Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
Навчально-науковий інститут філології
Кафедра української мови та прикладної лінгвістики

ДОНАВЧАННЯ МОДЕЛІ МАШИННОГО ПЕРЕКЛАДУ ДЛЯ ЛОКАЛІЗАЦІЇ АНГЛОМОВНИХ КОМІКСІВ УКРАЇНСЬКОЮ МОВОЮ

Кваліфікаційна робота
на здобуття ОС «бакалавр»
студента IV курсу
галузі знань 03 «Гуманітарні науки»,
спеціальність 035 «Філологія»
спеціалізації 035.10 «Прикладна лінгвістика», ОПП «Прикладна (комп'ютерна) лінгвістика та англійська мова»
Іллі Костянтиновича ШЕВЧЕНКА

Науковий керівник:
асистент кафедри української мови
та прикладної лінгвістики
Валентина РОБЕЙКО

«Допущено до захисту»
Протокол засідання кафедри
української мови та прикладної лінгвістики
від «5» 06 2026 року № 16

завідувач кафедри _____
к.філол.н., доц. Сергій РІЗНИК

КИЇВ – 2026

ЗМІСТ

АНОТАЦІЯ

Шевченко І. К. Донавчання моделі машинного перекладу для локалізації англомовних коміксів українською мовою.

Кваліфікаційна робота на здобуття ОС «бакалавр». – Київський національний університет імені Тараса Шевченка – Київ, 2026.

Робота присвячена вузькоспеціалізованій адаптації моделі машинного перекладу для відтворення текстів англомовних коміксів українською мовою. Актуальність зумовлена зростанням попиту на україномовний локалізований контент, низькоресурсністю мовної пари англійська-українська та відсутністю спеціалізованих систем перекладу для жанру коміксів.

Об’єктом дослідження є корпус текстів англомовних коміксів та їх відповідники українською мовою. Предметом – процес і результати донавчання нейромережевої моделі *«Dragoman»* на вузькоспеціалізованих текстових даних. Мета роботи полягає у визначенні ефективності доменної адаптації порівняно з базовою та альтернативною системами. Для цього проаналізовано жанр комікс, розглянуто еволюцію підходів до машинного перекладу, обґрунтовано вибір моделі, сформовано спеціалізований корпус, проведено донавчання, автоматичне оцінювання та лінгвістичний аналіз помилок, також застосовано модель в умовах реального пайплайну.

Дослідження базується на корпусних, експериментальних та порівняльних методах та лінгвістичному аналізі. Паралельний корпус на 2000 сегментів підготовлено стратифікованою вибіркою з «*COMICS TEXT+»*. Донавчання виконано методом LoRA-адаптації з 4-бітним квантуванням у *Google Colaboratory*. За метриками *BLEU*, *TER*, *chrF* та *COMET* донавчена модель переважає над базовими системами. Покращилося відтворення ідіом, розмовних реплік і кличного відмінка. Розроблено прикладний пайплайн із графічним інтерфейсом, що інтегрує систему розпізнавання тексту «*ComiQ*» та переклад. Наукова новизна полягає у першому досвіді донавчання української decoder-only моделі на корпусі коміксів.

Ключові слова: машинний переклад, донавчання, доменна адаптація, комікси, паралельний корпус, мовна пара англійська-українська.

ABSTRACT

Shevchenko I. K. Fine-Tuning a Machine Translation Model for the Localization of English Comics into Ukrainian.

Bachelor's qualification thesis. – Taras Shevchenko National University of Kyiv – Kyiv, 2026.

The thesis examines domain-specific fine-tuning of a machine translation model for rendering English-language comic book texts into Ukrainian. The study is motivated by growing demand for localised Ukrainian content, the low-resource nature of the English-Ukrainian language pair, and the absence of specialised translation systems for the comics genre.

The object of the study is a corpus of English-language comic book texts and their Ukrainian counterparts. The subject is the process and outcomes of fine-tuning the *Dragoman* neural machine translation model on domain-specific data. The aim is to assess the effectiveness of domain adaptation compared to baseline and alternative systems. To this end, the comics genre was analysed, the evolution of machine translation approaches was reviewed, the model selection was justified, a specialised corpus was compiled, fine-tuning and automatic evaluation were performed, a linguistic error analysis was conducted, and the model was tested in a real-world pipeline.

The study draws on corpus-based, experimental, and comparative methods alongside linguistic analysis. A 2,000-segment parallel corpus was prepared via stratified sampling from *COMICS TEXT+*. Fine-tuning was performed using LoRA adaptation with 4-bit quantisation in *Google Colaboratory*. Across *BLEU*, *TER*, *chrF*, and *COMET* metrics, the fine-tuned model outperformed the baseline systems. Improvements were observed in rendering idioms, colloquial utterances, and vocative case forms. An applied pipeline with a graphical interface integrating the *ComiQ* OCR system and translation was developed. The scientific novelty lies in the first fine-tuning of a Ukrainian decoder-only model on a comics corpus.

Keywords: machine translation, fine-tuning, domain adaptation, comics, parallel corpus, English-Ukrainian language pair.

ВСТУП

Стрімкий розвиток технологій машинного перекладу в сучасному світі викликав потребу у вирішенні нових, складніших завдань, спрямованих на покращення якості перекладу. Одним із таких викликів є автоматичний переклад мультимодальних текстів, зокрема тих, що поєднують вербальні та візуальні компоненти. До цієї категорії належать комікси – форма наративу, яка завдяки певній послідовності зображень у поєднанні з текстом утворює цілісну оповідь.

Цей жанр популяризувався й в українському культурному просторі, тому зростає необхідність до забезпечення адекватного перекладу таких текстів, зокрема й засобами машинного перекладу.

Актуальність цієї роботи зумовлена кількома чинниками. По-перше, спостерігається стійке зростання попиту на україномовний контент, зокрема локалізовані комікси та мангу. За даними презентації «Книжковий ринок України», описаними в статті О. Бойко (2024), ця тенденція особливо посилилася після початку повномасштабного вторгнення 2022 року, а 70% іноземної літератури в Україні перекладається саме з англійської мови. По-друге, мовна пара англійська-українська історично належить до низькоресурсних, відповідно для неї доступно значно менше якісних паралельних корпусів, що безпосередньо впливає на якості загальних систем машинного перекладу. По-третє, універсальні системи машинного перекладу не враховують жанрову специфіку коміксів, а автоматичний переклад мультимодальних текстів залишається відкритою науковою задачею.

Практичне значення одержаних результатів полягає в тому, що запропонована методологія донавчання та розроблений пайплайн можуть бути використані локалізаторами для автоматизації перекладу, що скорочує витрати часу й ресурсів на початкових етапах процесу. Для дослідників у галузі NLP відкрита методологія адаптації моделі та корпус створюють основу для подальших експериментів з іншими архітектурами. Інтеграція пайплайну з системою

розпізнавання тексту додатково підтверджує практичну застосовність рішення в реальних умовах локалізації.

Мета роботи – дослідити можливості донавчання нейронної моделі машинного перекладу «*Dragoman*» для перекладу англомовних коміксів українською мовою та визначити, наскільки вузькоспеціалізована адаптація покращує якість перекладу порівняно з базовою й альтернативною моделями.

Для досягнення поставленої мети були визначені такі **завдання**:

1. Проаналізувати структуру та особливості подання тексту в коміксах як мультимодальному жанрі.
2. Розглянути сучасні підходи до машинного перекладу та методи адаптації вузькоспеціалізованих текстів.
3. Обґрунтувати вибір моделі «*Dragoman*» на основі аналізу її архітектурних особливостей та порівняння з альтернативами.
4. Підготувати спеціалізований паралельний англо-український корпус текстів коміксів для донавчання моделі.
5. Здійснити процес донавчання моделі «*Dragoman*» на підготовленому корпусі.
6. Провести комплексну автоматичну оцінку якості перекладу за метриками, порівнявши результати базової моделі «*Dragoman*», «*OPUS-MT*» та донавченої моделі.
7. Проаналізувати типові помилки, характерні для перекладів коміксів різними моделями.
8. Оцінити застосовність натренованої моделі в умовах реального пайплайну з моделлю розпізнавання тексту.

Об’єктом дослідження у цій роботі виступає корпус текстів англомовних коміксів, взятих з відкритого набору даних «*COMICS TEXT+*», їх відповідники українською мовою.

Предметом дослідження є процес і результати донавчання нейромережевої моделі «*Dragoman*» на вузькоспеціалізованих текстах для покращення якості перекладу.

У роботі було використано низку **методів**, які дозволили здійснити дослідження процесу та результатів донавчання моделі машинного перекладу. Зокрема, аналіз та синтез застосовувалися під час опрацювання наукової літератури, що стосується специфіки коміксів та машинного перекладу. Корпусні та експериментальні методи були використані для підготовки спеціалізованого англо-українського корпусу та під час процесу донавчання моделі *«Dragoman*». Методи порівняльного аналізу було застосовано для кількісного зіставлення результатів базових архітектур та донавченої моделі з метою виявлення покращення якості перекладу. Також лінгвістичний аналіз допоміг класифікувати типові помилки моделей та дав змогу оцінити здатність системи адекватно відтворювати специфічну лексику коміксів і визначити практичну застосовність розробленого рішення.

Матеріалом дослідження слугують тексти англомовних коміксів із відкритого набору даних «*COMICS TEXT+*» , що містить діалогові репліки, підписи та звуконаслідувальні слова, витягнуті автоматично з графічних панелей. На основі цього масиву було сформовано спеціалізований паралельний англо-український корпус обсягом 2000 паралельних сегментів, де якісні фрагменти англійського тексту різної довжини відбиралися автоматично, а переклад виконувався моделлю «*Google Gemini Flash 3*» із застосуванням технік інженерії запитів для забезпечення жанрової відповідності.

Наукова новизна роботи полягає в тому, що вперше здійснено вузькоспеціалізоване донавчання україномовної нейромережевої моделі машинного перекладу «Dragoman» на корпусі текстів коміксів та проведено комплексну порівняльну оцінку якості перекладу до і після адаптації з використанням автоматичних метрик і лінгвістичного аналізу типових помилок.

Структура роботи. Дипломна робота складається зі вступу, трьох розділів, висновків, списку використаних джерел та додатків. Загальний обсяг роботи становить 57 сторінок.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ ДОСЛІДЖЕННЯ МАШИННОГО ПЕРЕКЛАДУ ТЕКСТІВ КОМІКСІВ

Перший розділ присвячено формуванню теоретичної основи дослідження. Розуміння специфіки об'єкта перекладу та інструментів його опрацювання є необхідною передумовою для обґрунтованого проведення експерименту. Розділ складається з трьох підрозділів: у першому розглядається комікс як мультимодальний жанр та об'єкт перекладознавчого аналізу, у другому – еволюція методів машинного перекладу та підходи до адаптації вузькоспеціалізованих текстів, у третьому – архітектура та особливості моделі «*Dragoman*», обраної як основа для донавчання.

1.1 Комікси як мультимодальний жанр і об'єкт перекладознавчого дослідження

«**Комікс**», або, як його ще називають в українському просторі, «*мальопис*», походить від однойменного англійського слова «*comic*» або «*comical*», що можна перекласти як «*комічний*», «*смішний*» тощо. Тож первинне призначення коміксу було розважити читача за допомогою смішних картинок та надписів. З часом цей жанр значно еволюціонував, збагатився тематично й жанрово. Паралельно з цим в окремих культурах сформувалися власні варіанти коміксу.

Якщо розглядати поняття «*комікс*» з лінгвістичного боку, то воно належить до креолізованого тексту – тобто того, що поєднує в собі вербальні та візуальні компоненти, утворюючи при цьому цілісний наратив. На думку дослідника Скотта Макклауда (2019) термін «*комікс*» є загальним і охоплює широке коло форм, від традиційних «*скетч-коміксів*» до «*комікс-книжок*» і «*графічних романів*». **Скетч-комікс** можна описати як короткі історії на декілька панелей, що публікуються в газетах чи журналах. **Комікс-книжка** є повноцінним виданням з послідовними малюнками та текстом на десятки чи сотні сторінок. **Графічний роман** являє собою збірку мальописів з цілісним, завершеним і часто складнішим сюжетом. Його також можна вважати «літературною» формою коміксу.

Комікс у сучасному його розумінні виник у 1897 році в США з друком перших скетчів під назвою «*Жовтий Хлопчак*» в місцевих газетах і з того часу заполонив американський медіапростір (Данкан, Сміт та Левіц, 2020). З плином історії різні форми коміксів, відмінні від оригіналу, почали з'являтися по всьому світу. Так, наприклад, існують франко-бельгійські «**bande dessinées**» (укр. «*намальовані смужки*»), які відомі високоякісними ілюстраціями, намальованими фарбами у широкому форматі та з палітуркою (Мильченко та Татарінова, 2020).

