERTScore/METEOR для MT), що дозволяє локалізувати джерела помилок;
- *аналіз каскадного поширення помилок* — можливість дослідити, яким чином помилки розпізнавання мовлення впливають на якість фінального перекладу;

Альтернативою каскадному підходу є end-to-end моделі, зокрема SeamlessM4T [12], що здійснюють переклад мовлення напряму, без проміжного текстового подання. Однак якість таких моделей для мовної пари українська–англійська на момент проведення дослідження поступається каскаду з окремих компонентів [12]. Крім того, end-to-end підхід унеможливлює якісний аналіз помилок, що є завданням цієї роботи.

4.1 ASR-компонент: Whisper medium

Для автоматичного розпізнавання мовлення обрано модель OpenAI Whisper medium [8] — мультимовну ASR-модель, натреновану на приблизно 680 000 годин слабко-розміченого аудіо з мережі. Модель підтримує 99 мов, включно з українською, та застосовує zero-shot підхід, що не потребує дотренування для конкретної мови чи домену.

Whisper medium (769 млн параметрів) обрано як баланс між якістю та обчислювальними вимогами; типове значення WER для українського мов-

лення — 10–15% на чистому аудіо. Розглянуті альтернативи: Whisper large-v3 (вища точність, але повільніший), Whisper turbo (швидший, нижча точність на малоресурсних мовах), wav2vec 2.0 (потребує fine-tuning на українських даних) та комерційний Google Speech-to-Text (обмежена відтворюваність). Для прискорення інференсу використано Faster-Whisper на базі CTranslate2 [17].

4.2 MT-компонент: OPUS-MT

Для етапу машинного перекладу обрано спеціалізовану NMT-модель Helsinki-NLP/opus-mt-uk-en [10], побудовану на архітектурі Transformer у фреймворку MarianNMT [16]. Модель натренована на корпусі OPUS, що включає CCMatrix, WikiMatrix, Europarl та інші паралельні дані. Розмір моделі — приблизно 74 млн параметрів, максимальна довжина контексту — 512 токенів. Очікуваний BLEU на бенчмарку FLORES-101 [18] для uk→en становить 25–30 залежно від домену. Модель повністю відкрита (ліцензія MIT), працює локально та забезпечує високу швидкість інференсу.

Вибір OPUS-MT зумовлений відкритістю (повна відтворюваність без залежності від комерційних API), спеціалізацією на парі uk→en та ефективністю на CPU. Серед альтернатив розглянуто NLLB-200 [11] (1,3 млрд параметрів, BLEU 35–38), однак OPUS-MT обрано як легковісний baseline, що дозволяє чітко виділити каскадний ефект без впливу масштабу моделі.

Як додаткове дослідження якості перекладу також залучено чотири великі мовні моделі (GPT-4o, Claude Sonnet 4, Claude Haiku 4.5 та Gemini 2.0 Flash), опис та результати яких наведено у підрозділі 5.6.

4.3 Реалізація pipeline

Pipeline (ASR + MT) реалізовано на Python із використанням бібліотек `faster-whisper` та Hugging Face Transformers (параметри OPUS-MT — за замовчуванням). Обробка корпусу (≈45 хв аудіо) на CPU зайняла 518 с, що відповідає коефіцієнту реального часу RTF = 0,26× (ASR — 361,3 с,

MT — 156,7 с), що підтверджує придатність системи для офлайн-обробки великих корпусів.

4.4 Архітектура запропонованого pipeline

Попередній аналіз помилок baseline-системи (підрозділ 5.3) виявляє, що 46,2% спотворень OPUS-MT припадає на власні назви, ще 16,5% — на географічні, 13,2% — на військову термінологію. Кореляційний аналіз (підрозділ 5.4) підтверджує каскадний ефект: $r = -0{,}59$ ($p = 0{,}042$) для CER–BLEU. Ця діагностика мотивує pipeline-level рішення, яке цілеспрямовано обробляє клас іменованих сутностей перед поданням у OPUS-MT.

4.4.1 Загальна схема

Запропонований pipeline розширює базовий трьома модулями, які втручаються у два моменти — перед етапом машинного перекладу та після нього:

$$x(t) \xrightarrow{\text{ASR}} \hat{y} \xrightarrow[+\,\text{dict}]{\text{NER}} \hat{y}' \xrightarrow[\text{switch}]{\text{code-}} \hat{y}'' \xrightarrow{\text{MT}} \hat{z}' \xrightarrow[\text{repair}]{\text{post-MT}} \hat{z}. \tag{1}$$

На відміну від підходів із корекцією ASR-транскрипції ($\hat{y} \to \hat{y}'$), що обмежуються виправленнями у межах вхідної мови, запропонований метод *обходить* обмеження OPUS-MT у передачі власних назв: канонічні англійські форми підставляються у вхідний український текст до перекладу, а після перекладу виконується корекція решток спотворень.

4.4.2 Доменний словник

Основою методу є словник $D : \text{UA} \to \text{EN}$, що відображає канонічні та відмінкові форми іменованих сутностей українською на їхні канонічні англійські переклади. Словник містить 176 записів у п'яти категоріях (таблиця 2), сформованих вручну на основі аналізу референсних перекладів корпусу.

Табл. 2: Структура доменного словника D

Категорія	**Записів**	**Приклади**
Персоналії (особи)	62	Рустем Умєров → Rustem Umerov, Кирило Буданов → Kyrylo Budanov, Михайло Федоров → Mykhailo Fedorov
Географічні назви	55	Кривий Ріг → Kryvyi Rih, Дніпровщина → Dnipropetrovsk region, Запоріжжя → Zaporizhzhia
Країни та міжнародні організації	30	Євросоюз → European Union, Сполучених Штатів Америки → United States of America
Установи та абревіатури	15	СБУ → SBU, ГУР → HUR, ДСНС → State Emergency Service
Військові підрозділи та інше	14	окремого штурмового батальйону → Separate Assault Battalion, Patriot PAC-3, PEARL
Загалом	**176**	—

4.4.3 Виявлення сутностей

Модуль виявлення працює у два проходи. Перший прохід шукає у ASR-тексті $\hat{y}$ точні підрядки, що відповідають ключам словника D, застосовуючи greedy-стратегію (співпадіння довшої форми має пріоритет над коротшою). Другий прохід обробляє невкриті позиції за допомогою фонетичної дистанції, що дозволяє знаходити спотворені варіанти іменованих сутностей, які могли утворитися внаслідок помилок ASR.

Фонетична транслітерація ϕ відображає кирилицю на латинську фонети-

чну форму зі згортанням еквівалентних звуків:

$$\phi : \text{и, і, й, ї} \to \text{i}; \quad \text{г, х} \to \text{h}; \quad \text{щ} \to \text{sh}; \quad \text{ь, ’} \to \varnothing; \quad \ldots \tag{2}$$

Дистанція між двома рядками обчислюється як нормована Левенштейн-дистанція їхніх транслітерованих форм:

$$d_\phi(w_1, w_2) = \frac{\mathrm{Lev}\big(\phi(w_1), \phi(w_2)\big)}{\max\big(|\phi(w_1)|,\ |\phi(w_2)|\big)}. \tag{3}$$

Вікно w довжиною 1–3 слова класифікується як спотворений варіант сутності $k \in D$, якщо $d_\phi(w, k) < \tau_\phi$, де поріг $\tau_\phi = 0{,}18$ визначено через мінімізацію WER на підвибірці з 3 сегментів.

4.4.4 Code-switching

Кожна виявлена сутність w у тексті $\hat{y}$ замінюється на свою канонічну англійську форму $D(w)$ у межах вхідної послідовності:

$$\hat{y}'' = \mathrm{replace}\big(\hat{y},\ w,\ D(w)\big) \quad \forall w \in E(\hat{y}), \tag{4}$$

де $E(\hat{y})$ — множина виявлених сутностей. Одержаний текст $\hat{y}''$ містить кириличний контекст та латинські власні назви.

Вибір такої стратегії зумовлений емпіричним спостереженням: альтернативний варіант із UPPERCASE-маркерами (`XQQAA`, `XQQAB`) як перша версія методу показав незадовільну стабільність — BPE-токенізатор MarianNMT розбиває такі токени на subwords. Пряма підстановка латинських канонічних форм, що належать до розподілу тренувальних даних моделі opus-mt-uk-en, дає значно кращу стійкість (кількісні результати — підрозділ 5.5).

4.4.5 Post-MT repair

OPUS-MT частково спотворює вставлені англійські форми через BPE-токенізацію та автоматичне перекомпонування: наприклад, *Rustem Umerov* переходить у *Rust Umerov*, *Sumy region* — у *Sumland*. Модуль post-MT repair відновлює канонічні форми у перекладі $\hat{z}'$ через fuzzy-зіставлення.

Для кожної канонічної форми $D(w)$ зі списку вставлених сутностей, якої немає у $\hat{z}'$ дослівно, алгоритм перебирає ngram-вікна довжини $n_{D(w)} - 1$, $n_{D(w)}$, $n_{D(w)} + 1$ (де $n_{D(w)}$ — кількість слів у $D(w)$) та обчислює нормовану Левенштейн-дистанцію на латиниці. Якщо для вікна v виконується $d_L(v, D(w)) < \tau_L$ (де $\tau_L = 0{,}30$), вікно замінюється на $D(w)$:

$$\hat{z} = \mathrm{replace}\big(\hat{z}',\, v^*,\, D(w)\big), \quad v^* = \arg\min_v d_L(v, D(w)). \tag{5}$$

4.4.6 Алгоритм

Табл. 3: Алгоритм entity-aware code-switching pipeline

Entity-aware code-switching pipeline
Вхід: UA-транскрипція $\hat{y}$, словник D, пороги τ_ϕ, τ_L.
Вихід: EN-переклад $\hat{z}$.
1. $E \leftarrow$ виявити сутності в $\hat{y}$ (точне співпадіння з D; phonetic: $d_\phi < \tau_\phi$).
2. $\hat{y}'' \leftarrow$ замінити кожне $w \in E$ на $D(w)$ у $\hat{y}$.
3. $\hat{z}' \leftarrow$ OPUS-MT($\hat{y}''$).
4. для кожного $D(w) \in E$, відсутнього в $\hat{z}'$ дослівно: якщо існує вікно $v^* = \arg\min_v d_L(v, D(w))$ з $d_L(v^*, D(w)) < \tau_L$, замінити v^* на $D(w)$.
5. Повернути $\hat{z}$.

Результати експериментального оцінювання у порівнянні з baseline та альтернативними методами наведено у підрозділі 5.5.

4.5 Система метрик оцінювання якості перекладу

Для об'єктивної оцінки якості машинного перекладу в даному дослідженні застосовано комплексний підхід, що поєднує автоматичні метрики та людську оцінку. Такий підхід дозволяє отримати всебічну характеристику якості перекладу, враховуючи як формальну відповідність референсному тексту, так і семантичну адекватність та природність мовлення.