Паралельно жанр також розвивався в азійських країнах. Популярною є «**манґа**» – унікальний японський різновид, якому властивий традиційний чорно-білий формат та прочитання справа наліво. Китайські «**маньхва**» та корейські «**манхва**» є більш сучасними варіаціями коміксу. Вони зображені у веб-форматі й орієнтовані на вертикальне прочитання, тобто прокручування сторінки вниз. (Manga Media, 2023)

Серед наукових досліджень, присвячених коміксу як мультимодальному жанру, можна виділити низку фундаментальних праць, що розкривають його структуру та культурний вплив: С. Макклауд «*Зрозуміти комікси*» (2019); Р. Данкан, М. Сміт та П. Левіц «*Сила Коміксів*» (2020); Н. Кон «*The Visual Language of Comics*» (2013); Т. Грунстін «*The System of Comics*» (2007).

С. Макклауд (2019) у своїй книзі пояснює роботу коміксу через сам комікс. Він пропонує таке визначення цього поняття: «суміжні статичні зображення та малюнки, вибудувані в наративі, покликані передавати інформацію і/чи викликати естетичний відгук» (2019, с.9). Особливу увагу автор приділяє процесу «додумування», що передає час, рух та емоцій через прогалини між панелями, які читач заповнює власною уявою. Макклауд також класифікує типи переходів між панелями та демонструє, як комікс досягає естетичної виразності завдяки взаємодії слова й образу. Загалом, ця робота стала революційною для теорії коміксів, оскільки вперше систематизувала їх структурні особливості.

У книзі **Р. Данкана**, **М. Сміта** та **П. Левіца** (2020) комікс розглядається як культурний феномен, аналізується історія, форма та культура навколо нього. Автори досліджують, як комікси трансформувалися від простих газетних скетчів до складних графічних романів, і який вплив вони мають на суспільство. Вони вводять поняття «інкапсуляції» і пояснюють, що це – «добір ключових моментів історії та розташування їхніх зображень у панелях» (2020, с.154). Також виділяють синтагматичний і парадигматичний процеси побудови сторінки, де синтагматика стосується послідовності панелей, які формують часову та просторову логіку оповіді, а парадигматика відображає вибір альтернативних способів візуалізації для передачі певних емоцій або ідей.

Н. Кон (2013) досліджує текст у коміксах з погляду когнітивної лінгвістики. Він вважає, що комікс має власну «візуальну мову» із власною граматикою, синтаксисом та морфологією, де текст є структурним компонентом цієї системи. Форма текстів, їхнє розміщення та навіть шрифти працюють як окремі слова у реченні та несуть певну інформацію.

Дослідник **Т. Грунстін** (2007) розглядає текст як невіддільну частину семіотичної системи коміксу, який постійно перебуває в динамічній взаємодії з візуальними компонентами. Так, наприклад, навіть короткі репліки персонажів можуть мати різне емоційне забарвлення залежно від візуального контексту, міміки персонажа та композиції кадру. Автор пропонує впровадити поняття «просторової топології» коміксу, тобто «погляду на комікс, який можна сформувати ще до ознайомлення з конкретним твором і з якого можливо уявити нове прочитання цього медіуму» (2007, с. 23).

У розглянутих працях зустрічається низка термінів на позначення окремих текстових елементів коміксу, які варто розглянути. А саме: «*мовна булька*», *текстові блоки*, *звуконаслідування* та *додаткові текстові елементи*.

Почнемо з основного засобу вербальної комунікації в коміксах – «**мовних бульок**». Термін також називають «*мовні кульки*», «*розмовні бульбашки*» тощо. Він

функціонує як вираження голосу або думки персонажу, якому ця булька належить. Структурно будову мовної бульки можна поділити на *зміст* (основний текст), *форма контуру* (впливає на сприйняття інтонації або характеру мовлення) і *хвостик* (вказує на мовця). Детальніше розглянемо саме **форми контуру**. Ніл Кон (2013) зазначає, що контур бульки функціонує як візуальна морфема, що змінює значення тексту. Наприклад, *зубчасті краї* можна вважати морфемою на позначення стану «гучності» або «агресії»; *хмароподібна форма* – морфема «думки» персонажа; *пунктирний контур* – морфема «шепоту» або «прихованості». З іншими формами можна ознайомитися нижче на *Рис. 1.1* – тут зображені основні та найуживаніші контури.

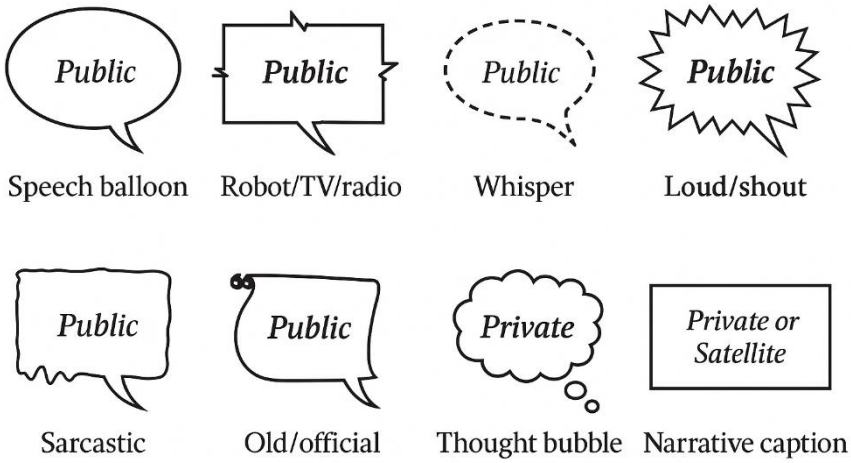

Рис. 1.1. Форми контурів. 1 – звичайне мовлення; 2 – звучання робота/телевізора/радіо; 3 – шепіт; 4 – гучно/крик; 5 – саркастично; 6 – старомодно/офіційно; 7 – думка; 8 – текстовий блок або монолог (Джерело: «The Visual Language of Comics»)

На відміну від мовних бульок, **текстові блоки** існують поза мовленням персонажа, але так само доповнюють історію. Зазвичай зображені у квадратних рамках, розміщених в одному з кутів панелі. Це слова самого автора або оповідача, який не є частиною історії, тобто знаходиться на іншому наративному рівні. Саме тому важливо не плутати їх із внутрішніми монологами персонажів, які зазвичай зображені ідентично, але належать до мовних бульок. (Forceville, Veale and Feyaerts, 2010)

Не менш важливим елементом коміксу є явище **звуконаслідування** або **ономатопеї**. Відповідно до назви, це слова, що імітують звуки, які не є частиною мовлення людини. Вони виконують функцію візуалізації та експресивності подій, що відбуваються в панелі. У коміксах звуконаслідування часто графічно стилізоване, щоб підкреслити дії (удари, вибухи, рухи тощо). Це допомагає створювати яскраві образи та покращує сприйняття сюжету. (Шеремет, 2021)

До останньої категорії віднесемо всі інші елементи, що не належать до описаних типів. Сюди входять: **вказівки часу й місця дії**; **написи всередині кадру** (вивіски, етикетки тощо); **авторські примітки**; **нумерація глав чи сторінок**. Такі елементи виконують роль художніх інструментів, які доповнюють загальну картину оповіді або спрощують її сприйняття для читача.

Окрім текстових елементів коміксу, варто зауважити, що сам текст також може бути насиченим мультимодальними компонентами. Вони сприяють реалізації базових функцій тексту в коміксах: інформативної, експресивної, апелятивної (звернення до читача), контактовстановної (підтримка взаємодії між персонажами), ідентифікаційної (розпізнавання мовця) та мислетвірної (Івасишин, 2019, с. 18).

Шрифтова гарнітура є найпоширенішим інструментом, що виконує не лише естетичну, а й прагматичну функцію. Наприклад, *жирний шрифт* виділяє ключові слова; *розмір, колір* і *варіації шрифту* можуть створювати особливий стиль мовлення персонажа або ідентифікувати його (Івасишин, 2019). **Регістр символів** також впливає на сприйняття інформації читачем. Здебільшого текст у коміксах подається у великих літерах, оскільки в такому вигляді всі символи вирівнянні й не перекривають елементи сусідніх рядків. Це було особливо важливо на ранніх етапах розвитку коміксу, коли технології друку не дозволяли якісно відтворювати складне рукописне оформлення (Pellitteri, 2024). **Графічні символи** виступають допоміжними засобами для реалізації експресивної, апелятивної та ідентифікаційної функцій тексту. Це неалфавітні зображення або умовні позначення, які мають смислове навантаження.

Таким чином, комікси як мультимодальний жанр характеризуються складним поєднанням вербальних та візуальних компонентів, які формують цілісний наратив. Цей жанр охоплює різноманітні форми, збагачені культурними особливостями. Структурні особливості коміксу, зокрема специфічні типи текстових елементів та мультимодальні компоненти формують унікальне мовне середовище, яке вимагає особливого підходу до аналізу і перекладу.

1.2 Еволюція методів МП та адаптація вузькоспеціалізованих текстів

Машинний переклад (далі – МП) – це технологія, яка дозволяє перетворити текст, написаний однією мовою, на іншу, використовуючи для цього комп'ютерні алгоритми. Стрімкий розвиток алгоритмів штучного інтелекту та машинного навчання популяризував застосування систем МП у сучасному середовищі. Їх активно використовують в різних сферах – від бізнесу до освіти. Головна перевага МП – здатність швидко й доступно подолати мовні бар'єри. (Tabsharani, 2024)

За словами науковиці К. Рябової (2023), історія машинного перекладу почалася у 1930-х роках із примітивних механічних пристроїв. Однак справжнього поштовху розвитку сфери надав меморандум «Переклад», написаний Ворреном Вівером у 1949 році. В цій праці автор порівняв завдання перекладу з дешифруванням повідомлень, що було популярне в роки Другої Світової. У 1954 році Леон Достерт провів першу успішну демонстрацію машинного перекладу. Однак лінгвіст Бар-Хіллел у 1960 році заявив, що знання лінгвістики та комп'ютерних систем на той період не дозволяли створити якісний автоматичний переклад. Консультативний комітет з автоматичної обробки мови (*Automatic Language Processing Advisory Committee*), створений в 1964 році, обмежив фінансування, яке виділяв уряд США, на дослідження в цьому напрямі, через що розвиток галузі було припинено на тривалий час.

Другий етап розвитку почався у 1970-1980-х роках, з появою систем-інтрелінгв (проєкт CETA, де використовували проміжну мову «pivot language» для перекладу між російською та французькою) та синтаксичних систем МП.

Паралельно з'явилися перші комерційно успішні системи, наприклад, «*Systran*», яка початково розроблялася для військових потреб, а згодом стала загальнодоступною.

З 1990-х років виокремлюються кілька провідних підходів. Популярними стають корпусні методи, водночас активно розвиваються системи на основі правил. Також проводять експерименти з системами МП «на основі діалогу» та розробкою технологій автоматичного усного перекладу, що поєднують розпізнавання мовлення і лінгвістичне тлумачення.

До початку 2000-х років у галузі сформувалося чотири покоління систем МП:

1. Прямий переклад, заснований на правилах
2. Не прямий переклад, заснований на правилах
3. Корпусні методи
4. Гібридні системи

Дослідник М. Шераґуї (2012) розглядає їх детальніше. Перші два покоління пов'язує те, що вони базуються на використанні лінгвістичних правил для перекладу тексту, такий підхід має відповідну назву англійською мовою (*Rule-based Machine Translation).* **Прямий переклад** вважається найпростішим підходом, у якому перетворення тексту здійснюється безпосередньо з мови оригіналу на цільову мову. В цьому випадку система перетворює всі слова в тексті на леми, знаходить відповідник в словнику мовної пари та формує текст перекладу. Для такої системи необхідні лише великий словник, модулі морфологічного аналізу та генерації. Недолік такого підходу в тому, що він не враховує глибокий синтаксичний або семантичний аналіз, через це застосовується тільки для споріднених мов. **Непрямі переклади** можна поділити на два типи: *інтерлінгвальний* та *трансферний*. Перший перетворює текст оригінальною мовою спочатку на інтерлінгву, тобто мову-посередник, яка є універсальною семантичною репрезентацією. І лише з цієї інтерлінгви створюється текст мовою перекладу. Найкраще це підходить для багатомовних систем. Другий тип замість проміжної

мови створює певну структуру мови оригіналу, наприклад, синтаксичне дерево, а потім перетворює її на відповідну структуру вже мовою перекладу і в результаті генерує текст. Основною проблемою цього підходу є потреба у створенні систем граматики і словників, що потребує багато часу і ресурсів. Крім того, додавання нових правил може зламати систему.