4.5.1 Автоматичні метрики

BLEU (Bilingual Evaluation Understudy). Метрика BLEU [5] базується на порівнянні n-грам між машинним та референсним перекладами:

$$\text{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right), \quad BP = \min\left(1,\ e^{1-r/c}\right), \tag{6}$$

де p_n — модифікована n-грамна точність, $w_n = 1/N$ — ваги (стандартно $N = 4$), BP — штраф за стисліcть, c та r — довжини кандидата й референсу. BLEU обрано як стандартну метрику для порівняння MT-систем, що добре корелює з людською оцінкою на корпусному рівні [9].

BERTScore. BERTScore [6] оцінює семантичну подібність на рівні контекстуалізованих ембедингів BERT. Для кожного токена обчислюється максимальна косинусна подібність із токенами референсу; з цих значень формуються precision, recall та F_1:

$$F_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{7}$$

На відміну від BLEU, BERTScore враховує семантичну еквівалентність перефразувань та синонімів, що особливо важливо для оцінки окремих речень.

METEOR. METEOR [7] враховує не лише точний збіг, але й стемінг та синоніми (WordNet). Метрика обчислює гармонійне середнє точності й повноти уніграм зі штрафом за фрагментацію:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}), \quad F_{\text{mean}} = \frac{10PR}{R + 9P}, \quad \text{Penalty} = 0{,}5 \cdot \left(\frac{\#\text{chunks}}{\#\text{matches}}\right)^3 \tag{8}$$

METEOR демонструє кращу кореляцію з людською оцінкою на рівні речень, особливо для морфологічно багатих мов.

4.5.2 Людська експертна оцінка

Автоматичні метрики доповнено ручною анотацією помилок за типами (підрозділ 2.6: пропуски, спотворення, додавання) з виокремленням критичних помилок, що змінюють зміст повідомлення (розділ 5). Така тріангуляція забезпечує діагностичну інформацію, недоступну для агрегованих автоматичних метрик.

4.6 Дизайн експерименту

4.6.1 Процедура оцінювання

Автоматичні метрики обчислюються посегментно з використанням стандартизованих інструментів: BLEU — через SacreBLEU [9] з токенізацією "13a"; BERTScore — з моделлю `microsoft/deberta-xlarge-mnli` та параметром `rescale_with_baseline=True`; METEOR — реалізація NLTK із WordNet. Експертне оцінювання включає анотацію помилок за класифікацією підрозділу 2.6 (з фіксацією типу, серйозності та контексту) та попередню оцінку міжфразової когерентності для OPUS-MT за 5-бальною шкалою.

4.6.2 Автоматична оцінка міжфразової когерентності

Ручна оцінка когерентності виконується одним анотатором на обмеженій підвибірці пар сегментів і має суб'єктивний характер. Для підвищення надійності та масштабованості оцінювання когерентності розроблено автоматичну методологію, яка реалізує функцію $C_i = g(\hat{z}_i, \hat{z}_{i+1}, z_i^*, z_{i+1}^*)$, введену у розділі 5, як композицію чотирьох незалежних компонентів. Методологія застосована до 12 сегментів кожної з чотирьох MT-систем (OPUS-MT, GPT-4o, Claude, Gemini) та референсного перекладу.

Компонент 1: Флюентність на основі мовної моделі. Для кожного сегмента $\hat{z}$ обчислюється perplexity за моделлю `distilgpt2`:

$$\mathrm{PPL}(\hat{z}) = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log P(w_t \mid w_{<t})\right), \tag{9}$$

де $w_1, \ldots, w_T$ — послідовність токенів. Чим нижча perplexity, тим природніша мовна послідовність [1]. Для забезпечення порівнюваності між сегментами perplexity системи нормується відносно perplexity референсного перекладу того ж сегмента:

$$F_{\text{flu}}(\hat{z}) = \min\left(1, \frac{\mathrm{PPL}(z^*)}{\mathrm{PPL}(\hat{z})}\right). \tag{10}$$

Значення $F_{\text{flu}} = 1$ відповідає флюентності на рівні референсу або вищій.

Компонент 2: Точність іменованих сутностей. Іменовані сутності категорій PERSON, GPE, LOC, ORG та NORP виокремлюються моделлю `en_core_web_sm` бібліотеки spaCy. Для нормалізованих сутностей системи E_{sys} та референсу E_{ref} обчислюється F_1:

$$F_{\text{ent}} = \frac{2 \cdot P_{\text{ent}} \cdot R_{\text{ent}}}{P_{\text{ent}} + R_{\text{ent}}}, \quad P_{\text{ent}} = \frac{|E_{\text{sys}} \cap E_{\text{ref}}|}{|E_{\text{sys}}|}, \quad R_{\text{ent}} = \frac{|E_{\text{sys}} \cap E_{\text{ref}}|}{|E_{\text{ref}}|}. \tag{11}$$

Метрика прямо відображає здатність системи зберігати референти (імена, географічні назви, організації) між сегментами.

Компонент 3: Збереження дискурсивних маркерів. Дискурсивні маркери (*however*, *therefore*, *because*, *meanwhile* тощо — усього 50 одиниць у словнику) є лексичними сигналами риторичної структури тексту. Для системи обчислюється частка маркерів референсу, що збереглися в перекладі:

$$F_{\text{disc}} = \frac{|M_{\text{sys}} \cap M_{\text{ref}}|}{|M_{\text{ref}}|}. \tag{12}$$

Компонент 4: Семантична зв'язність сусідніх речень. Для кожного сегмента речення $s_1, \ldots, s_n$ кодуються моделлю `sentence-transformers/all-mpnet-base-v2` у векторні представлення $\mathbf{e}_1, \ldots, \mathbf{e}_n$,
після чого обчислюється середня косинусна подібність сусідніх речень:

$$F_{\text{intra}} = \frac{1}{n-1} \sum_{i=1}^{n-1} \frac{\langle \mathbf{e}_i, \mathbf{e}_{i+1} \rangle}{\|\mathbf{e}_i\| \, \|\mathbf{e}_{i+1}\|}. \tag{13}$$

Цей показник відображає тематичну неперервність викладу й застосовується як проксі тематичної зв'язності (а не повної когерентності).

Композитна оцінка C_{auto}. Чотири нормовані компоненти комбінуються у композитну оцінку, лінійно відображену на 5-бальну шкалу для зіставності з ручною оцінкою:

$$C_{\text{auto}} = 1 + 4 \cdot \big(0{,}40\, F_{\text{flu}} + 0{,}30\, F_{\text{ent}} + 0{,}15\, F_{\text{disc}} + 0{,}15\, F_{\text{intra}}\big). \tag{14}$$

Більша вага флюентності та точності сутностей обґрунтована попереднім ручним аналізом помилок (підрозділ 5.3), який показав, що саме спотворення власних назв та поламана англійська найсильніше впливають на сприйняття перекладу читачем.

Валідація методології. Для перевірки узгодженості автоматичної оцінки з ручною обчислюється кореляція Пірсона та Спірмена між середнім C_{auto} для OPUS-MT по аудіозаписах та ручними оцінками з таблиці **??**.

4.6.3 Статистичні методи

Для кожної метрики обчислюються середнє, стандартне відхилення та медіана. Зв'язки між метриками ASR та MT досліджуються коефіцієнтами кореляції Пірсона та Спірмена [15]; відмінності між групами систем — критерієм Манна-Уітні [14] (попарно) та критерієм Краскела-Уолліса (для множинного порівняння). Рівень значущості — $\alpha = 0{,}05$. Каскадний ефект досліджується через кореляційний аналіз між WER/CER та метриками

якості перекладу. Усі параметри моделей, версії бібліотек та результати задокументовані у відкритому репозиторії [30].

Табл. 4: Зведена таблиця методів оцінювання

Аспект оцінки	**Метод**	**Рівень**
Лексичний збіг	BLEU	Сегмент/корпус
Семантична подібність	BERTScore	Сегмент
Морфологічний збіг	METEOR	Сегмент
Типи помилок	Анотація помилок (MQM)	Сегмент
Когерентність (якісна)	Експертна оцінка	Пара сегментів
Когерентність (автомат.)	C_{auto} (perplexity, entity F1, discourse markers, cos_intra)	Сегмент
Якість ASR	WER, CER	Сегмент

4.7 Висновки до розділу 4

У четвертому розділі описано методологію дослідження: реалізовано каскадний pipeline (Whisper medium + OPUS-MT) з RTF = 0,26×, обґрунтовано вибір системи метрик (BLEU, BERTScore, METEOR, ручна анотація помилок) та розроблено автоматичну композитну оцінку міжфразової когерентності C_{auto}. Описано дизайн експерименту та статистичні методи аналізу (кореляції Пірсона/Спірмена, Манна-Уітні, Краскела-Уолліса).

5 Результати експериментального дослідження

У цьому розділі наведено результати експериментального оцінювання якості автоматичного перекладу усного мовлення з української на англійську мову.

5.1 Результати автоматичного розпізнавання мовлення

Якість автоматичного розпізнавання мовлення (ASR) оцінено за допомогою метрик WER (Word Error Rate) та CER (Character Error Rate). Порівняння здійснено між транскрипцією, згенерованою моделлю Whisper medium, та референсною ручною транскрипцією.

5.1.1 Загальні результати ASR

Зведену статистику метрик якості ASR наведено в таблиці 5.

Табл. 5: Зведена статистика метрик якості ASR

Метрика	**Mean**	**Std**	**Min**	**Max**	**Median**
WER	8,62%	2,82%	4,71%	13,39%	7,49%
CER	2,10%	1,10%	0,66%	4,97%	1,93%

Середнє значення WER = 8,62% свідчить про високу якість розпізнавання українського мовлення моделлю Whisper. Для порівняння, за даними бенчмарків [8], типове значення WER для української мови становить 10–15% на чистому мовленні. Отриманий результат є кращим за середній, що пояснюється високою якістю аудіозаписів (один мовець, відсутність шуму).

Низьке значення CER = 2,10% вказує на те, що більшість помилок ASR є помилками на рівні слів (заміни, пропуски), а не на рівні символів, що характерно для помилок у власних назвах та термінології.

5.1.2 Посегментний аналіз ASR

Детальні результати для кожного сегмента наведено в таблиці 6.

Табл. 6: Результати ASR за сегментами

Сегмент	**WER**	**CER**	**Слів (реф.)**	**Subst.**	**Del.**
В_Еміратах_seg001	6,16%	1,72%	341	18	2
В_Еміратах_seg002	4,71%	0,66%	191	8	0
Діалог_з_Америкою_seg001	6,70%	1,59%	388	24	2
Діалог_з_Америкою_seg002	12,88%	4,97%	163	15	2
Знаємо_що_росіяни_seg001	11,45%	2,46%	358	32	4
Знаємо_що_росіяни_seg002	9,67%	2,14%	362	32	1
Командна_ланка_seg001	13,39%	2,93%	351	36	9
Наша_стратегія_seg001	10,14%	2,29%	345	27	7
Наша_стратегія_seg002	6,86%	1,16%	277	19	0
Отримуємо_хороші_seg001	7,64%	2,22%	445	28	4
Поведінка_seg001	6,52%	1,34%	368	19	4
Поведінка_seg002	7,34%	1,70%	218	13	3
Загалом	**8,62%**	**2,10%**	**3807**	**271**	**38**

Найвищий WER спостерігається для сегментів «Командна ланка» (13,39%) та «Діалог з Америкою seg002» (12,88%). Аналіз цих сегментів показав, що підвищений рівень помилок пов'язаний із:

- високою щільністю власних назв та військової термінології;
- швидшим темпом мовлення;
- наявністю специфічних абревіатур (ГУР, СБУ, ДСНС).