Наступне покоління – **корпусні методи** (англ. *Corpus-Based Machine Translation*). Підхід до перекладу з використанням великих двомовних корпусів для навчання моделей. Напрям виник як альтернативний підхід правилам, щоб вирішити проблему збору лінгвістичного матеріалу. Виділяють два основних типи: *статистичний машинний переклад* (англ. *Statistical Machine Translation*) і *переклад на основі прикладів* (англ. *Example-Based Machine Translation*). Завдяки використанню паралельних текстових корпусів для вивчення відповідностей між мовами, такий підхід дозволяє швидко побудувати модель для будь-якої мовної пари за наявності достатньої кількості даних.

Гібридні системи поєднують у собі всі попередні підходи. Основна ідея полягає в тому, щоб навчити модель правил перетворення за допомогою використання невеликого корпусу. Такі системи є ефективними при обмежених ресурсах, адже поєднують структурованість правил із гнучкістю статистичного навчання. Розробник може сам вирішувати на що більше звернути увагу аби отримати бажаний результат від моделі.

Останнім часом також розвивається новий підхід, яким на сьогодні користуються усі сучасні перекладачі (*Google Translate, Bing Translator, DeepL* тощо). Цей підхід називається **нейронний МП**. Сучасні системи НМП пройшли кілька послідовних архітектурних етапів, кожен з яких суттєво покращував якість перекладу. Першим етапом стало використання рекурентних нейронних мереж (англ. *Recurrent Neural Network*). Рекурентні підходи революціонізували моделювання послідовностей завдяки здатності обробляти тексти послідовно, а саме слово за словом зберігаючи «пам’ять» про попередній контекст. Саме на їх

основі у 2014 році було запропоновано архітектуру *seq2seq*, де один енкодер стискав вхідне речення у вектор фіксованої довжини, а інший декодер розгортав цей вектор у речення перекладу. Удосконалена версія таких мереж, LSTM (*Long short-term memory*) подолала головну проблему ранніх рекурентних моделей. Вони перестали забувати важливу інформацію з початку речення до моменту генерації його кінця, що суттєво покращило зв'язність і граматичну коректність перекладу. Водночас згорткові нейронні мережі (англ. *Convolutional Neural Network*) запропонували альтернативний підхід, обробляючи всю послідовність паралельно та використовуючи шаровані згорткові блоки для поступового охоплення більшої кількості контексту. Проте справжнім переломним моментом стала архітектура *Transformer* (Vaswani *et al.*, 2017), де механізм масштабованої точкової уваги дозволяє модулю енкодера будувати контекстуалізовані репрезентації всього речення одночасно, а декодеру – генерувати переклад, врахувавши значущість кожного вхідного токена незалежно від відстані між ними. Еволюція практичного впровадження цих технологій чітко простежується на прикладі платформи *Google Translate*, яка у 2016 році здійснила радикальний перехід зі застарілого статистичного підходу на нейронний машинний переклад на основі архітектури *LSTM*, що підвищило показник метрики *BLEU* на 5-10 одиниць, а переклади стали більш природними. Згодом, після публікації архітектури *Transformer* у 2017, розробники почали поетапно інтегрувати й цей підхід, що заклало основу для сучасних високоефективних систем перекладу. (Dong, 2026)

Архітектура *Transformer* представлена двома основними підходами. Класичним енкодер-декодером (англ. *encoder-decoder*) і виключно декодером (англ. *decoder-only*). Питання про те, яка архітектура є більш ефективною для задач перекладу, залишається досі актуальним. Автори однієї статті (Fu *et al*., 2023) провели порівняльний аналіз двох підходів і виявили між ними принципові структурні відмінності. В архітектурі encoder-decoder вхідне речення та речення перекладу обробляються окремо, де енкодер формує контекстне представлення

джерела, яке декодер потім використовує через механізм перехресної уваги (англ. cross-attention) на кожному кроці генерації. Натомість decoder-only модель об'єднує вхідний та вихідний тексти в єдину послідовність і опрацьовує їх одним модулем, що зменшує розмір моделі та спрощує навчання на великих обсягах немаркованих даних. Проте такий підхід породжує проблему деградації уваги (англ. *attention degeneration problem*). Дедалі більше токенів перекладу модель генерує, її увага дедалі менше зосереджується на вхідному реченні, що призводить до передчасного завершення генерації та появи галюцинацій (англ. *hallucination*) – граматично правильних, але семантично спотворених перекладів.

Попри значний прогрес у розвитку архітектур НМП, навіть найсучасніші моделі мають труднощі при перекладі вузькоспеціалізованих текстів. Наприклад, медична або юридична документація, а у контексті цього дослідження й тексти коміксів. Саме тому для цього застосовується процес доменної адаптації (англ. *domain adaptation*) – цілеспрямованого налаштування попередньо навченої моделі на нові типи текстів без необхідності перенавчання з нуля. Як зазначає Сондерс (2022), доменна адаптація охоплює будь-який підхід, спрямований на покращення перекладів попередньо створеної системи для лексики певної галузі чи особливої тематики мови. Найпоширенішим методом є тонке настроювання (англ. *fine-tuning*) – це донавчання моделі на невеликому паралельному корпусі, створеному на текстах відповідного домену (англ. *domain*). Попри простоту та ефективність цього підходу, він приховує два серйозних ризики. По-перше, **надмірне донавчання** на малому доменному корпусі може спричинити перенавчання (англ. *overfitting*), коли модель досягає високої якості перекладу на наданих прикладах, але не справляється з перекладом речень поза навчальним корпусом. По-друге, навчання на нових даних може спричинити **катастрофічне забування** (англ. *catastrophic forgetting*), коли модель покращує якість перекладу цільового домену коштом поступової втрати знань, набутих під час початкового навчання на загальних даних. Для подолання цих обмежень дослідники розробили альтернативні стратегії, зокрема

методи ефективного налаштування параметрів (англ. *Parameter-Efficient Fine-Tuning*) та адаптацію низького рангу (англ. *Low-Rank Adaptation*), які замість повного оновлення ваг моделі навчають лише невелику кількість додаткових параметрів, зберігаючи при цьому базові знання незмінними.

Отже, розвиток машинного перекладу пройшов тривалий шлях від примітивних систем до нейронних архітектур. Хоча сучасні моделі подолали обмеження попередніх систем і здатні забезпечувати якісний переклад загальних текстів, вони все ще зазнають труднощів при роботі з вузькоспеціалізованими матеріалами, зокрема текстами коміксів. Саме тому розуміння еволюції архітектурних підходів і притаманних їм проблем є необхідною теоретичною основою для подальшого аналізу можливостей доменної адаптації моделей МП до перекладу цього жанру.

1.3 Архітектура та особливості моделі «Dragoman»

Модель «*Dragoman*» була розроблена командою українських дослідників (Paniv *et al*., 2024) та представлена на воркшопі UNLP 2024 (*Ukrainian Natural Language Processing*) (Romanyshyn *et al*., 2024). Головна мотивація авторів полягала в тому, що для розбудови великих мовних моделей для української мови необхідно значно розширити корпуси навчальних даних, а оскільки більшість готових наборів даних для донавчання існує переважно англійською мовою, наявність якісної системи перекладу є критично важливою умовою для прискорення цього процесу.

В основі архітектури моделі «*Dragoman*» лежить базова попередньо натренована багатомовна модель *Mistral-7B-v0.1* (Jiang *et al*., 2023), яка показала найкращі показники серед розглянутих альтернатив під час експериментів із навчанням на кількох прикладах. Розробники використали для побудови моделі підхід умовного мовного моделювання, де головною задачею перекладу є максимізація правдоподібності цільового українського речення. На вхід подаються англійські речення, які оформленні у вигляді шаблону псевдоінструкції (*[INST] translated sentence [/INST]*). Токени вхідного речення маскуються під час

обчислення перехресної ентропії, що дозволяє моделі зосередитися виключно на генерації цільового тексту.

Для ефективного навчання без необхідності оновлення всіх параметрів великої моделі автори застосували метод навчання з низькоранговою адаптацією (англ. *Low-Rank Adaptation*) (Hu *et al*., 2022) та квантування ваг у форматі nf4 (Dettmers *et al*., 2023). При цьому оптимізації підлягали лише додаткові параметри адаптера, тоді як основні ваги попередньо натренованої моделі залишалися незмінними. Такий підхід оптимізував використання відеопам'яті на споживчому графічному процесорі до 24 Гб, що робить відтворення результатів доступним для широкого кола дослідників.

Процес навчання «Dragoman» складався з **двох** послідовних етапів:

1. Модель навчалася на відфільтрованому підкорпусі набору даних «*Paracrawl*» (Bañón *et al*., 2020), що містить паралельні речення українською та англійською, зібрані автоматично з текстів Інтернету. Початкова версія корпусу складалася з 13,354,365 речень, але значну кількість шуму: повторювані прогнози погоди, тексти навігаційних меню сайтів, низькоякісні машинні переклади та речення неукраїнською мовою тощо. Тому автори виконали багаторівневу евристичну фільтрацію:

- за допомогою бібліотеки *gcld3* було відфільтровано усі речення, які не були визначені бібліотекою як українські чи англійські;
- було обчислено порогове значення перплексії на основі двох монолінгвальних моделей, де речення з нетиповим значенням за показником бітів на символ вилучалися. *Перплексія* – це метрика, що вимірює рівень невизначеності мовної моделі під час передбачення наступного токена в послідовності;
- фільтрація на основі косинусної подібності ембедингів речень LaBSE (*Language-Agnostic BERT Sentence Embedding*) (Feng *et al*., 2022) була

використана, щоб відсіяти речення, які були погано розмічені в паралельному корпусі;

- перевірка довжин речень була застосована для виявлення пар, в яких сильно різняться довжини.

На виході оптимальним виявився підкорпус обсягом 2,9 мільйона пар, відсортованих за схожістю *LaBSE*-фільтру від найменш схожих до найбільш схожих. Модель натренована на цьому підкорпусі показала найвищий показник за метрикою BLEU – 30.37 (з 100), ніж корпуси іншого обсягу (див. *Табл. 1.1*).

Dataset	Pairs	Filters				Example Order	Best BLEU ↑
		Lang	BPC	LaBSE	Len diff		
1m unfiltered	963k	-	-	-	-	Random	28.26
1m filtered	958k	En/Uk	<3.33	>0.91	<50	Random	29.47
3m filtered	2.9m	En/Uk	<3.25	>0.85	<50	By LaBSE score, dissimilar first	**30.37**
8m filtered	8m	En/Uk	<5	>0.5	<50	By LaBSE score, dissimilar first	30.19

Табл. 1.1. Результати за метрикою BLEU для різних підкорпусів.

2. Далі модель донавчали на наборі даних «*Extended Multi30k-Uk*» (Saichyshyna *et al*., 2023). На цьому етапі автори розробили алгоритм перехресної перевірки K-Fold для фільтрації датасету за перплексією. Суть методу полягає в тому, що для кожного речення обчислюється його логарифмічна правдоподібність за моделлю, яка не бачила цього речення під час навчання, та після чого речення з низьким показником відсіюється. Це дозволило позбутися семантично неправильних речень та залишити 17 тисяч найкращих зразків для фінального донавчання.

Такий підхід до архітектури даних дав відчутний приріст якості генерації тексту. І як зазначили самі автори, кінцева версія моделі «*Dragoman*», яка була побудована на базі decoder-only, успішно перевершила попередні найсучасніші моделі типу encoder-decoder за метрикою *BLEU* (див Табл. 1.2) на тестовому наборі даних «*FLORES-101 devtest*» (Goyal *et al*., 2022).

Model	BLEU ↑
Finetuned	
Dragoman P, 10 beams (section 3)	30.4
Dragoman PT, 10 beams (section 4)	**32.3**
Zero shot and few shot (section 5)	
Llama 2 7B 2-shot, 10 beams	20.1
Mistral-7B-v0.1 2-shot, 10 beams	24.9
gpt-4 10-shot	29.5
gpt-4-turbo-preview 0-shot	30.4
Pretrained encoder-decoder	
NLLB-3B, 10 beams	30.6
OPUS-MT, 10 beams	**32.2**

Табл 1.2. Результати оцінювання моделей за метрикою BLEU для перекладу en-uk.

Вибір моделі «*Dragoman*» як основи для подальшого донавчання в контексті цього дослідження є обґрунтованим з кількох принципових міркувань. По-перше, модель та весь код для її навчання й оцінювання розміщені у відкритому репозиторії на GitHub-сторінці, що забезпечує повну відтворюваність та можливість адаптації під специфічні потреби дослідження. По-друге, на відміну від загальних багатомовних систем, «*Dragoman*» цілеспрямовано оптимізована саме для пореченнєвого перекладу з англійської на українську мову, що максимально відповідає задачі перекладу діалогових реплік і підписів коміксів, які загалом подаються у вигляді коротких, семантично щільних одиниць тексту. По-третє, «*Dragoman*» є фактично першою публічно доступною decoder-only моделлю машинного перекладу, розробленою українськими дослідниками з фокусом на українську мову, що надає дослідженню додаткової наукової та практичної цінності в контексті розвитку україномовних технологій обробки природної мови.

РОЗДІЛ 2. ФОРМУВАННЯ ПАРАЛЕЛЬНОГО КОРПУСУ ДЛЯ ДОНАВЧАННЯ

Другий розділ описує практичну підготовку мовних даних, які необхідні для процесу донавчання. Формуванню даних було приділено окрему увагу, щоб отримати якісний та репрезентативний корпус здатний виконати задачу доменної адаптації моделі. Розділ містить два підрозділи: перший охоплює процес відбору, фільтрації та перекладу текстів із набору «*COMICS TEXT+*» для створення навчального корпусу, другий – підготовку тестового корпусу на основі офіційного перекладу коміксу «*Amazing Spider-Man: The Clone Conspiracy*», який використовувався для оцінювання якості перекладу.

2.1 Формування спеціалізованого паралельного корпусу для донавчання

Основою для формування корпусу став відкритий набір даних «*COMICS TEXT+*» (Soykan, Yuret and Sezgin, 2022) – масштабна колекція англійських текстів, автоматично витягнутих зі сторінок коміксів за допомогою системи оптичного розпізнавання символів. Корпус розміщений у форматі *CSV,* і містить такі поля: *номер коміксу*, *номер сторінки*, *номер панелі*, *номер текстової комірки*, *позначка типу тексту* (діалог або оповідь), сам *текст* та *координати* його розташування на панелі за осями X і Y. Загальний обсяг набору становить понад 2 мільйони рядків. Принципово важливо, що одиницею корпусу є рядок, а не речення у традиційному лінгвістичному розумінні. Текст розміщений в мовних бульках рідко утворює синтаксично завершені речення і не має чіткої міжреченнєвої пунктуації. Крім того, система оптичного розпізнавання визначає координати окремих текстових фрагментів і відповідно до них ділить усі знахідки на окремі рядки.

comic_no	page_no	panel_no	textbox_no	dialog_or_narration	text	x1	y1	x2	y2
0	0	1	0	1	account of - s wiggins	11	17	361	190
0	0	1	1	1	no . . uu ioc	855	5	1067	154
0	0	1	2	1	i account uk ss wiggins man ho oke aw of vity !	1	30	264	644
0	2	0	0	1	him i plas ! him vn !	0	0	106	132
0	2	1	0	1	i don ' t understand this , woozy ! he just won ' t fall down !	752	511	957	676
0	2	1	1	2	i his career as as crime - fighter , plastic man , fearless agent of the f . b . i . has encountered every variety of aw - breaker . the grifter the grafter . the smuggler and the traitor ! but never before has he coped with the likes of the fabulous weightless wiggins . . . the man who broke the gravity	0	657	380	1133
0	3	0	0	1	now if i put this here . - - - and this one here . . .	129	21	363	134
0	3	1	0	1	and adjust this spring . . .	20	20	272	85
0	3	2	0	1	that does it ! it ' s finished and ready for the final test !	28	9	317	119
0	3	3	0	1	i ' ve been working on it for days ! ! i must get a breath of air and clear my head before i begin my test !	26	1	314	157

Рис. 2.1. Вигляд набору даних «COMICS TEXT+».

З огляду на практичні обмеження процесу навчання моделі цільовий обсяг корпусу було визначено у 2000 паралельних сегментів. Однак для забезпечення цього обсягу після перекладу для первинної вибірки було встановлено розмір 2500 рядків з урахуванням подальших втрат. Для розв'язання цієї задачі було застосовано скрипт (див. *Додаток А*) зі стратифікованою вибіркою (англ. *stratified sample*). З оригінального набору було взято два поля: *текст* та *позначка типу тексту*. На першому етапі скрипт застосовує фільтр довжини, де рядки, що містять менш як 5 або понад 40 слів, відкидаються. Надто короткі фрагменти не несуть достатнього перекладного контексту, а надто довгі є нетиповими для жанру й можуть свідчити про помилки розпізнавання. Після фільтрації рядки розподілялися на чотири категорії:

1. Короткі діалоги (5-9 слів) – 15% вибірки.
2. Середні діалоги (10-20 слів) – 50% вибірки.
3. Довгі діалоги (21-40 слів) – 25% вибірки.
4. Рядки оповіді – 10% вибірки.

Такий підхід було використано, щоб забезпечити репрезентативність корпусу щодо різних типів коміксного тексту та запобігти домінуванню однієї групи текстів у тренувальних даних. Після формування вибірки рядки перемішувалися у випадковому порядку.

Для автоматичного перекладу відібраних відібраних англійських рядків на українську мову було вирішено використовувати велику мовну модель «*Google Flash 3*» (Google, 2026) завдяки її здатності виконувати якісний переклад українською у великих обсягах. Додатково було застосовано інженерію запитів для роз'яснення моделі жанрової специфіки текстів коміксу (*Рис. 2.2*). Оскільки через недосконаліть автоматичного розпізнавання оригінальний корпус містить численні артефакти та рядки, непридатні до перекладу, у запиті було передбачено умову, за якої модель може пропускати такі рядки (*Рис. 2.2*). Саме тому початкова вибірка становила 2500 рядків, а після фільтрації пар фінальний обсяг паралельного корпусу склав 2000 сегментів.

```
You are a comics transnlator (en-uk). Translate comic book dialogue to Ukrainian. I will give you lines extracted from random comics using ocr system, because of that there are a lot of trash in those line.
Rules:
- DO NOT translate: ocr artifacts, untraslatable words, etc. Instead SKIP those elements or SKIP those lines fully.
- Try to traslate naturally and correctly even if there is mistakes in orginal lines.
- Keep informal/colloquial register (not literary)
- Sound effects stay as-is OR use Ukrainian equivalents (BOOM → БУМ)
- Short exclamations must feel natural in Ukrainian speech
- Do NOT over-translate — if English is 3 words, Ukrainian ~3-5 words max
- No footnotes, no explanations, orginal line - translation only. For example:
{source} - {your translation} (keep {})
```

Рис. 2.2. Запит для моделі «Flash 3» перед початком перекладу.

```
{how come ? anything wrong out to the hallidays ?} - {Як так? Щось не те у Галлідейв?}
{that ' s what you think , but i have other ideas !} - {Це ти так думаєш, а у мене інші плани!}
{take this one of me , pop do ilook like a real avidal trainer ?} - {Сфоткай мене, тату! Я схожий на справжнього дресирувальника?}
{like dis . . . oof !} - {Ось так... Ох!}
{he sure did a swell job on these mugs i .} - {Він добряче відробив цим пикам.}
{ia strange prisoner is brought to the village warriors where tygra rules as queen !} - {Дивного полоненого приводять до воїнів селища, де королевою править Тайгра!}
{good ! things might get hot if anybody spots us !} - {Добре! Нам непереливки буде, якщо нас хтось помітить!}
{he ' s hard up for dough offer him a hundred bucks an ' he ' ll grab it} - {Йому капець як потрібні гроші, запропонуй йому сотню баксів, і він вхопиться за них.}
{what made you think that they might be hiding ' em in that old mine phil ?} - {З чого ти взяв, Філе, що вони можуть ховати їх у тій старій шахті?}
{i don ' t like being shot at , friend !} - {Не люблю, коли в мене стріляють, друже!}
```

Рис. 2.3. Фрагмент корпусу після перекладу моделлю «Flash 3».

Далі відбувся процес форматування корпусу відповідно до вимог моделі «*Dragoman*» для процесу донавчання. Кожен сегмент оформлювався у вигляді інструкційного шаблону (*[INST] translated sentence [/INST]*), а корпус загалом записувався у форматі *JSON* (*Рис. 2.4.*). Весь процес виконувався автоматично за допомогою окремого скрипту (див. *Додаток А*).

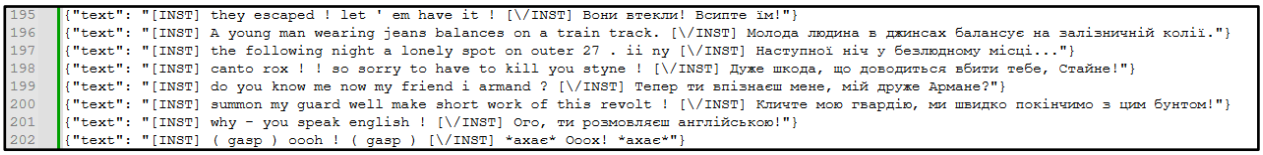

Рис. 2.4. Фрагмент корпусу після форматування до вимог моделі «Dragoman».

Останнім етапом підготовки корпусу стало змішування корпусу коміксів з паралельними парами загальної тематики, які початково використовувалися для навчання базової моделі «*Dragoman*», а саме відкритого набору даних «*Extended Multi30k-Uk*». До 2000 сегментів комікcного тексту скриптом було додано 1000 випадково відібраних рядків зі згаданого набору, рівномірно розподілених по всьому корпусу. Мета такого змішування полягає у реалізації стратегії змішаного донавчання (англ. *mixed fine-tuning*). Цей підхід запобігає явищу катастрофічного забування та дозволяє моделі зберігати загальну якість перекладу, набуту під час попереднього навчання, водночас адаптуючись до специфіки комікcного тексту.

2.2 Формування паралельного тестового корпусу

Для проведення оцінки якості базової моделі «*Dragoman*», моделі «*OPUS-MT*» та донавченої моделі на текстах коміксу було використано паралельний корпус, сформований на основі англомовного коміксу «*Amazing Spider-Man: The Clone Conspiracy*» (Gage and Slott, 2017) та офіційного українського перекладу видавництвом «*Fireclaw*» (2026). Попередньо корпус було створено в межах курсової роботи «*Аналіз особливостей перекладу сучасних коміксів*» (2024) та повторно використано в курсовій роботі «*Оцінка якості машинного перекладу текстів коміксу*» (2025). Результатом роботи стали тексти зазначеного коміксу, вирівняні за окремими текстовими елементами (мовні бульки, текстові блоки,

надписи). З метою обмежити обсяг дослідження було використано лише паралельні тексти першого з десяти випусків, що має обсяг 231 рядків. Також, для зручності використання еталона при оцінюванні метриками текст англійською та українською розділено на окремі файли (див. *Рис. 2.5* та *Рис. 2.6*).

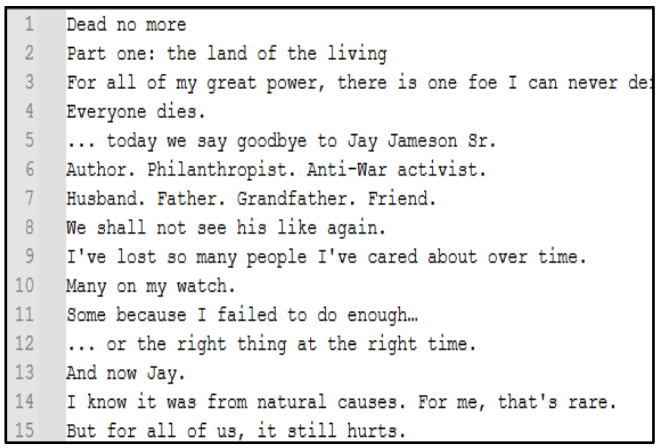

Рис. 2.5. Фрагмент оригінального тексту коміксу.