Найкращі результати отримано для сегментів «В Еміратах seg002» (WER = 4,71%) та «В Еміратах seg001» (WER = 6,16%), що містять переважно загальновживану лексику.

Візуалізацію результатів ASR за сегментами наведено на рис. 1.

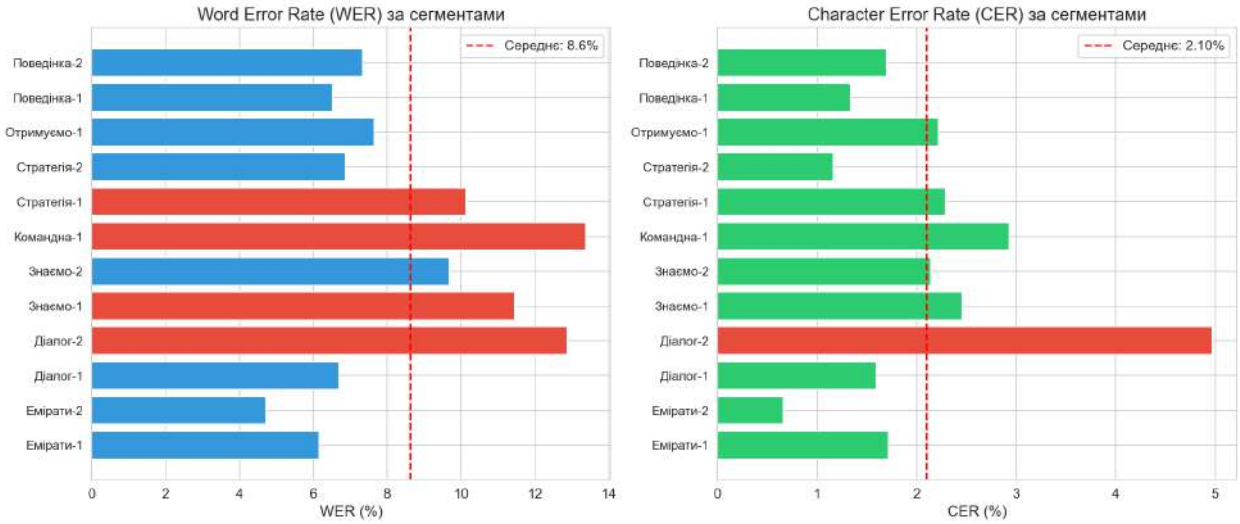

Рис. 1: Метрики якості ASR (WER та CER) за сегментами корпусу

5.1.3 Аналіз типів помилок ASR

З 3807 слів референсу зафіксовано 271 заміну (82,6% усіх помилок), 38 видалень (11,6%) та 19 вставок (5,8%). Переважання замін типове для помилок у власних назвах, де модель генерує фонетично подібні, але неправильні варіанти («Рустему Міров» замість «Рустем Умєров»). Розподіл типів помилок візуалізовано на рис. 2.

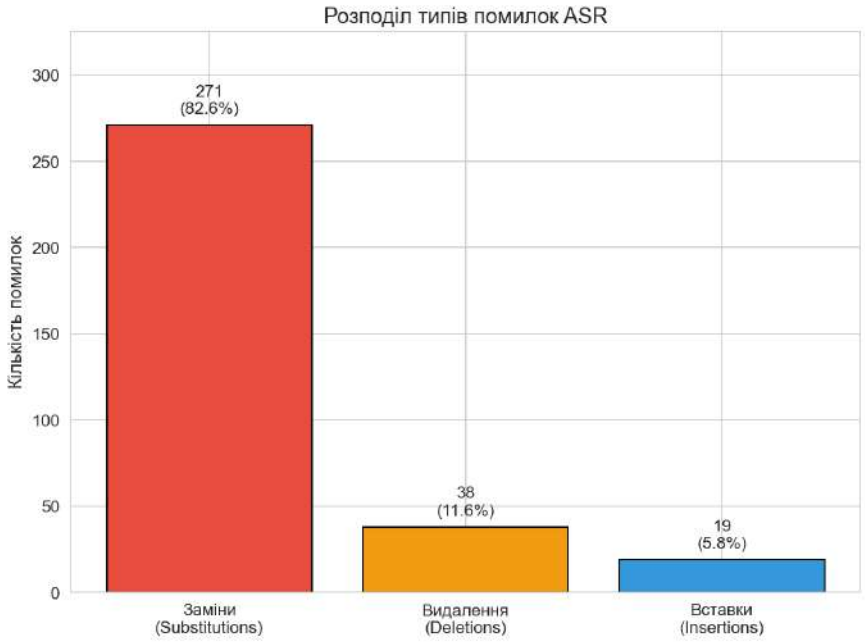

Рис. 2: Розподіл типів помилок ASR

5.2 Результати машинного перекладу

Якість машинного перекладу (MT) оцінено за допомогою трьох автоматичних метрик: BLEU, BERTScore та METEOR. Порівняння здійснено між перекладом, згенерованим моделлю Helsinki-NLP/opus-mt-uk-en, та референсним людським перекладом.

5.2.1 Загальні результати MT

Зведену статистику метрик якості перекладу наведено в таблиці 7.

Табл. 7: Зведена статистика метрик якості MT

Метрика	**Mean**	**Std**	**Min**	**Max**	**Median**
BLEU	23,12	4,73	12,37	31,44	23,02
BERTScore-F1	0,397	0,080	0,228	0,506	0,405
METEOR	0,452	0,061	0,302	0,528	0,476

BLEU = 23,12 є типовим для пари uk–en (очікувані 25–30 на FLORES-101 для загального тексту [18]); дещо нижчий результат пояснюється доменною специфікою корпусу. BERTScore-F1 = 0,397 з параметром `rescale_with_baseline=True` відповідає помірній семантичній подібності; розбіжність із BLEU зумовлена чутливістю BERTScore до стилістичних відмінностей між дослівним перекладом OPUS-MT та більш вільним референсом. METEOR = 0,452 вищий за BLEU за рахунок врахування синонімів та стемінгу.

5.2.2 Посегментний аналіз MT

Детальні результати для кожного сегмента наведено в таблиці 8.

Табл. 8: Результати МТ за сегментами

Сегмент	**BLEU**	**BERTScore-F1**	**METEOR**
В_Еміратах_seg001	21,00	0,354	0,476
В_Еміратах_seg002	22,43	0,342	0,395
Діалог_з_Америкою_seg001	31,44	0,506	0,528
Діалог_з_Америкою_seg002	12,37	0,228	0,302
Знаємо_що_росіяни_seg001	23,32	0,333	0,481
Знаємо_що_росіяни_seg002	24,18	0,437	0,477
Командна_ланка_seg001	22,71	0,361	0,477
Наша_стратегія_seg001	24,58	0,386	0,425
Наша_стратегія_seg002	19,25	0,451	0,412
Отримуємо_хороші_seg001	21,89	0,424	0,463
Поведінка_seg001	25,60	0,441	0,479
Поведінка_seg002	28,63	0,504	0,507
Середнє	**23,12**	**0,397**	**0,452**

Найвищі показники якості отримано для сегментів:

- «Діалог з Америкою seg001» (BLEU = 31,44, BERTScore-F1 = 0,506);
- «Поведінка seg002» (BLEU = 28,63, BERTScore-F1 = 0,504).

Найнижчі показники — для сегмента «Діалог з Америкою seg002» (BLEU = 12,37, BERTScore-F1 = 0,228), що корелює з найвищим WER для цього сегмента (12,88%). Це підтверджує гіпотезу про каскадний ефект помилок ASR на якість МТ.

Візуалізацію метрик якості МТ за сегментами наведено на рис. 3.

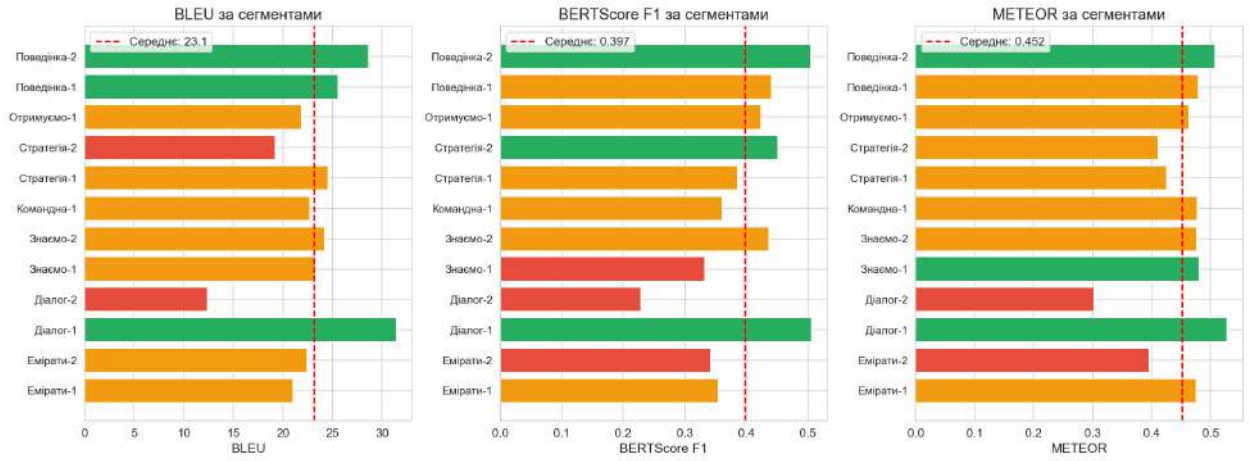

Рис. 3: Метрики якості MT (BLEU, BERTScore, METEOR) за сегментами корпусу. Кольорове кодування: зелений — вище середнього, жовтий — близько до середнього, червоний — нижче середнього

5.3 Результати ручного оцінювання якості перекладу

Для детального аналізу якості машинного перекладу OPUS-MT (baseline-модель) проведено ручну анотацію помилок із класифікацією за типами [22]: пропуски (omissions), спотворення (distortions) та додавання (additions).

5.3.1 Класифікація помилок перекладу

За результатами порівняння машинного перекладу (MT) з референсним людським перекладом ідентифіковано 114 помилок різних типів. Розподіл помилок за категоріями наведено в таблиці 9.