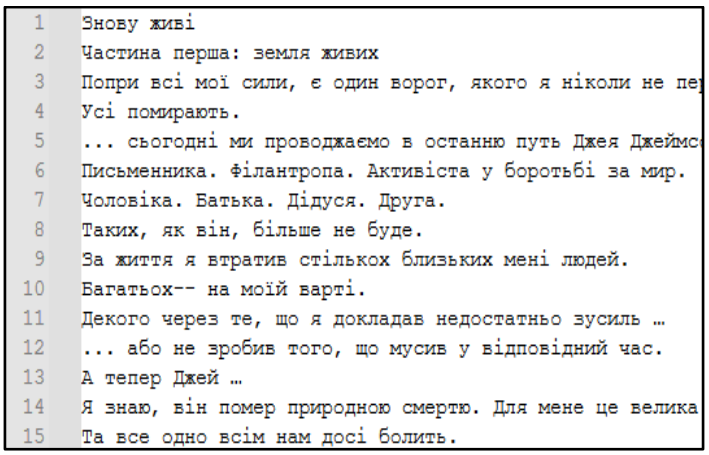

Рис. 2.6. Фрагмент експертного перекладу тексту коміксу.

Викликом у досліджуваному коміксі стало те, що весь текст подано у верхньому регістрі. Це збереглося й при побудуванні паралельного корпусу. Для забезпечення еталонного корпусу паралельних перекладів текстів коміксу виникла потреба попередньо нормалізувати обидва тексти (англійський та український). Для цього тексти було збережено в окремі текстові файли та за допомогою спеціальних скриптів (див. *Додаток А*), написаних програмною мовою *Python* із використанням відповідних сторонніх бібліотек, було автоматизовано процес нормалізації текстів:

- Для оригінального тексту, написаного англійською мовою, було застосовано бібліотеку *TrueCase* (Fury, 2021). Вона використовує статистичну модель на основі N-грам, яка визначає ймовірний регістр слів за контекстом. А також опирається на словникові бази, зібрані з великих англомовних корпусів, щоб визначити найчастотнішу форму написання слів.

- Для українського офіційного перекладу окремої бібліотеки для визначення регістру слів не існує, тому в роботі було використано лінгвістичну модель *uk_core_news_lg* (корпус українських новинних текстів обсягом приблизно 220 МБ), що входить до складу бібліотеки *spaCy* (Honnibal *et al.*, 2020). Для цього текст спочатку розбивається на окремі слова, які аналізуються за частинами мови. Після цього, з урахуванням контексту, відновлюється типовий для української мови регістр усіх слів.

Якість комп'ютерної обробки обох текстів виявилася різною. Бібліотека *TrueCase* справилася з англійським текстом краще, ніж методи на основі бібліотеки *spaCy* для української мови. Це можна пояснити тим, що в оригіналі текст написано англійською мовою, тому іменовані сутності (імена персонажів, локації тощо) краще представлені у словникових базах. Крім того, обсяг навчального корпусу для англійської мови перевищує український, що також впливає на точність відновлення регістру.

РОЗДІЛ 3. ПРОЦЕС ДОНАВЧАННЯ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

Третій розділ є центральним компонентом роботи, де теоретичні знання та підготовлені дані реалізуються в конкретному дослідницькому процесі. Розділ складається з чотирьох підрозділів: у першому описано технічну реалізацію донавчання моделі «*Dragoman*» із зазначенням конфігурації гіперпараметрів та динаміки навчання, у другому – оцінювання якості перекладу за автоматичними метриками *BLEU*, *TER*, *chrF* та *COMET*, у третьому – лінгвістичний аналіз типових помилок і покращень, у четвертому – практична реалізація донавченої моделі в пайплайні з графічним інтерфейсом.

3.1 Донавчання моделі «Dragoman»

Для проведення експерименту з донавчання моделі «*Dragoman*» було обрано хмарне обчислювальне середовище *Google Colaboratory* (Google, 2026) – онлайн-платформа, що надає вільний доступ до графічного процесора *Tesla T4* з обсягом відеопам’яті 16 ГБ, який здатний забезпечити достатню потужність для проведення процесу донавчання за відносно невеликий проміжок часу та не потребує локального налаштування. Для програмної реалізації використовувалася низка бібліотек мови програмування *Python*, які безпосередньо пов’язані з машинним навчанням, зокрема:

- *Transformers* (Wolf *et al*., 2020) – для завантаження базової моделі перекладу «*Mistral-7B-v0.1*» та загального токенізатора. Також налаштування параметрів квантування через клас *BitsAndBytesConfig*;
- *PEFT* (Mangrulkar *et al*., 2022) – для підключення LoRA-адаптера моделі «Dragoman» та підготовки моделі до навчання в режимі квантування;
- *TRL* (von Werra *et al*., 2020) – для організації процесу донавчання та конфігурації усіх гіперпараметрів навчання;
- *PyTorch* (Paszke *et al*., 2019) – для роботи з тензорами та графічним процесором;
- *Datasets* (Lhoest *et al*., 2021) – для завантаження та обробки навчального корпусу.

Також для зберігання проміжних і фінального результатів донавчання було під'єднано хмарне сховище *Google Drive*. Для автентифікації на платформі *Hugging Face Hub* та отримання доступу до ваг моделі використовувався персональний токен доступу, що є стандартною вимогою платформи при завантаженні моделей через програмний інтерфейс. Повний скрипт донавчання знаходиться в *Додатку Б*.

Першим етапом роботи було завантажено підготовлений паралельний корпус з текстами коміксів. Для забезпечення можливості моніторингу якості навчання та своєчасного виявлення ефектів перенавчання датасет було розділено на дві частини: **навчальну** та **валідаційну**. Валідаційна вибірка складала 5% від загального обсягу, що відповідає 150 прикладам, а навчальна – 2850 прикладів. Параметр *seed=42* забезпечив відтворюваність розбиття даних при повторних запусках, щоб до валідаційної вибірки завжди потрапляли ідентичні приклади.

Оскільки модель «*Mistral-7B-v0.1*» містить близько 7 мільярдів параметрів, її завантаження із повною точністю потребувало б понад 28 ГБ відеопам'яті, що значно перевищує ресурси обраного графічного процесора. Тому наступним етапом стало застосування 4-бітного квантування ваг. Аналогічну конфігурацію використовували й автори моделі «*Dragoman*» під час її первинного навчання (Paniv *et al*., 2024). Це забезпечує архітектурну сумісність та дозволяє скоротити необхідний обсяг відеопам'яті приблизно до 5-6 ГБ.

Третій етап донавчання передбачав послідовне завантаження трьох ключових компонентів: токенізатора, базової моделі та адаптера «*Dragoman*». Спочатку ініціалізувався токенізатор моделі «*Mistral-7B-v0.1*» з вимкненим автоматичним додаванням токена початку послідовності, оскільки інструкційний формат вхідних даних (*[INST] translated sentence [/INST]*) вже містить необхідні службові токени. Після цього завантажувалася сама базова модель із застосуванням раніше визначеної конфігурації квантування та готувалася до навчання за допомогою функції *prepare_model_for_kbit_training*, що дозволило мінімізувати пікове споживання відеопам'яті. Завершальним кроком стало підключення LoRA-

адаптера «*Dragoman*» поверх підготовленої базової моделі через клас *PeftModel* з параметром *is_trainable=True*, який переводить завантажений адаптер у режим навчання і дозволяє оновлювати ваги моделі.

Останній етап налаштування гіперпараметрів охоплював визначення як загальних параметрів організації процесу навчання, так і специфічних параметрів оптимізації. З головних параметрів, які варто виділити:

- *learning_rate* (швидкість навчання) – визначає розмір кроку оновлення ваг моделі під час оптимізації;
- *lr_scheduler_type* (планувальник швидкості навчання) – стратегія зміни темпу навчання протягом тренування;
- *per_device_train_batch_size* (розмір батчу) – визначає кількість прикладів, після обробки яких оновлюються ваги моделі;
- *num_train_epochs* (кількість епох) – кількість повних проходів через навчальний корпус
- *warmup_ratio* (коефіцієнт прогріву) - частка кроків навчання, протягом яких швидкість навчання поступово зростає від нуля до заданого значення.

Після серії експериментів з різними комбінаціями гіперпараметрів емпіричним шляхом було визначено оптимальну конфігурацію, за якої модель демонструвала стабільне навчання без ознак перенавчання чи розбіжності оптимізатора:

Гіперпараметр	**Конфігурація**	**Обґрунтування**
learning_rate	2e-5	Відповідає значенню, що використовувалося авторами «Dragoman».
lr_scheduler_type	cosine	Забезпечує плавне затухання темпу навчання, що стабілізує процес оптимізації на пізніх етапах.

per_device_train_batch_size	1	Мінімальний розмір батчу для запобігання вичерпанню відеопам'яті.
num_train_epochs	3	Забезпечує достатню кількість оновлень ваг для адаптації до доменної лексики коміксів.
warmup_ratio	0.03	Перші 3% кроків темп навчання зростає від нуля, що запобігає нестабільності на старті.

Табл. 3.1. Значення основних гіперпараметрів донавчання.

Після завантаження усіх моделей та встановлення гіперпараметрів почався процес донавчання, який загалом тривав понад 5 годин та складався з 537 кроків. Під час першого запуску було виконано одну епоху навчання з метою первинної оцінки поведінки моделі та коректності обраної конфігурації гіперпараметрів. Після аналізу результатів та підтвердження стабільної динаміки навчання було прийнято рішення продовжити тренування ще на дві епохи, використовуючи механізм відновлення з контрольної точки (англ. *checkpoint*), де проміжні результати зберігалися кожні 100 кроків на хмарному сховищі *Google Drive*. По закінченні фінальна модель також була збережена на хмарному сховищі. Наразі фінальна модель також розміщена у відкритому репозиторії на платформі *Hugging Face Hub* (див. *Додаток В*).

Аналіз процесу навчання засвідчив стабільне та поступове зниження значення функції втрат протягом усього тренування. Функція втрат (англ. *loss function*) – це математична міра, що кількісно відображає розбіжність між передбаченнями моделі та еталонними значеннями, і саме її мінімізація є метою процесу навчання. На початку першої епохи значення навчальної втрати, що обчислюється безпосередньо на даних, які модель використовує для оновлення своїх ваг, становило 3.53, після чого спостерігалося її поступове зниження (див. *Табл. 3.2*). До кінця першої епохи – 1.45, до кінця другої – 1.31, до кінця третьої –

1.02. Паралельно з цим середня точність передбачення токенів на навчальній вибірці зросла з 50.6% на першому кроці до 78.9% наприкінці навчання. Натомість валідаційна втрата, що вимірюється на незалежній вибірці даних, визначалася тільки на контрольних точках і завдяки параметру *load_best_model_at_end=True* точка з найменшим показником була автоматично збережена як фінальний результат донавчання (див. *Табл. 3.3*).

Етап тренування	**Навчальна втрата**	**Точність (%)**
Епоха 0.05	3.53	50.6
Епоха 1.0	1.45	70.6
Епоха 1.5	1.32	72.6
Епоха 2.0	1.31	72.8
Епоха 2.5	1.03	78.3
Епоха 3.0	1.02	78.9

Табл. 3.2. Показники функції втрат та точності на навчальній вибірці.

Контрольна точка (кількість кроків)	**Валідаційна втрата**
100	1.552
200	1.526
300 (фінальний адаптер)	**1.466**
400	1.497
500	1.497

Табл. 3.3. Показники валідаційної втрати за контрольними точками.

3.2 Оцінювання систем МП автоматичними метриками

Для комплексного оцінювання якості перекладу донавченої моделі у порівнянні зі стандартною моделлю «*Dragoman*» та альтернативною нейронною моделлю МП «*OPUS-MT*» (Tiedemann and Thottingal, 2020) було використано чотири автоматичні метрики. Сюди входять як класичні підходи до вимірювання

якості перекладу (*BLEU*, *TER*, *chrF*), так і сучасні нейронні методи оцінювання (*COMET*).

- Метрика **BLEU** (*Bilingual Evaluation Understudy*) є найпоширенішою. Вона була представлена у 2002 році й до сьогодні є актуальною. Метрика порівнює N-грами перекладу з його експертним варіантом, та обчислює геометричне середнє N-грамної точності. Стандартний алгоритм враховує від уніграми до чотири-грами, а також може накладати штраф за короткий переклад, що застосовується, коли речення перекладу вийшло коротшим, ніж в експертному варіанті. (Papineni *et al.*, 2002)
- Метрика **TER** (*Translation Edit Rate*) вимірює якість машинного перекладу, підраховуючи мінімальну кількість редагувань (вставок, видалень, замін і перестановок слів), які потрібні, щоб довести машинну версію перекладу до її експертного варіанту. (Snover *et al.*, 2006)
- Ще один алгоритм – **chrF** (*Character-level F-score*), який відрізняється від попередньої метрики *BLEU* тим, що замість N-грам слів використовує N-грами символів. Метрика обчислює F-міру (точність і повноту збігів) між машинним і експертним перекладом на рівні символів, що робить її незалежною від токенізації й особливо ефективною для мов із багатою морфологією. (Popović, 2015)
- Також, з появою та розвитком нейронних мереж почали домінувати й метрики, створені на їх основі. Одною з таких метрик є **COMET** (*Crosslingual Optimized Metric for Evaluation of Translation*), яка використовує багатомовні трансформери та об’єднує інформацію з вихідного речення, машинного та експертного перекладу. Щоб надати точну оцінку модель навчається на різних типах людських оцінок. Базова версія метрики все ще потребує еталонного варіанту, але існує й тип системи без цього – *COMET-QE* (*Quality Estimation*). (Rei *et al.*, 2020)

Усі зазначені метрики, за винятком *TER*, мають шкалу оцінювання від 0 до 1, де вищі значення свідчать про кращу якість перекладу. Натомість *TER* працює за зворотним принципом. Метрика підраховує мінімальну кількість редагувань, необхідних для перетворення автоматичного перекладу на еталонний варіант, тому нижчі значення відповідають вищій якості перекладу, а теоретична нижня межа дорівнює нулю.

Основне тестування по всім метрикам відбувалося за тестовим корпусом, який було описано в *Розділі 2.2*. Проте варто зазначити, що обсяг тестового корпусу виявився недостатнім для статистично коректного застосування класичних N-грамних метрик оцінювання. Усі обчислення здійснювалися за допомогою скриптів (див. *Додаток Г*), реалізованих мовою програмування *Python* та бібліотеками *sacreBLEU* (Post, 2026) і *Unbabel-COMET* (Rei *et al*., 2025).

Результати оцінювання трьох систем машинного перекладу за чотирма метриками наведені в *Таблиці 3.4*:

Система	**BLEU ↑**	**TER ↓**	**chrF ↑**	**COMET ↑**
OPUS-MT	0.109	0.858	0.284	0.757
Dragoman	0.165	0.799	0.385	0.797
Comics LoRA	**0.190**	**0.762**	**0.409**	**0.815**

Табл. 3.4. Результати оцінювання за автоматичними метриками на тестовому корпусі.

Спостерігаємо стійку перевагу донавченої моделі над обома базовими системами за всіма метриками. Модель «*OPUS-MT*» показала найнижчі результати, що можна пояснити її архітектурними особливостями (encoder-decoder). Так, наприклад, у порівнянні модель «*Dragoman*» суттєво перевершує «*OPUS-MT*», через кращу прилаштованість архітектури decoder-only для пореченнєвого перекладу з англійської на українську. Натомість донавчена модель демонструє покращення відносно базової моделі «*Dragoman*». Хоча класичні метрики не показали прийнятних результатів, через малий обсяг тестової вибірки та

нездатність метрик вдало аналізувати контекст, що критично важливо в текстах коміксів, значення *BLEU* покращилося на 15.2%. Найбільш показовим є результат за нейронною метрикою *COMET*, де приріст становить 0.018 пункти – від 0.797 до 0.815. Оскільки *COMET* оцінює семантичну еквівалентність перекладу, а не лише лексичний збіг з еталоном, це свідчить про те, що донавчання на корпусі текстів коміксу дозволило моделі краще відтворювати змістове наповнення реплік, характерних для цього жанру.

3.3 Лінгвістичний аналіз помилок у МП

Варто також розглянути конкретні приклади перекладу, щоб якісно доповнити результати автоматичного оцінювання та виявити типові помилки, які вдалося подолати в процесі донавчання, а також ті, що залишилися невирішеними.

Загалом моделі вдалося подолати основні труднощі перекладу текстів коміксу. По-перше, це випадки, де донавчена модель суттєво покращила як лексичний вибір, так і синтаксичну структуру речення (див. *Табл. 3.5*).

Оригінал	**Dragoman**	**Comics LoRA**
I ‘ve lost so many people I’ve cared about **over time.**	Я втратив стільки людей, про яких мені **було дороге з часом**.	**За цей час** я втратив так багато людей, про яких дбав.
You were the one who **talked him out** of the treatments...	Ви були тим, хто **переконував його від** лікування...	Ти сам **переконував його відмовитися від** лікування...
I used that man as my **guinea pig**, Anna.	Я використав цього чоловіка як свого **морського свиняти**, Анна.	Я використав того чоловіка як **піддослідного**, Ганно.

See? This is what's nice about having a new Electro…	Бачте, що приємно в тому, щоб мати новий електромобіль.	Бачиш? Ось у чому приємність нового Електро...
I--I had video of it on my **Webware**.	Я, я мав відеозапис на своєму **веб-сайті**.	У мене... у мене було відео на моєму **веб-пристрої**.

Табл. 3.5. Приклади покращення лексичної та синтаксичної точності.

Головною проблемою базової моделі «*Dragoman*» було зберігання дослівної структури англійського речення, що призводить до граматично некоректного результату, тоді як донавчена модель перебудовує речення природною українською мовою. Також помітне покращення у відтворенні фразеологічних одиниць та ідіом, де донавчена модель правильно відтворює значення цих одиниць. Наприклад, у третьому прикладі базова модель буквально перекладає ідіому «*to use as guinea pig*» як «*використати як морського свиняти*», тоді як донавчена модель використовує відповідний еквівалент «*використав як піддослідного*». Покращилося розпізнавання іменних сутностей. Як наведено в четвертому прикладі, базова модель «*Dragoman*» інтерпретує власну назву «*Electro*» як частину слова «*електромобіль*», що можна вважати за галюцинацію, адже модель припустила, що в цьому контексті слово було недописано чи написано неправильно.

По-друге, треба виділити проблему перекладу коротких розмовних реплік та усталених виразів, характерних для жанру коміксу. Це було однією з головних проблем невдалого перекладу. З наявних прикладів (див. *Табл. 3.6*) можна сказати, що модель точно покращилася в цьому плані.

Оригінал	**Dragoman**	**Comics LoRA**
Here we go.	Ось ми йдемо.	Ось і почалося.
One sec…	Один секунда...	Хвилинку...

Security!	Безпека!	Охорона!

Табл. 3.6. Приклади застосування покращення відтворення коротких реплік.

Основна проблема такої лексики в тому, що її семантика є ситуативною та ідіоматичною. Часто відірвані від основного контексту такі репліки створюють перешкоду до якісного перекладу, через відсутність розуміння попереднього або наступного контексту в моделі. Саме тому базова модель, яка не була навчана на подібних текстах, абсолютно не справлялася з поставленою задачею і натомість перекладала дослівно. Донавчена модель демонструє значно вищу здатність до прагматично коректного відтворення подібних виразів, що безпосередньо пов'язано з доменною специфікою навчального корпусу.

По-третє, відбулися покращення на рівні морфології, де модель почала правильно відмінювати іменники у звертаннях (див. *Табл. 3.7*).

Оригінал	**Dragoman**	**Comics LoRA**
Dad, don't.	**Тато**, не роби.	**Тату,** не роби.
Jonah, please…	**Йона**, будь ласка...	**Джоне**, будь ласка...
Too slow, **dolt**!	Занадто повільно, **дурня**!	Занадто повільно, **дурню**!

Табл. 3.7. Приклади застосування кличного відмінка у звертаннях.

Тексти коміксів складаються з діалогів, тому наявність великої кількості звертань неминуча. Базова модель повністю ігнорувала застосування кличного відмінка, що є граматичною помилкою. Донавчена модель змогла подолати цей бар'єр і доволі ефективно справляється з цією задачею, як з власними назвами, так і звичайними іменниками.

Водночас були певні систематичні помилки, які не вдалось повноцінно усунути після донавчання. Аналіз результатів виявив дві стійкі проблеми, що залишилися невирішеними. Першою є некоректне відтворення звуконаслідування (див. *Табл. 3.8*). Хоча помітні деякі зміни в тому як модель сприймає ці одиниці,

але через повну відсутність усталених відповідників у паралельних корпусах загальної тематики та недостатню кількість у навчальних даних, результат виявився нестабільним. Натомість модель або залишає оригінальні англійські форми без змін, або занадто узагальнює окремі вирази, що виглядає недоречним. В окремих випадках може сприймати звуконаслідування як окремі слова і спробувати перекласти їх. Другою невирішеною проблемою стала обробка навмисно розірваних реплік (див. *Табл. 3.9*), що є типовим стилістичним прийомом коміксу, де речення обривається на позначення переривання мовлення. Частково причиною є обраний підхід до перекладу, а саме пореченнєвий, через що модель не може об'єднати репліки, якщо ті поділені на рядки. Також моделі бракує знань для відтворення неповного речення або фрази. Модель намагається завершити речення самостійно, але це призводить до семантичних спотворень, як показано у прикладах. Також навіть серед труднощів, які вдалося подолати моделі, трапляються випадки галюцинацій або граматично неправильних перекладів.

Оригінал	**Comics LoRA**
WAKK	WAKK
Gnh!	Ой!
Arghh!	Ой!
ZRACKZZ	ЗРАКИ!

Табл. 3.8. Приклади невдалого звуконаслідування.

Оригінал	**Comics LoRA**
Spidey? You're breaking u--	Людина-павук? Ти ламаєш...
Idn't get that last par--	Ти не захопив останню частину!

Табл. 3.9. Приклади невдалого перекладу розірваних реплік.

Тож попри подолання моделлю основних проблем, які попередньо виникали у базовій моделі, варто розуміти, що модель все ще не досконала і потребує доопрацювань в майбутньому. Такий досвід варто використати, щоб наголошувати на конкретних слабких місцях моделі та впровадити рішення, щоб підвищити загальну якість перекладу.

3.4 Практична реалізація моделі в умовах реального пайплайну

Для інтеграції донавченої моделі перекладу «*Dragoman*» у прикладне програмне забезпечення було розроблено наскрізний пайплайн (англ. *pipeline*) обробки сторінок коміксів. Він охоплює три послідовні етапи: розпізнавання англійського тексту безпосередньо із зображення за допомогою системи OCR (*Optical Character Recognition*), передачу витягнутого тексту до натренованої нейромережевої моделі та генерацію перекладу українською мовою, а також візуальне накладання результатів на оригінальне зображення.

З огляду на потребу великої кількості обчислювальних ресурсів для послідовного запуску моделі розпізнавання тексту та моделі перекладу, було вирішено створити клієнт-серверну архітектуру. Нейромережеві обчислення винесені на хмарний сервер (*Google Colaboratory*), а локальний клієнт виконує функції графічного інтерфейсу користувача, керування файлами та фінального відображення перекладеного тексту поверх оригіналу.

Варто детальніше оглянути кожен з етапів та описати процес роботи пайплайну:

1. Локальний застосунок передає зображення сторінки коміксу на сервер для виконання OCR. Для цього було використано HTTP (*Hypertext Transfer Protocol*) запит на захищену адресу сервера, який був розгорнутий у середовищі *Google Colaboratory* за допомогою тунелювання *Ngrok* (Ngrok, 2026). На сервері виконується сам процес витягування тексту. Для цього було обрано систему розпізнавання тексту, яка налаштована саме для комікснних панелей – «*ComiQ*» (Kanishq, 2026). Це спеціалізована система, яка поєднує стандартні моделі OCR та алгоритми сегментації за допомогою штучного інтелекту для точного визначення меж текстових бульок на панелі. Завдяки такій архітектурі система «*ComiQ*» не лише розпізнає літери, а й ефективно локалізує просторові координати текстових блоків у вигляді обмежувальних рамок (англ. *bounding boxes*) у форматі (*[ymin, xmin, ymax, xmax]*), що мінімізує ризик спотворення тексту через нестандартне

художнє оформлення коміксу. Сама система дає вибір моделей OCR, які можна взяти за основу для процесу розпізнавання, в нашому випадку використовується модель «*EasyOCR*» (JaidedAI, 2020). Окрім цього, для алгоритму сегментації система потребує доступу до зовнішніх мультимодальних моделей за допомогою ключа прикладного програмного інтерфейсу (англ. *API key*). У межах цієї роботи для проведення сегментації було використано «*Gemini API*» (Google, 2026). Після закінчення процесу розпізнавання на виході система видає: *координати текстових блоків*, *текст*, *номер панелі*, *номер бульки, тип тексту* та його *стиль* (див. *Рис. 3.1*). Далі з цих даних власне текст передається моделі перекладу, а координати зберігаються для майбутньої задачі відображення перекладу поверх сторінки.


```
{'text_box': [1521, 1502, 1634, 1789], 'text': '*YEP! IT JUST HAPPENED IN CLONE CONSPIRACY #1 OUT NOW!
--NICK.', 'panel_id': '1', 'text_bubble_id': '1-6', 'type': 'narration', 'style': 'normal'}
{'text_box': [1815, 1467, 1925, 1735], 'text': 'I AM NO MERE CLONE, SIMULACRUM, OR HOLOGRAM.', 'panel_id': '2',
'text_bubble_id': '2-1', 'type': 'dialogue', 'style': 'normal'}
{'text_box': [1952, 1564, 2065, 1749], 'text': 'IT IS I. THE ONE TRUE OTTO OCTAVIUS.', 'panel_id': '2',
'text_bubble_id': '2-2', 'type': 'dialogue', 'style': 'normal'}
{'text_box': [2603, 1433, 2722, 1648], 'text': 'THE MAN WHO CHEATED DEATH ITSELF!', 'panel_id': '3',
'text_bubble_id': '3-1', 'type': 'dialogue', 'style': 'normal'}
```


Рис. 3.1. Приклад виведення результату розпізнавання моделі «ComiQ».