Табл. 9: Розподіл помилок перекладу за типами

Тип помилки	**Кількість**	**%**	***R*-коефіцієнт**
Спотворення (distortions)	91	79,8%	$R_{\text{dist}} = 1,00$
Пропуски (omissions)	19	16,7%	$R_{\text{omit}} = 0,83$
Додавання (additions)	4	3,5%	$R_{\text{add}} = 0,33$
Загалом	**114**	**100%**	—

Коефіцієнт R (rate) обчислено як частку сегментів, у яких виявлено що-

найменше одну помилку відповідного типу: $R = n_{\text{affected}}/N$, де $N = 12$ — загальна кількість сегментів. Значення $R_{\text{dist}} = 1,00$ свідчить про наявність спотворень у всіх 12 сегментах корпусу. Високе значення $R_{\text{omit}} = 0,83$ вказує на систематичні пропуски інформації в більшості сегментів.

5.3.2 Категорії спотворень

Детальний аналіз 91 спотворення дозволив виокремити шість основних категорій помилок (таблиця 10).

Табл. 10: Категорії спотворень у машинному перекладі

Категорія	**Кількість**	**%**	**Приклад**
Власні назви	42	46,2%	Rustema Mirov → Rustem Umerov
Географічні назви	15	16,5%	Crete → Kryvyi Rih
Семантичні помилки	14	15,4%	our killers → our warriors
Військова термінологія	12	13,2%	425th ice storm regiment → 425th Assault Battalion
Абревіатури	7	7,7%	MBI → Ministry of Internal Affairs
Одиниці виміру	1	1,1%	20,000 dollars → 20,000 hryvnias

Власні назви (46,2%). Найбільшу частку спотворень становлять помилки у власних назвах. Модель ASR некоректно розпізнає українські імена та прізвища, що призводить до каскадного ефекту в MT. Характерні приклади:

- «Rustem Umerov» → «Rustema Mirov»;
- «Yevhenii Khmara» → «Eugene Cloud»;
- «Keir Starmer» → «Cyrus Starmer»;

- «Andrii Sybiha» → «Andriy Sebiges».

Географічні назви (16,5%). Значну проблему становить переклад українських топонімів. Виявлено критичні помилки, коли назви українських міст замінювалися на географічно віддалені об'єкти:

- «Kryvyi Rih» → «Crete»;
- «Kryvyi Rih» → «Crimea»;
- «Mykolaiv region» → «Nicholasland»;
- «Poltava region» → «Valvas region».

Семантичні помилки (15,4%). До цієї категорії належать спотворення, що суттєво змінюють зміст повідомлення:

- «our warriors» → «our killers» (зміна конотації);
- «We also reviewed the situation» → «The eye doctor figured out»;
- «resilience» → «sewers»;
- «debris clearing» → «invasion of the wrecks».

Військова термінологія (13,2%). Помилки в назвах військових підрозділів є критичними для точності повідомлення:

- «425th Separate Assault Battalion» → «425th ice storm regiment»;
- «414th Unmanned Systems Brigade» → «414 Tachimadora»;
- «93rd Separate Mechanized Brigade» → «93 individual mechanized team».

5.3.3 Критичні помилки

Серед виявлених спотворень ідентифіковано 7 критичних помилок, які суттєво впливають на сприйняття повідомлення або містять фактичні неточності (таблиця 11).

Табл. 11: Критичні помилки машинного перекладу

Сегмент	**MT**	**Референс**	**Вплив**
B_Еміратах seg002	20,000 dollars	20,000 hryvnias	Фактична помилка (валюта)
Діалог seg001	our killers today	our warriors today	Зміна конотації
Знаємо seg001	Crete	Kryvyi Rih	Географічна плутанина
Наша_страт. seg001	Crimea	Kryvyi Rih	Політично чутлива помилка
Отримуємо seg001	Sewers	resilience	Повне спотворення
B_Еміратах seg001	The eye doctor figured out	We also reviewed	Семантична деструкція
Діалог seg002	425th ice storm regiment	425th Assault Battalion	Ідентифікація підрозділу

5.3.4 Аналіз пропусків

Виявлено 11 пропусків інформації в машинному перекладі. Найчастіше пропускаються:

- деталі військових підрозділів та операцій (5 випадків);
- уточнюючі коментарі та оцінки (4 випадки);
- назви конкретних організацій (2 випадки).

5.3.5 Автоматична оцінка міжфразової когерентності

Композитну метрику C_{auto} (підрозділ 4.6.2), що комбінує чотири компоненти — флюентність (F_{flu}), F_1-міру іменованих сутностей (F_{ent}), збереження дискурсивних маркерів (F_{disc}) та семантичну зв'язність сусідніх речень (F_{intra}) — застосовано до всіх 12 сегментів кожної з чотирьох MT-систем (OPUS-MT та три LLM-системи: GPT-4o, Claude Sonnet 4, Gemini 2.0 Flash; детальний порівняльний аналіз LLM — у підрозділі 5.6) та референсного перекладу (таблиця 12).

Табл. 12: Автоматична оцінка когерентності за системами (n = 12)

Система	C_{auto}	**PPL**	F_{flu}	F_{ent}	F_{disc}	F_{intra}
OPUS-MT	$3{,}34 \pm 0{,}46$	59,21	0,73	0,48	0,70	0,318
GPT-4o	$4{,}19 \pm 0{,}17$	44,46	0,95	0,86	0,81	0,312
Claude	$3{,}90 \pm 0{,}18$	55,80	0,76	0,85	0,79	0,302
Gemini	$4{,}18 \pm 0{,}21$	42,88	0,99	0,84	0,78	0,300
Reference	$4{,}59 \pm 0{,}01$	42,29	1,00	1,00	1,00	0,316

Візуалізацію розподілів усіх компонентів за системами наведено на рис. 4, композитну оцінку — на рис. 5, посегментний розподіл — на рис. 6.

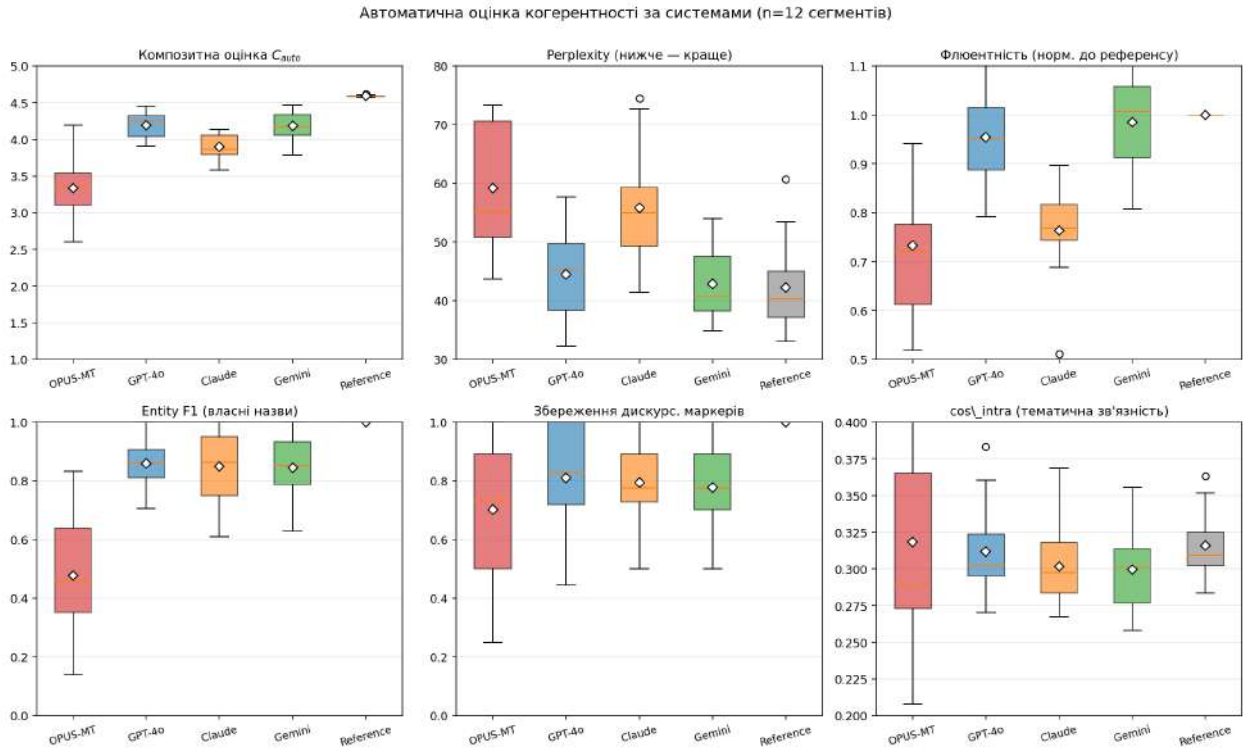

Рис. 4: Розподіл компонентів автоматичної оцінки когерентності (C_{auto}, perplexity, F_{flu}, F_{ent}, F_{disc}, F_{intra}) за системами