2. Процес перекладу реалізовано в межах єдиного серверного середовища, щоб уникнути додаткових затримок під час обміну даними між модулями. Для перекладу використовуються ваги навчаного LoRA-адаптера, які накладаються поверх базової архітектури моделі «*Mistral-7B-v0.1*». Отриманий з системи OCR текст необхідно попередньо нормалізувати. Оскільки оригінальні репліки часто записуються суцільними великими літерами, варто примусово перевести весь текст у нижній регістр. Це оптимізує роботу токенізатора і запобігає появі контекстуальних галюцинацій при генерації перекладу. Далі виконується стандартний процес пореченнєвого перекладу за інструкційним шаблоном моделі «*Dragoman*». По закінченню локальний клієнт отримує перекладені речення і виводить їх у текстове вікно в графічному інтерфейсі (див. *Рис. 3.2*).

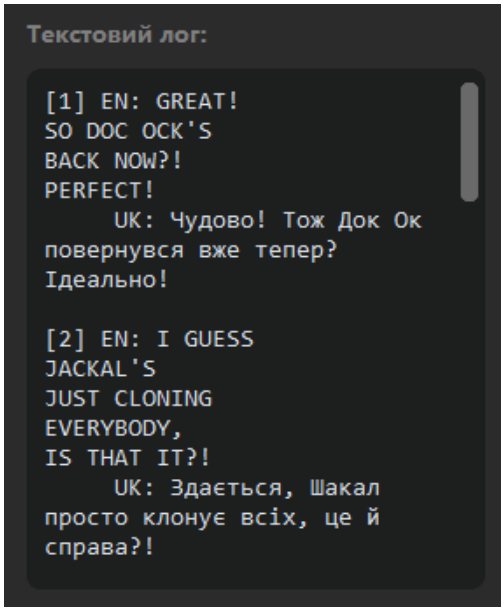

Рис. 3.2. Приклад виведення перекладу в графічному інтерфейсі.

3. Фінальним етапом роботи пайплайну є накладання перекладених реплік поверх сторінки коміксу. Після первинного завантаження файлу в графічний інтерфейс користувачем, оригінальне зображення сторінки виводиться на центральне полотно та автоматично масштабується, а після завершення процесу перекладу користувач має можливість увімкнути режим відображення перекладу в боковій панелі керування застосунку. Перекладений текст розташовується в обмежувальних рамках за координатами, які попередньо надала система OCR, а регістр тексту змінюється на верхній, щоб відповідати формату коміксів (див. *Рис. 3.3* та *Рис. 3.4*).

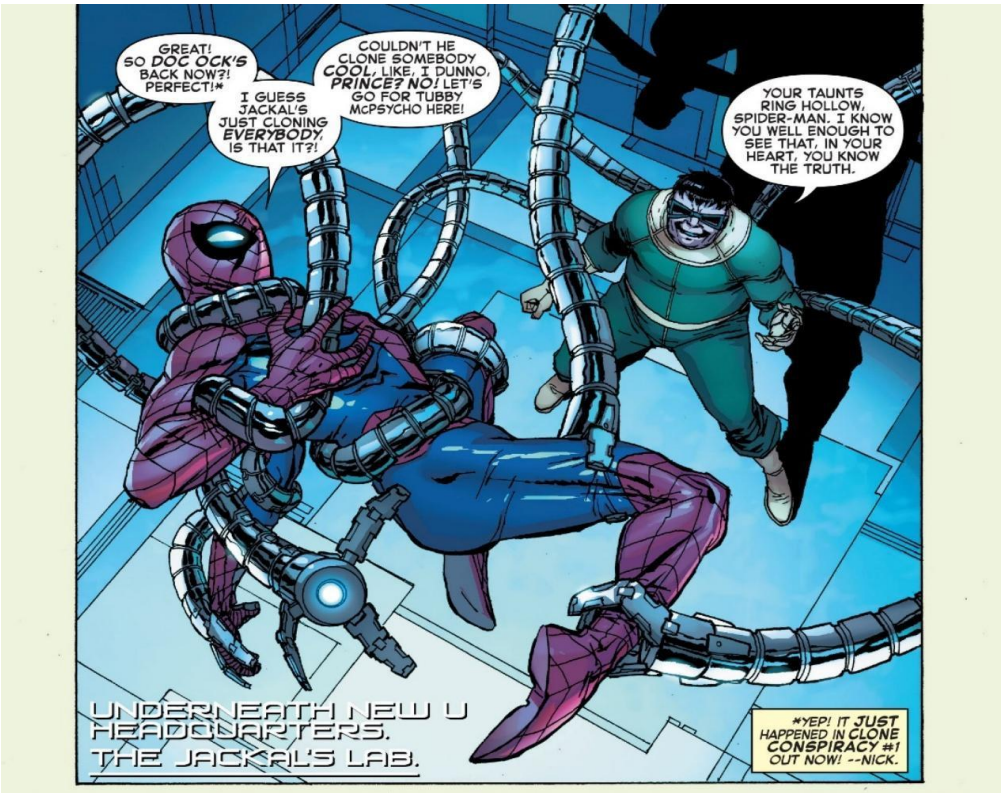

Рис. 3.3. Оригінальний фрагмент сторінки коміксу.

Рис. 3.4. Фрагмент сторінки коміксу з увімкненим накладанням перекладу.

Повне виконання пайплайну для обробки однієї сторінки коміксу займає 1.5-2 хвилини, залежно від обсягу тексту на сторінці.

Реалізація програмного застосунку, що інтегрує всі компоненти пайплайну в єдину систему з графічним інтерфейсом користувача, написана мовою програмування *Python* з використанням бібліотеки *CustomTkinter* (Schimansky, 2023) – сучасного розширення стандартної бібліотеки *Tkinter*, що має більш гнучку можливість налаштування дизайну інтерфейсу. Файли застосунку розміщені у відкритому репозиторії на платформі *GitHub* (див. *Додаток Ґ*).

Файлова структура застосунку виглядає таким чином:

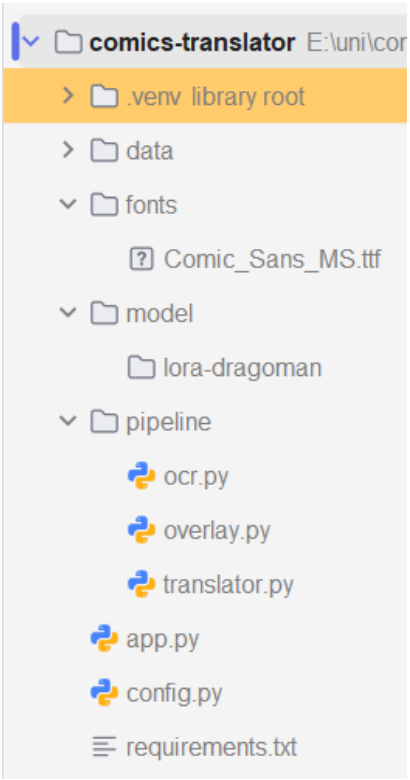

Рис. 3.5. Структура застосунку.

У теці *pipeline* знаходяться головні модулі для реалізації описаного пайплайну. Модуль *pipeline/ocr.py* посилає запит на сервер для виконання процесу розпізнавання тексту та витягування текстових фрагментів із зображення сторінки коміксу разом із їхніми координатами. Модуль *pipeline/translator.py* відправляє запит на виконання перекладу тексту отриманого з системи OCR і повернення

паралельних текстів розділених за блоками або бульками. І відповідно модуль *pipeline/overlay.py* виконує функції реалізації накладання отриманого перекладу поверх сторінки коміксу. Головний модуль *app.py* інтегрує всі компоненти та реалізує графічний інтерфейс. Конфігураційний файл config.py зберігає всі основні налаштування: шляхи до моделей, ключі доступу до зовнішніх сервісів та параметри відображення. Це забезпечує зручне налаштування застосунку без зміни основного коду. Також присутні допоміжні теки: *font* – зберігає файл шрифту *Comic Sans*, який застосовується при накладанні; *model* – локальне розташування ваг донавчаного адаптера; *data* – файли та зображення, які застосовані в графічному інтерфейсі. Файл *requirements.txt* містить список зовнішніх бібліотек та залежностей для коректної роботи скриптів.

Графічний інтерфейс застосунку поділений на дві основні панелі (*Рис. 3.6*). Ліва панель містить усі елементи керування: кнопка завантаження зображення сторінки коміксу з підтримкою різних форматів (*jpg*; *jpeg*; *png*; *webp*), кнопка запуску системи OCR, кнопка запуску перекладу, перемикач відображення оверлею з перекладом, кнопка збереження результату та кнопка скидання усіх дій. Нижче розташоване текстове поле, в якому виводяться проміжні результати, отримані з моделей. Кнопка-перемикач справа зверху дозволяє користувачеві перейти в окрему вкладку, де розміщені різні налаштування, як посилання на сервер. Права частина інтерфейсу займає основне полотно для відображення завантаженої сторінки коміксу. Після виконання перекладу поверх зображення накладаються напівпрозорі бульки з українським перекладом, розташовані відповідно до координат, отриманих від модуля OCR. Перемикач «Показати переклад» дозволяє миттєво вмикати та вимикати відображення накладання, щоб була можливість порівняти оригінал з перекладом без повторного запуску пайплайну. Саме зображення можна масштабувати і прокручувати. Допоміжний рядок у нижній частині вікна інформує користувача про поточний стан виконання операцій.

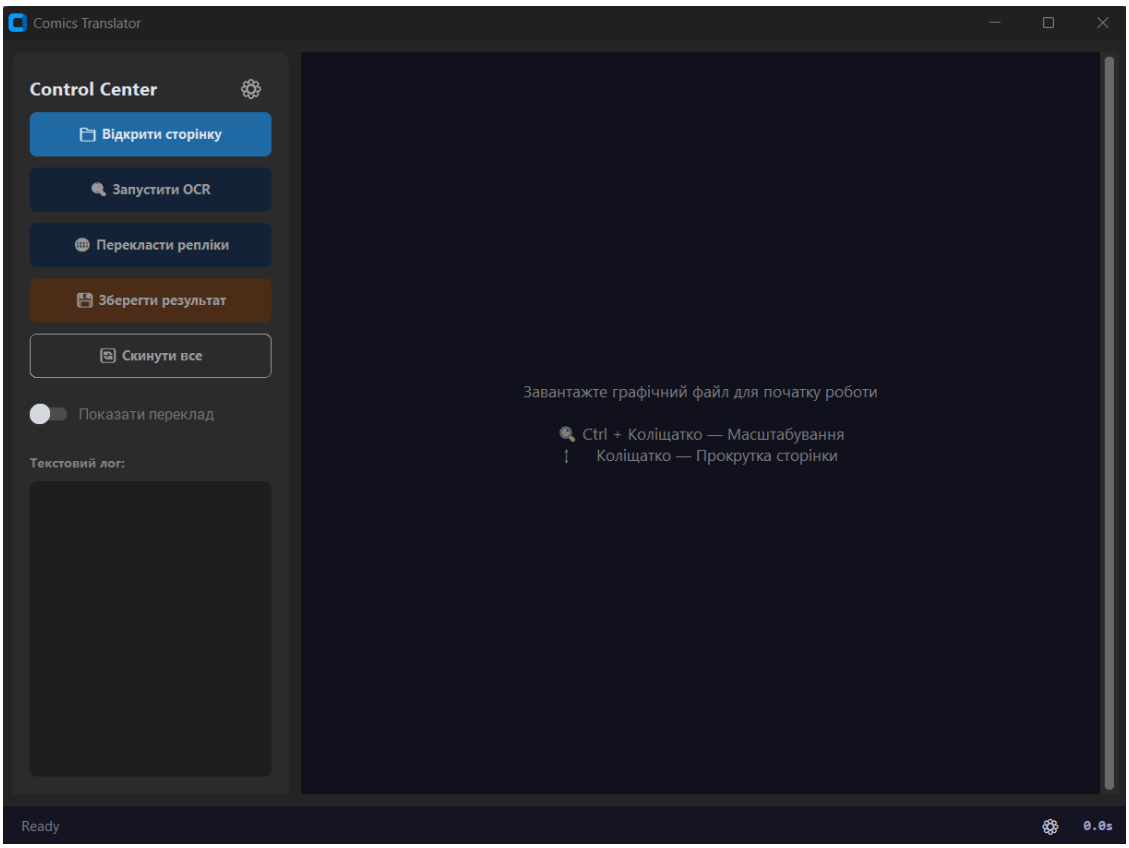

Рис. 3.6. Повний вигляд графічного інтерфейсу.