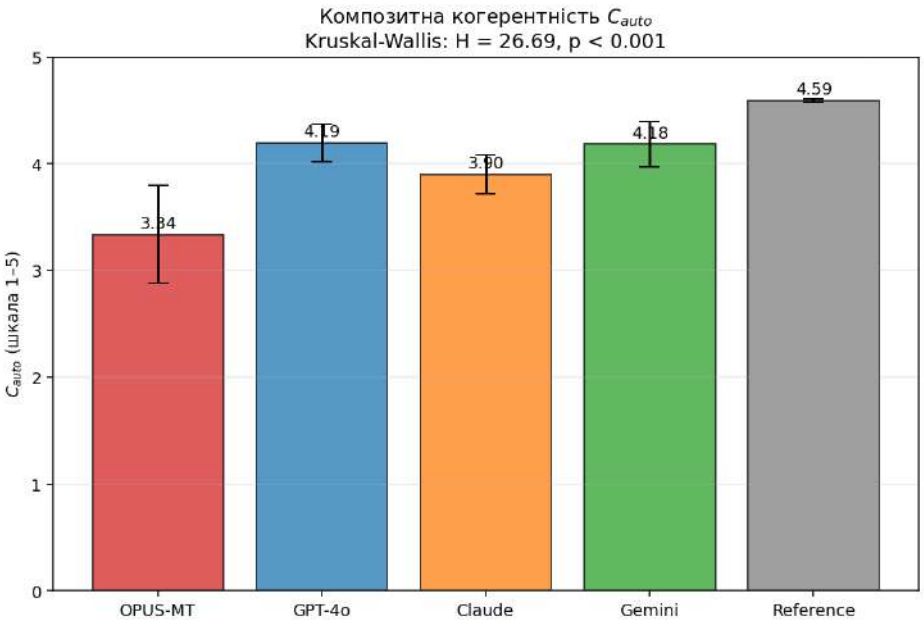

Рис. 5: Композитна оцінка C_{auto} для чотирьох MT-систем та референсу

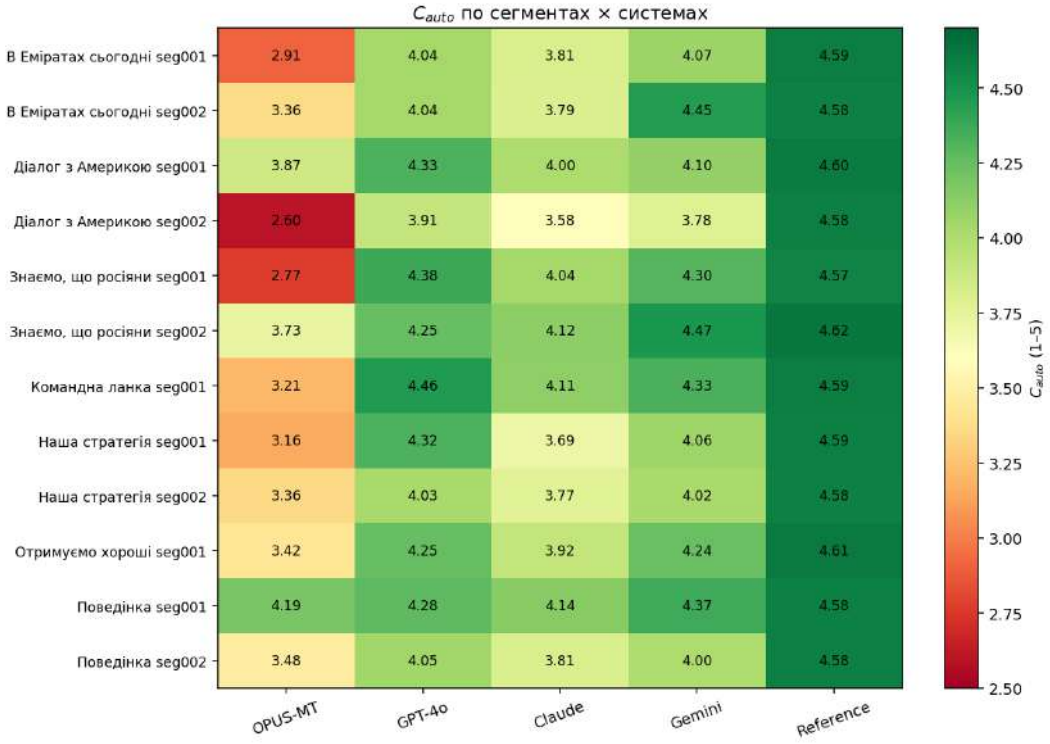

Рис. 6: Посегментний розподіл C_{auto}: системи × сегменти

Статистична значущість. Тест Краскела-Уолліса виявив статистично значущі відмінності у композитній оцінці C_{auto} між чотирма MT-системами: $H = 26{,}69$, $p < 0{,}001$. Значущі відмінності зафіксовано також для трьох з чотирьох компонентів: perplexity ($H = 20{,}17$, $p = 0{,}0002$), F_{flu} ($H = 23{,}87$,

$p < 0{,}001$), F_{ent} ($H = 20{,}99$, $p = 0{,}0001$). Для збереження дискурсивних маркерів ($H = 1{,}60$, $p = 0{,}66$) та тематичної зв'язності ($H = 0{,}78$, $p = 0{,}85$) системи статистично не відрізняються — усі MT-моделі однаково передають базові дискурсивні зв'язки та тематичну неперервність викладу.

Попарне порівняння OPUS-MT з кожною LLM-системою за критерієм Манна-Уітні (таблиця 13) підтверджує, що перевага LLM є статистично значущою для всіх трьох систем.

Табл. 13: Попарне порівняння OPUS-MT з LLM за C_{auto} (критерій Манна-Уітні)

Порівняння	U	p	**Значущість**
OPUS-MT vs GPT-4o	5	0,0001	$p < 0{,}001$
OPUS-MT vs Claude	20	0,0029	$p < 0{,}01$
OPUS-MT vs Gemini	7	0,0002	$p < 0{,}001$

Інтерпретація. Перевага LLM локалізована у двох компонентах: **флюентність** ($F_{\text{flu}} = 0{,}95$–$0{,}99$ vs $0{,}73$ для OPUS-MT, що відповідає PPL ≈ 43 vs 59,2) та **точність іменованих сутностей** ($F_{\text{ent}} = 0{,}84$–$0{,}86$ vs $0{,}48$). Останнє кількісно підтверджує висновок ручної анотації про те, що спотворення власних назв є основною слабкістю спеціалізованої NMT-моделі (46,2 % усіх помилок, таблиця 10). Для дискурсивних маркерів та тематичної зв'язності LLM не дають статистично значущої переваги: ці компоненти визначаються змістом оригіналу, а не якістю перекладу.

Валідація. Автоматичну оцінку перевірено через кореляцію з попередньою ручною оцінкою 5 пар сегментів OPUS-MT за 5-бальною шкалою (середні значення $\{3, 2, 3, 4, 4\}$ для відповідних аудіозаписів, $\bar{x}_{\text{ручн.}} = 3{,}2$): Пірсон $r = 0{,}55$ ($p = 0{,}33$), Спірмен $\rho = 0{,}79$ ($p = 0{,}11$). Сильна рангова кореляція підтверджує узгодженість методів; відсутність формальної значущості пояснюється малим $n = 5$. Близькість середніх ($\bar{x}_{\text{ручн.}} = 3{,}2$, $\bar{x}_{C_{\text{auto}}} = 3{,}34$) посилює надійність висновків.

5.3.6 Підсумок ручного оцінювання

Ручна анотація показала домінування спотворень (79,8% помилок), переважно у власних (46,2%) та географічних (16,5%) назвах — категоріях, недостатньо представлених у навчальних даних NMT-моделі. Виявлено 7 критичних помилок, що суттєво змінюють зміст. Помилки розпізнавання власних назв безпосередньо призводять до некоректного перекладу, оскільки OPUS-MT не має механізму корекції вхідних помилок. Когерентність OPUS-MT — помірна ($C_{\text{auto}} = 3{,}34$, ручна оцінка 3,2 з 5).

5.4 Аналіз каскадного ефекту

Для дослідження впливу помилок ASR на якість MT проведено кореляційний аналіз між метриками. Використано два методи: коефіцієнт кореляції Пірсона і коефіцієнт Спірмена. Результати наведено в таблиці 14.

Табл. 14: Кореляція між метриками ASR та MT

	Пірсона (r)			**Спірмена (ρ)**		
	BLEU	BERTScore	METEOR	BLEU	BERTScore	METEOR
WER	$-0,38$	$-0,52$	$-0,27$	$-0,11$	$-0,33$	$-0,02$
CER	$-0,59^*$	$-0,65^*$	$-0,51$	$-0,21$	$-0,54$	$-0,06$

* — статистично значущий зв'язок ($p < 0,05$)

Виявлено статистично значущі від'ємні кореляції CER vs BLEU ($r = -0{,}59$, $p = 0{,}042$) та CER vs BERTScore ($r = -0{,}65$, $p = 0{,}022$). Це підтверджує каскадний ефект: помилки на рівні символів (CER), переважно у власних назвах і термінології, мають сильніший вплив на якість перекладу, ніж помилки на рівні слів (WER). Відсутність значущості для кореляцій Спірмена пояснюється малим розміром вибірки ($n = 12$). Кореляційні залежності візуалізовано на рис. 7.

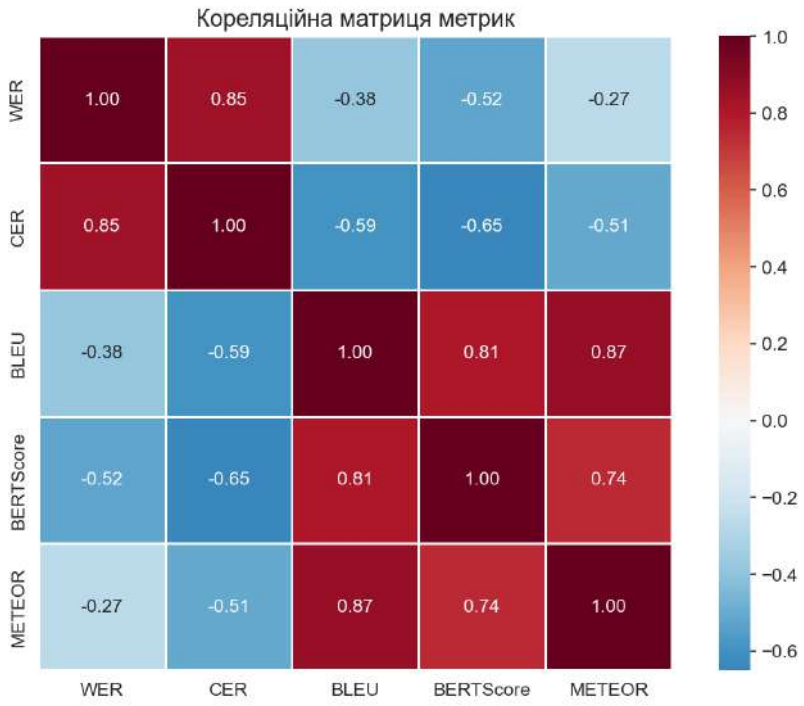

Рис. 7: Кореляційна матриця (коефіцієнт Пірсона) між метриками ASR та MT

5.5 Інтервенційне зменшення каскадного ефекту: експериментальне оцінювання запропонованого pipeline

Кореляційний аналіз у попередньому підрозділі встановив асоціативний зв'язок між помилками ASR та якістю MT ($r = -0{,}59$, $p = 0{,}042$ для CER–BLEU). Однак асоціація не є каузальністю: із кореляції не випливає, що виправлення ASR *призведе* до покращення MT. Для причинно-наслідкової перевірки каскадного ефекту проведено інтервенційний експеримент — порівняння baseline OPUS-MT із набором методів пост-ASR корекції та запропонованим pipeline (підрозділ 4.4).

5.5.1 Дизайн експерименту

На корпусі з $n = 12$ сегментів порівнюються шість варіантів pipeline на базі OPUS-MT:

1. **Baseline:** OPUS-MT на вихідній ASR-транскрипції.
2. **Exact-match correction:** заміна заздалегідь визначених пар «помилка–канонічна форма» у ASR, далі OPUS-MT.
3. **Phonetic correction:** виправлення ASR через фонетичну дистанцію (транслітерація + Левенштейн) до канонічних форм словника, далі OPUS-MT.
4. **LLM correction:** prompt-based корекція ASR великою мовною моделлю з наданим словником, далі OPUS-MT.
5. **Code-switching (52-dict):** запропонований pipeline із базовим ручним словником.
6. **Code-switching (176-dict):** запропонований pipeline із розширеним словником, побудованим ручним аналізом референсних перекладів корпусу.

Як емпірична верхня межа якості паралельно наведено чотири LLM у прямому uk→en режимі — Claude Haiku 4.5, Claude Sonnet 4, Gemini 2.0 Flash та GPT-4o (детальний аналіз LLM-систем — підрозділ 5.6) — для визначення мінімального LLM-класу, з яким конкурує запропонований локальний метод.

Для забезпечення коректності порівняння всі варіанти MT обробляються єдиною процедурою (OPUS-MT із чанкуванням вхідного тексту по 400 символів), що дає дещо нижчий BLEU baseline (19,01) порівняно з посегментним перекладом розділу 5 (23,12); такий вибір усуває методологічну варіативність між процедурами та забезпечує справедливість парного порівняння.

Оцінка виконана за метриками BLEU, BERTScore-F1, METEOR та Entity F1 (точне зіставлення нормалізованих іменованих сутностей PERSON/GPE /LOC/ORG, виокремлення через `en_core_web_sm`). Статистична значущість відмінностей BLEU від baseline перевіряється критерієм Вілкоксона для пов'язаних вибірок.

5.5.2 Результати

Зведені результати наведено у таблиці 15.

Табл. 15: Порівняння MT-варіантів: автоматичні метрики і Entity F1 ($n = 12$)

Варіант	**BLEU**	**BERT-F1**	**METEOR**	**Entity F1**	p (Wilcoxon)
OPUS-MT (baseline)	19,01	0,330	0,416	0,447	—
OPUS-MT (exact-match)	19,13	0,344	0,421	0,455	0,234
OPUS-MT (phon. correction)	18,93	0,330	0,415	0,450	0,719
OPUS-MT (LLM correction)	19,51	0,340	0,429	0,451	0,151
OPUS-MT (code-switch, 52)	19,62	0,346	0,418	0,515	0,080
OPUS-MT (code-sw, 176)	**20,33**	**0,350**	0,423	**0,610**	**0,046***
Claude Haiku 4.5 (direct)	55,25	0,724	0,722	0,837	0,0002**
Claude Sonnet 4 (direct)	55,46	0,710	0,730	0,849	0,0002**
Gemini 2.0 Flash (direct)	56,92	0,739	0,747	0,845	0,0002**
GPT-4o (direct)	57,00	0,725	0,746	0,860	0,0002**

* $p < 0{,}05$; ** $p < 0{,}001$ (порівняння BLEU з baseline).

Візуалізацію результатів за усіма метриками наведено на рис. **??**. Посегментний розподіл Entity F1 — на рис. 8. Динаміка збереження канонічних форм у перекладі до та після post-MT repair — на рис. 9.

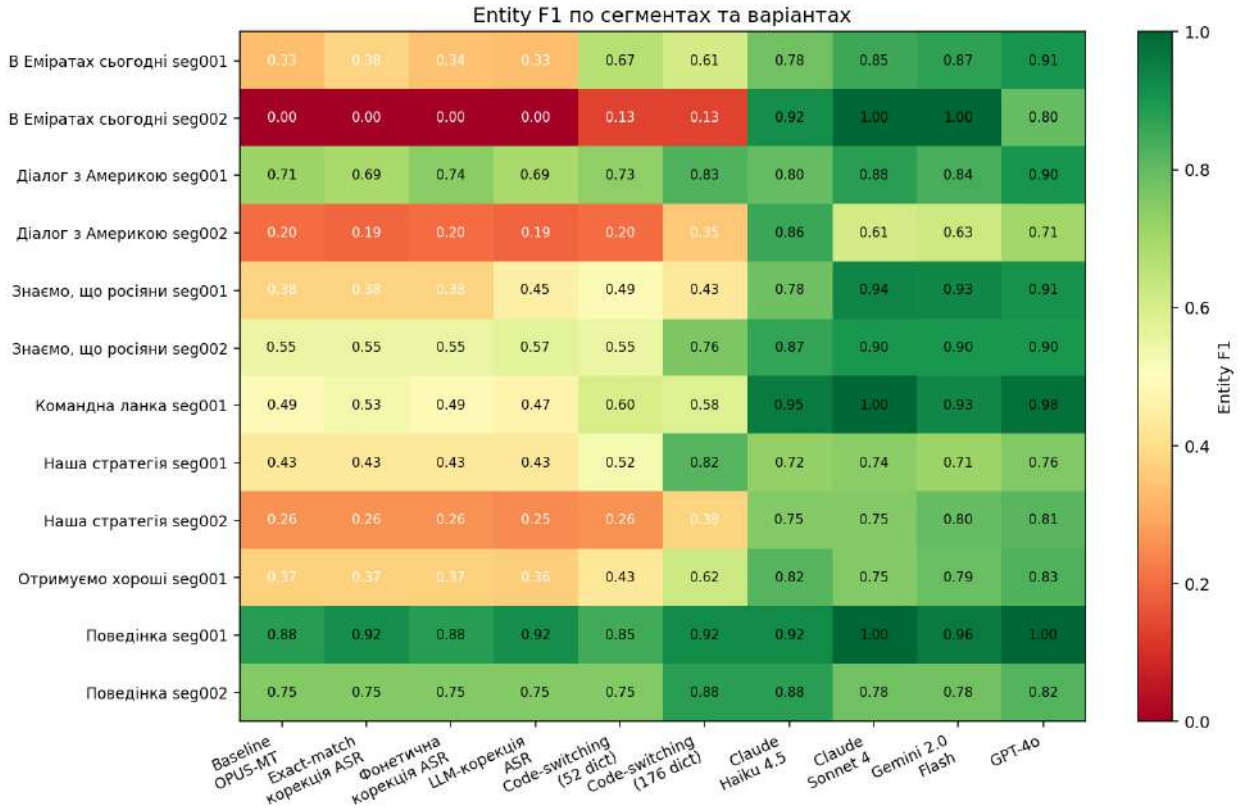

Рис. 8: Посегментний розподіл Entity F1 за варіантами MT

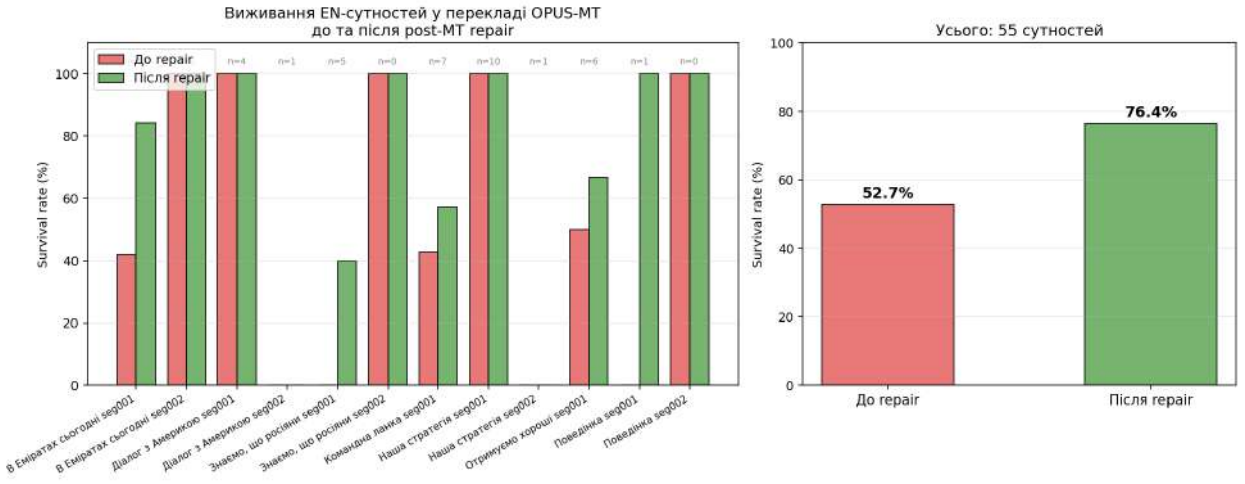

Рис. 9: Збереження канонічних англійських форм іменованих сутностей у перекладі OPUS-MT до та після post-MT repair. Ліворуч — посегментна статистика; праворуч — загальний survival rate на корпусі

5.5.3 Ефективність post-MT repair

Без модуля post-MT repair OPUS-MT зберігає лише 62,5% вставлених англійських форм (із розширеним словником; для малого — 52,7%). Post-MT repair виявляє 29 спотворень і успішно виправляє всі, піднімаючи survival rate до 75,0% (таблиця 16). Характерні приклади: *Sumland* → *Sumy region*, *Chernigivs region* → *Chernihiv region*, *Rust Umerov* → *Rustem Umerov*.

Табл. 16: Survival rate іменованих сутностей у перекладі OPUS-MT

Варіант	**Виявлено**	**pre-repair**	**Repairs**	**post-repair**
Code-switching (52-dict)	55	52,7%	14	76,4%
Code-switching (176-dict)	160	62,5%	29	75,0%

5.5.4 Аналіз складових Entity F1

Розширений словник забезпечує приріст як precision, так і recall (таблиця 17): у середньому правильно передається 10,25 сутностей на сегмент проти 7,08 у baseline (з 17,75 у референсі), що закриває 39,6% розриву до GPT-4o (+8,00 сутностей). Стандартне відхилення BLEU між сегментами знижується з 4,34 до 2,30 — метод забезпечує більш передбачувану якість між сегментами.

Табл. 17: Розкладення Entity F1 на precision, recall та кількість правильно переданих сутностей ($n = 12$)

Варіант	**Precision**	**Recall**	**Entity F1**	**avg правильних (з 17,75/сегм)**
OPUS-MT (baseline)	0,486	0,420	0,447	7,08
Code-switching (52)	0,586	0,475	0,515	8,33
Code-switching (176)	**0,663**	**0,582**	**0,610**	**10,25**
GPT-4o (direct MT)	0,871	0,851	0,860	15,08

5.5.5 Висновки до підрозділу

Методи, що обмежуються виправленням ASR-транскрипції (exact-match, phonetic, LLM-correction), дають статистично незначущий приріст BLEU ($p > 0{,}15$) та $< 2\%$ приросту Entity F1. Це суперечить інтуїтивному очікуванню із кореляційного аналізу ($r = -0{,}59$): самі виправлення ASR не відновлюють втрачену інформацію про іменовані сутності, оскільки OPUS-MT спотворює їх навіть при коректному вхідному тексті.

Запропонований code-switching pipeline не «виправляє» OPUS-MT, а *обходить* його слабкість: у поєднанні з розширеним словником (176 записів) він дає єдиний серед локальних методів статистично значущий приріст BLEU (+6,9%, $p = 0{,}046$), найбільший приріст Entity F1 (+36,5%) та закриває 40% розриву до GPT-4o. На посегментному рівні метод перевершує Claude Haiku 4.5 на 17% сегментів та рівний йому ще на 17% — усього на 33% корпусу локальний метод конкурує з найменшою з протестованих LLM.

У середньому метод усе ж поступається усім LLM за Entity F1 (0,610 проти 0,837–0,860), що відображає перевагу моделей зі значно ширшими контекстуальними знаннями про світ та окреслює три обмеження та напрямки подальших досліджень:

- *Залежність від словника*: сутності поза D не виявляються; автоматична екстракція з паралельних корпусів (cross-lingual NER alignment) — природне продовження.
- *Survival rate* 75%: чверть вставлених форм видозмінюється понад поріг fuzzy-репару; потенційне підсилення — forced-decoding constraints у MarianMT або fine-tuning моделі на code-switched прикладах.
- *Обсяг експерименту* ($n = 12$) обмежує статистичну потужність; масштабування на більший корпус потрібне для верифікації висновків.

5.6 Порівняння з LLM-системами перекладу

Після аналізу результатів основного pipeline (Whisper + OPUS-MT) та каскадного ефекту постає питання: чи можна суттєво покращити якість MT-

ланки, замінивши спеціалізовану NMT-модель на сучасну велику мовну модель? Для відповіді на це питання проведено додатковий порівняльний експеримент із трьома провідними LLM-системами.

5.6.1 Опис LLM-систем

Для порівняння обрано три провідні LLM: GPT-4o (OpenAI) [13], Claude Sonnet 4 (Anthropic) та Gemini 2.0 Flash (Google). Усі три — авторегресивні Transformer-моделі загального призначення; за існуючими оцінками [27, 23], LLM-переклад для пари uk–en часто перевищує спеціалізовані NMT-системи завдяки кращому розумінню контексту та disambiguation. Зведену характеристику наведено у таблиці 18.

Табл. 18: Порівняльна характеристика MT-систем

Характеристика	OPUS-MT	GPT-4o	Claude S. 4	Gemini 2.0
Тип	Спеціаліз. NMT	LLM	LLM	LLM
Архітектура	Enc-Dec	Dec-only	Dec-only	Dec-only
Ліцензія	MIT	API	API	API
Локальна робота	Так	Ні	Ні	Ні
Контекст (токени)	512	128K	200K	1M

5.6.2 Умови експерименту

Для LLM-систем використано параметр `temperature=0` для забезпечення максимальної детермінованості. Усі LLM отримували однакове завдання і умови. Вхідними даними слугували ті ж самі ASR-транскрипції, що використовувались для OPUS-MT, що забезпечує коректність порівняння.

5.6.3 Результати автоматичних метрик

Зведені результати оцінювання якості перекладу наведено в таблиці 19.

Табл. 19: Порівняння якості перекладу: OPUS-MT vs LLM

Модель	**BLEU**	**BERTScore-F1**	**METEOR**
OPUS-MT	23,12	0,397	0,452
GPT-4o	**57,00**	0,725	0,746
Claude Sonnet 4	55,46	0,710	0,730
Gemini 2.0 Flash	56,92	**0,739**	**0,747**

Результати демонструють суттєву перевагу LLM-систем над спеціалізованою MT-моделлю OPUS-MT за всіма метриками:

- **BLEU**: покращення на 140–147% (з 23,12 до 55–57);
- **BERTScore-F1**: покращення на 79–86% (з 0,397 до 0,71–0,74);
- **METEOR**: покращення на 62–65% (з 0,452 до 0,73–0,75).

Статистичну значущість різниці підтверджено кількома непараметричними тестами. Критерій Манна-Уітні [14] показав $U = 0$ та $p < 0,001$ для всіх порівнянь OPUS-MT з кожною LLM за кожною метрикою, що свідчить про повну відсутність перетину розподілів (розмір ефекту $r = 1,0$). Тест Вілкоксона для парних порівнянь ($W = 0$, $p < 0,001$) підтвердив ці результати. Тест Крускала-Волліса для порівняння трьох LLM між собою не виявив статистично значущих відмінностей ($H = 1,08$, $p = 0,58$ для BLEU; $H = 2,37$, $p = 0,31$ для BERTScore; $H = 1,02$, $p = 0,60$ для METEOR), що підтверджує приблизну рівноцінність LLM-систем для даної задачі.

5.6.4 Аналіз покращень

Візуалізацію відносного покращення LLM над OPUS-MT наведено на рис. 10.

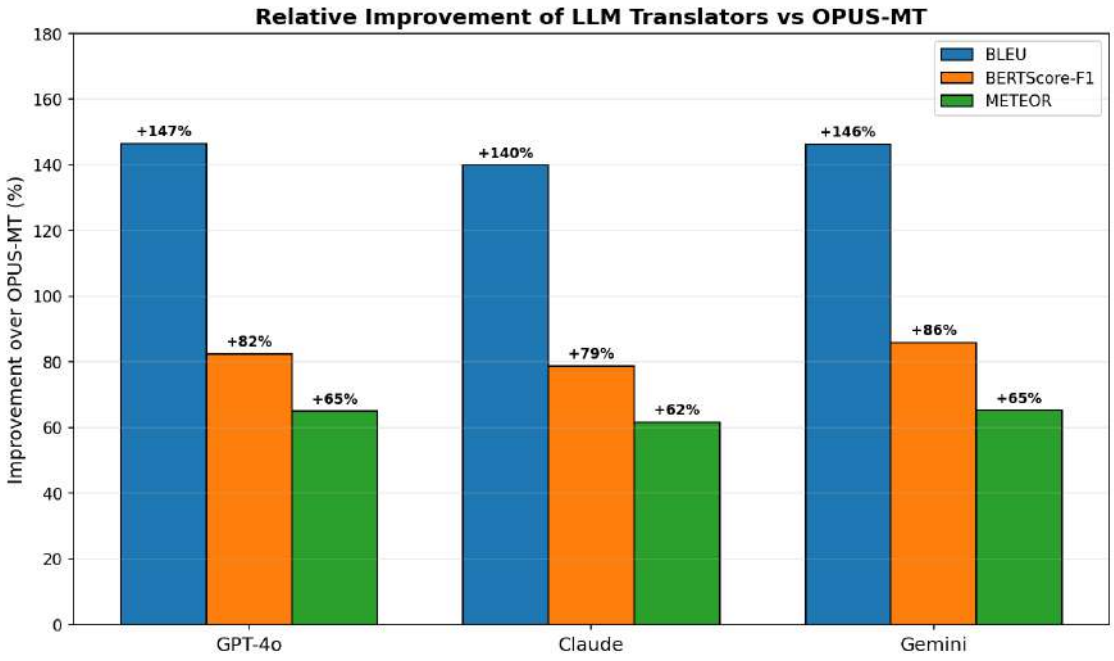

Рис. 10: Відносне покращення якості перекладу LLM порівняно з OPUS-MT (%)

Серед LLM спостерігається незначна варіація: GPT-4o найвищий за BLEU (57,00), Gemini 2.0 Flash — за BERTScore-F1 (0,739) та METEOR (0,747); Claude Sonnet 4 трохи нижчий за обома метриками. Різниця між LLM статистично незначуща ($p > 0{,}05$), тобто вони приблизно рівноцінні для цієї задачі.

5.6.5 Посегментний аналіз

Порівняння BLEU за окремими сегментами наведено на рис. 11.

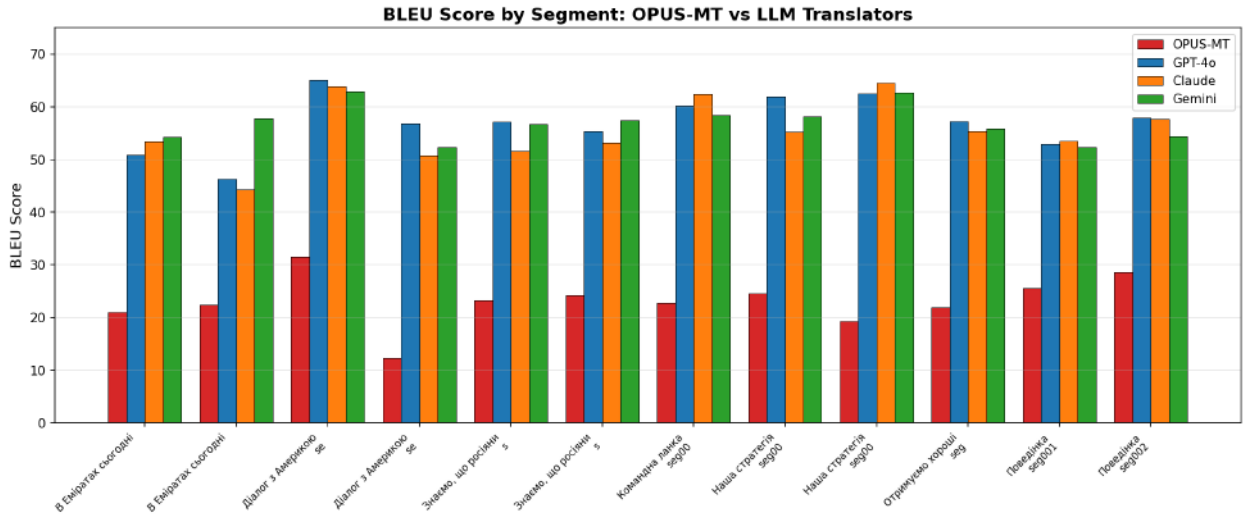

Рис. 11: BLEU за сегментами: OPUS-MT vs LLM-системи

Мовні моделі демонструють стабільно вищі результати на всіх сегментах корпусу. Особливо помітна перевага на сегментах із високою щільністю власних назв та термінології («Командна ланка», «Наша стратегія»), де OPUS-MT показував найнижчі результати.

5.6.6 Якісний аналіз перекладу власних назв

Ключовою перевагою LLM є коректна обробка власних назв. Приклади порівняння наведено в таблиці 20.

Табл. 20: Порівняння перекладу власних назв: OPUS-MT vs LLM

Оригінал (ASR)	**OPUS-MT**	**GPT-4o**	**Референс**
Рустему Міров	Rustema Mirov	Rustem Umerov	Rustem Umerov
Кривирі	Crete	Kryvyi Rih	Kryvyi Rih
генерал Гнатов	General Gnatov	General Hnatov	General Hnatov
ГУР	MBI	HUR (Defense Intelligence)	HUR

LLM демонструють здатність розпізнавати й коригувати помилки ASR у власних назвах, застосовувати стандартну транслітерацію українських топонімів та правильно розшифровувати абревіатури завдяки контекстуаль-

ним знанням, отриманим під час навчання на великих корпусах з українським та міжнародним новинним контентом.

5.6.7 Обговорення результатів

Перевага LLM над OPUS-MT [10] пояснюється масштабом навчальних даних, наявністю «знань про світ» (українські політики, географія, військова термінологія), можливістю інструкційного налаштування та довшим контекстом, що покращує когерентність. Обмеження — вища латентність (4–15 с на сегмент vs $<$ 1 с для OPUS-MT), залежність від API та інтернет-з'єднання, вища вартість і потенційні проблеми з конфіденційністю.

5.6.8 Висновки щодо LLM-перекладу

Експеримент підтвердив значну перевагу LLM-систем над спеціалізованими MT-моделями для перекладу політичного дискурсу з української на англійську:

- BLEU підвищується з 23 до 55–57 (+147%);
- BERTScore-F1 зростає з 0,397 до 0,71–0,74 (+82%);
- Критично покращується обробка власних назв та термінології;
- Композитна оцінка когерентності C_{auto} зростає з 3,34 (OPUS-MT) до 3,90–4,19 (LLM), $p < 0{,}01$ за Манна-Уітні для всіх трьох пар (підрозділ 5.3.5). Локалізовано джерело переваги: вища флюентність (F_{flu} зростає з 0,73 до 0,95–0,99) та точність іменованих сутностей (F_{ent} зростає з 0,48 до 0,84–0,86), при цьому тематична зв'язність та збереження дискурсивних маркерів практично однакові для всіх систем.

Для практичного застосування рекомендується використання LLM як MT-компонента. Для real-time застосувань OPUS-MT залишається прийнятною альтернативою з можливим пост-редагуванням.

5.7 Висновки до розділу 5

Whisper medium демонструє WER = 8,62% (кращий за типові 10–15% для української). OPUS-MT забезпечує BLEU = 23,12, що є типовим для пари uk–en. Виявлено статистично значущий каскадний ефект ($r = -0{,}59$ для CER–BLEU, $p < 0{,}05$), і ручна анотація ідентифікувала 114 помилок OPUS-MT, з яких 79,8% — спотворення (переважно власних назв). Розроблена автоматична метрика когерентності C_{auto} підтверджує статистично значущу перевагу LLM над OPUS-MT (3,90–4,19 vs 3,34; $p < 0{,}001$), з локалізацією джерела переваги у флюентності та точності іменованих сутностей. Запропонований entity-aware code-switching pipeline з розширеним словником (176 записів) дав BLEU = 20,33 (+6,9% над baseline, Wilcoxon $p = 0{,}046$) та Entity F1 = 0,610 (+36%), закривши 40% розриву до GPT-4o за точністю передачі власних назв. Додаткове дослідження порівняння із LLM-системами показало, що GPT-4o, Claude Sonnet 4, Gemini 2.0 Flash та Claude Haiku 4.5 забезпечують BLEU = 55–57 ($p < 0{,}001$) — суттєву перевагу за рахунок контекстуальних знань про світ, яку не здатні відтворити локальні методи пост-корекції ASR.

6 Висновки

У роботі проведено комплексне дослідження якості україно-англійського автоматичного перекладу усного мовлення з акцентом на семантичну точність та міжфразову когерентність. Дослідження виконано на корпусі політичного дискурсу із використанням каскадної архітектури ASR (Whisper medium) + MT (OPUS-MT). Додатково проведено порівняльний аналіз із чотирма великими мовними моделями: GPT-4o, Claude Sonnet 4, Gemini 2.0 Flash та Claude Haiku 4.5.

Основні результати:

1. **Якість ASR та MT.** Whisper medium досягає WER = 8,62%, CER = 2,10% (краще за типові 10–15% для української). OPUS-MT забезпечує BLEU = 23,12, BERTScore-F1 = 0,397, METEOR = 0,452 — типові значення для пари uk–en та відповідні бенчмарку FLORES.

2. **Перевага LLM.** GPT-4o, Claude Sonnet 4, Gemini 2.0 Flash та Claude Haiku 4.5 суттєво перевищують OPUS-MT: BLEU = 55–57 (+139–147%), BERTScore-F1 = 0,71–0,74, METEOR = 0,72–0,75; $p < 0{,}001$. Перевага LLM зумовлена ширшими контекстуальними знаннями про світ, що дозволяє правильно передавати власні назви без доменного словника.

3. **Каскадний ефект.** Виявлено статистично значущу від'ємну кореляцію між помилками ASR та якістю перекладу: $r = -0{,}59$ для CER–BLEU ($p = 0{,}042$) та $r = -0{,}65$ для CER–BERTScore ($p = 0{,}022$).

4. **Типові помилки.** Ручна анотація виявила 114 помилок OPUS-MT (79,8% — спотворення); найпроблемніші категорії: власні назви (46,2%), географічні назви (16,5%), військова термінологія (13,2%). Виявлено 7 критичних помилок зі зміною змісту.

5. **Когерентність.** Розроблено композитну метрику C_{auto} (флюентність, F_1 іменованих сутностей, дискурсивні маркери, семантична зв'язність). LLM значущо переважають OPUS-MT (3,90–4,19 vs 3,34, Краскел-Уолліс

$H = 26{,}69$, $p < 0{,}001$); перевага локалізована у флюентності та точності іменованих сутностей.

6. **Запропонований метод entity-aware code-switching.** Розроблено pipeline-level рішення, що виявляє іменовані сутності у UA-тексті ASR, підставляє канонічні англійські форми у вхідний текст перед OPUS-MT та виконує post-MT repair через fuzzy-зіставлення. З розширеним доменним словником (176 записів) метод дає +6,9% BLEU над baseline (Вілкоксон $p = 0{,}046$ — єдиний серед локальних методів зі статистично значущим приростом) та +36,5% Entity F1, закриваючи 40% розриву до GPT-4o за точністю передачі власних назв. Інтервенційно підтверджено, що основне обмеження якості каскадного pipeline — сама MT-модель: методи, обмежені корекцією ASR (exact-match, фонетична, LLM-correction), не дають статистично значущого приросту BLEU.

Практичне значення. Методологія застосовна для порівняльного аналізу систем перекладу у сфері дипломатичної комунікації. Результати порівняння OPUS-MT з LLM надають практичні рекомендації щодо вибору MT-компонента: для real-time — OPUS-MT з можливим пост-редагуванням, для офлайн-обробки з вищими вимогами до якості — LLM-системи.

Обмеження. Розмір корпусу (12 сегментів, ≈45 хв) обмежує статистичну потужність кореляційного аналізу. Ручна анотація виконана одним анотатором без обчислення IAA; для когерентності цей недолік компенсовано автоматичною метрикою C_{auto}, валідованою кореляцією з ручною оцінкою (Спірмен $\rho = 0{,}79$).

Напрямки подальших досліджень: автоматичне формування доменного словника через cross-lingual NER-alignment паралельних корпусів; forced-decoding constraints у MarianNMT або fine-tuning на code-switched даних для підвищення survival rate; гібридний pipeline (LLM-MT + post-MT repair) для підсилення найкращих LLM; порівняння запропонованого методу з моделями попередніх поколінь (NLLB-200, M2M-100) та комерційними сервісами (Google Translate, DeepL); розширення корпусу для підвищення статистичної потужності.

Література

[1] Jurafsky D., Martin J. H. *Speech and Language Processing*. 3rd ed. — Draft. Stanford University, 2023.

[2] Vaswani A., Shazeer N., Parmar N. et al. Attention Is All You Need // *Advances in Neural Information Processing Systems (NeurIPS)*. — 2017.

[3] Hinton G., Deng L., Yu D. et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition // *IEEE Signal Processing Magazine*. — 2012.

[4] Goodfellow I., Bengio Y., Courville A. *Deep Learning*. MIT Press, 2016.

[5] Papineni K., Roukos S., Ward T., Zhu W.-J. BLEU: a Method for Automatic Evaluation of Machine Translation // *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. — 2002.

[6] Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT // *Proceedings of the International Conference on Learning Representations (ICLR)*. — 2020.

[7] Banerjee S., Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments // *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. — 2005.

[8] Radford A., Kim J. W., Xu T. et al. Robust Speech Recognition via Large-Scale Weak Supervision // *Proceedings of the International Conference on Machine Learning (ICML)*. — 2023.

[9] Post M. A Call for Clarity in Reporting BLEU Scores // *Proceedings of the Third Conference on Machine Translation (WMT)*. — 2018.

[10] Tiedemann J., Thottingal S. OPUS-MT — Building open translation services for the World // *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT)*. — 2020.

[11] NLLB Team, Costa-jussà M. R. et al. No Language Left Behind: Scaling Human-Centered Machine Translation // *arXiv preprint arXiv:2207.04672*. — 2022.

[12] Barrault L., Chung Y.-A., Meglioli M. C. et al. SeamlessM4T — Massively Multilingual & Multimodal Machine Translation // *arXiv preprint arXiv:2308.11596*. — 2023.

[13] OpenAI. GPT-4 Technical Report // *arXiv preprint arXiv:2303.08774*. — 2023.

[14] Mann H. B., Whitney D. R. On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other // *The Annals of Mathematical Statistics*. — 1947. — Vol. 18, No. 1. — P. 50–60.

[15] Spearman C. The Proof and Measurement of Association between Two Things // *The American Journal of Psychology*. — 1904. — Vol. 15, No. 1. — P. 72–101.

[16] Junczys-Dowmunt M., Grundkiewicz R., Dwojak T. et al. Marian: Fast Neural Machine Translation in C++ // *Proceedings of ACL 2018, System Demonstrations*. — 2018.

[17] Klein G., Hernandez F., Nguyen V. CTranslate2: Efficient Inference with Transformer Models // *GitHub repository*. — 2020.

[18] Goyal N., Gao C., Chaudhary V. et al. The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation // *Transactions of the Association for Computational Linguistics*. — 2022.

[19] Sperber M., Paulik M. Speech Translation and the End-to-End Promise: Taking Stock of Where We Are // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. — 2020.

[20] Iranzo-Sánchez J., Silvestre-Cerdà J. A., Jorge J. et al. Direct Speech Translation for Automatic Subtitling // *Proceedings of Interspeech*. — 2021.

[21] Bentivogli L., Cettolo M., Federico M., Cattoni R. Machine Translation Quality: Human vs. Cascade Effects in End-to-End Speech Translation // *Proceedings of the International Workshop on Spoken Language Translation (IWSLT).* — 2021.

[22] Lommel A., Uszkoreit H., Burchardt A. Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics // *Tradumàtica: tecnologies de la traducció.* — 2014. — No. 12. — P. 455–459.

[23] Zhu W., Liu H., Dong Q. et al. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis // *Findings of NAACL.* — 2024.

[24] Baevski A., Zhou Y., Mohamed A., Auli M. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations // *Advances in Neural Information Processing Systems (NeurIPS).* — 2020.

[25] Tiedemann J. Parallel Data, Tools and Interfaces in OPUS // *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC).* — 2012.

[26] Zoph B., Yuret D., May J., Knight K. Transfer Learning for Low-Resource Neural Machine Translation // *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).* — 2016.

[27] Hendy A., Abdelrehim M., Sharaf A. et al. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation // *arXiv preprint arXiv:2302.09210.* — 2023.

[28] Ruiz N., Federico M. Assessing the Impact of Speech Recognition Errors on Machine Translation Quality // *Proceedings of the 11th Conference of the Association for Machine Translation in the Americas (AMTA).* — 2014.

[29] Anastasopoulos A., Chiang D. A Survey on End-to-End Speech Translation // *arXiv preprint arXiv:2104.06494.* — 2022.

[30] Мур І. Ukrainian Speech Translation Research: вихідний код та дані експерименту // *GitHub repository*. — 2025. — https://github.com/IllyaMoore/ukrainian-speech-translation-research