ВИСНОВКИ

Дослідження, проведене у цій роботі, ставило за мету вирішення актуальної задачі вузькоспеціалізованої адаптації нейромережевої моделі машинного перекладу для відтворення текстів англомовних коміксів українською мовою. У ході написання дипломної роботи було успішно досягнуто всіх поставлених цілей та розв'язано всі завдання. Це передбачало як теоретичне осмислення специфіки коміксів та труднощів доменної адаптації, так і практичне виконання всіх етапів донавчання моделі, оцінювання отриманих результатів та реалізації пайплайну.

У теоретичній частині роботи, з урахуванням фундаментальних праць С. МакКлауда, Н. Кона, Р. Данкана та інших, було розглянуто комікс як мультимодальний жанр із характерною системою текстових елементів, таких як: мовні бульки, текстові блоки, звуконаслідування тощо. Встановлено, що кожен із цих елементів функціонує в тісній взаємодії з візуальним контекстом і формує унікальне мовне середовище, яке вимагає особливого підходу до перекладу. Проаналізовано еволюцію підходів до машинного перекладу від систем на основі правил до сучасних архітектур *Transformers*, зокрема розглянуто відмінності між підходами encoder-decoder та decoder-only. Детально охарактеризовано модель *«Dragoman»* як першу публічно доступну decoder-only модель машинного перекладу з англійської на українську мову, розроблену українськими дослідниками, та обґрунтовано доцільність її використання як основи для подальшого донавчання.

У практичній частині роботи для навчання моделі було сформовано спеціалізований паралельний корпус обсягом 2000 рядків на основі відкритого набору даних *«COMICS TEXT+»* з використанням стратифікованої вибірки та автоматичного перекладу моделлю *«Google Gemini Flash 3»*. Для запобігання явищу катастрофічного забування корпус було доповнено 1000 рядками загальної тематики з набору даних *«Extended Multi30k-Uk»* та оформлено відповідно до інструкційного шаблону моделі *«Dragoman»*. Для оцінювання якості перекладу

сформовано окремий тестовий корпус на основі офіційного українського перекладу коміксу *«Amazing Spider-Man: The Clone Conspiracy»*, обсягом 231 рядок, попередньо нормалізований засобами автоматичної обробки природної мови. Процес донавчання моделі *«Dragoman»* було реалізовано в середовищі *Google Colaboratory* із застосуванням 4-бітного квантування та LoRA-адаптації упродовж трьох епох. У ході навчання спостерігалося стабільне зниження функції втрат з 3.53 до 1.02, а точність передбачення токенів зросла з 50.6% до 78.9%, що свідчить про ефективну адаптацію моделі до специфіки коміксного тексту без ознак перенавчання.

Комплексне оцінювання якості перекладу здійснено за чотирма автоматичними метриками (*BLEU*, *TER*, *chrF*, *COMET*). Хоча обсяг тестової вибірки був дещо замалий для класичних метрик, вони підтвердили стійку перевагу донавченої моделі над базовими системами за всіма показниками. Зокрема, за метрикою *BLEU* можна спостерігати відносне покращення на 15%. Приріст за нейронною метрикою *COMET* становив з 0.797 до 0.815. Оскільки метрика *COMET* оцінює семантичну еквівалентність перекладу, отримані результати свідчать про те, що донавчання дозволило моделі краще відтворювати змістове наповнення реплік, характерних для жанру коміксів. Це також підтвердив лінгвістичний аналіз результатів перекладу, який засвідчив, що модель суттєво покращила відтворення ідіоматичних та фразеологічних виразів, коротких розмовних реплік, а також граматично коректне використання кличного відмінка у звертаннях. Натомість було виявлено дві стійкі проблеми, які потребують подальшого вирішення: нестабільне відтворення звуконаслідування та некоректна обробка навмисно розірваних реплік.

Для підтвердження практичної застосовності розробленого рішення було реалізовано наскрізний пайплайн обробки сторінок коміксів, що інтегрує систему оптичного розпізнавання тексту *«ComiQ»* та донавчану модель перекладу в рамках єдиної клієнт-серверної архітектури з графічним інтерфейсом користувача.

Розроблений застосунок забезпечує автоматичне розпізнавання, переклад та візуальне накладання перекладеного тексту поверх оригінального зображення сторінки, що демонструє реальну придатність системи до практичного застосування в умовах локалізації коміксів.

Таким чином, наукова новизна кваліфікаційної роботи полягає в тому, що було вперше реалізовано вузькоспеціалізоване донавчання україномовної нейромережевої decoder-only моделі машинного перекладу на цільовому корпусі текстів коміксів, а також проведено порівняльну оцінку якості локалізації шляхом поєднання автоматичних метрик та лінгвістичного аналізу типових помилок. Водночас рекомендації щодо подальшого дослідження цієї теми полягають у розширенні обсягу навчального корпусу та його збагаченні прикладами звуконаслідування та розірваних реплік для подолання виявлених обмежень моделі, а також у вивченні ефективності альтернативних методів доменної адаптації та верифікації архітектури на ширшому наборі графічних видань задля об’єктивного оцінювання її загальної узагальнювальної здатності.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Бойко, О., (2024). *“УІК: Укрліт видають меншими накладами, ніж книжки іноземних авторів”, Читомо* [онлайн]. Режим доступу: https://chytomo.com/uik-ukrlit-vydaiut-menshymy-nakladamy-nizh-knyzhky-inozemnykh-avtoriv/.
2. Данкан, Р., Сміт, М. та Левіц, П., (2020). *Сила Коміксів*. Переклад з англ. Д. Скорбатюк. Київ: ArtHuss.
3. Івасишин, М. Р., (2019). *Мультимодальність англомовного коміксу: лінгвальний і екстралінгвальний виміри*. Автореферат дис. канд. філол. наук. Львівський національний університет імені Івана Франка, Львів.
4. Макклауд, С., (2019). *Зрозуміти Комікси*. Переклад з англ. Я. Стріха. Київ: РІДНА МОВА.
5. Мильченко, Л. та Татарінова, Л., (2020). *Особливості сприйняття візуальної книги. Комікси. Манга. Графічний роман*. Вісник Книжкової палати, (12), с. 10-15.
6. Рябова, К. О., (2023). *Історія розвитку систем машинного перекладу (від ранніх етапів до початку XXI століття)*. Вісник Київського національного лінгвістичного університету. Серія Філологія, 26(2), с. 101-111.
7. Шевченко, І. К., (2024). *Аналіз особливостей перекладу сучасних коміксів*. [Неопублікована курсова робота]. Київський національний університет імені Тараса Шевченка, Київ.
8. Шевченко, І. К., (2025). *Оцінка якості машинного перекладу текстів коміксу*. [Неопублікована курсова робота]. Київський національний університет імені Тараса Шевченка, Київ.
9. Шеремет, Є. А., (2021). *Структурно-семантичні та функційні особливості ономатопеї в англомовних коміксах*. Магістерська робота. Запорізький національний університет, Запоріжжя. Режим доступу: https://dspace.znu.edu.ua/jspui/handle/12345/6438.

10. Bañón, M., Chen, P., Haddow, B., Heafield, K., Hoang, H., Esplà-Gomis, M., Forcada, M.L., Kamran, A., Kirefu, F., Koehn, P., Ortiz-Rojas, S., Pla, L., Ramírez-Sánchez, G., Sarrías, E., Strelec, M., Thompson, B., Waites, W., Wiggins, D. and Zaragoza, J., (2020). *ParaCrawl: Web-Scale Acquisition of Parallel Corpora*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4555-4567).
11. Chéragui, M. A., (2012). *Theoretical Overview of Machine Translation*. ICWIT, pp.160-169.
12. Cohn, N., (2013). *The Visual Language of Comics: Introduction to the Structure and Cognition of Sequential Images*. London: Bloomsbury Academic.
13. Dettmers, T., Pagnoni, A., Holtzman, A. and Zettlemoyer, L., (2023). *QLoRA: Efficient finetuning of quantized LLMs*. Advances in Neural Information Processing Systems, 36, pp.10088-10115.
14. Dong, X., (2026). *Recent Progress of Neural Machine Translation*. In Proceedings of the 5th International Conference on Computer, Artificial Intelligence and Control Engineering (pp. 12-17).
15. Feng, F., Yang, Y., Cer, D., Arivazhagan, N. and Wang, W., (2022). *Language-agnostic BERT Sentence Embeddings*. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 878-891).
16. Fireclaw Ukraine, (2026). [online] Available at: https://fireclaw.com.ua/.
17. Forceville, C., Veale, T., & Feyaerts, K. (2010). *Balloonics: The visuals of balloons in comics*. In J. Goggin & D. Hassler-Forest (Eds.), *The rise and reason of comics and graphic literature: Critical essays on the form* (pp. 56-73). Jefferson, NC: McFarland.
18. Fu, Z., Lam, W., Yu, Q., So, A.M.C., Hu, S., Liu, Z. and Collier, N., (2023). *Decoder-Only or Encoder-Decoder? Interpreting Language Model as a Regularized Encoder-Decoder*. arXiv preprint arXiv:2304.04052.

19. Fury, D., (2021). *TrueCase: A python true casing utility that restores case information for texts* [online]. GitHub. Available at: https://github.com/daltonfury42/truecase.
20. Gage, C. and Slott, D., (2017). *Amazing Spider-Man: The Clone Conspiracy*. New York: Marvel Enterprises.
21. Google, (2026). *Google AI Studio* [online]. Available at: https://aistudio.google.com.
22. Google, (2026). *Google Colaboratory* [online]. Available at: https://colab.research.google.com.
23. Google, (2026). *Google Gemini* [online]. Available at: https://gemini.google.com/app.
24. Goyal, N., Gao, C., Chaudhary, V., Chen, P. J., Wenzek, G., Daudin, F., Khassanov, S., Peng, J., Popuri, A., Goyal, R. and Zhou, S., (2022). *Flores-101: A multilingual machine translation evaluation benchmark*. Transactions of the Association for Computational Linguistics, 10, pp. 522-538.
25. Groensteen, T., (2007). *System of Comics*. Jackson: University Press of Mississippi.
26. Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A., (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. [online] Available at: https://doi.org/10.5281/zenodo.1212303.
27. Hu, E. J., Shen, Y., Wallis, P., Allen-Zellmer, J., Yuan, L., Wang, Q., Hu, C. and Chen, W., (2022). *LoRA: Low-rank adaptation of large language models*. In International Conference on Learning Representations (ICLR).
28. JaidedAI, (2020). *EasyOCR: Ready-to-use OCR with 80+ supported languages and all popular writing scripts including: Latin, Chinese, Arabic, Devanagari, Cyrillic, etc*. [online]. GitHub. Available at: https://github.com/jaidedai/easyocr.
29. Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Devwial, D., de las Casas, D., Florian, B., Koppula, R., Stock, P., Scao, T. L. and Lavril, T., (2023). *Mistral 7B*. arXiv preprint arXiv:2310.06825.
30. Kanishq, V., (2026). *ComiQ: Comic-Focused Hybrid OCR Library* [online]. GitHub. Available at: https://github.com/StoneSteel27/ComiQ.

31. Lhoest, Q., Del Moral, A.V., Jernite, Y., Thakur, A., Von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L. and Davison, J., (2021). *Datasets: A community library for natural language processing*. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 175-184).
32. Manga Media, (2023). *Чим відрізняються манга, манхва та маньхва?* [онлайн]. Режим доступу: https://manga-media.com.ua/articles/manga-manhva-manhua.
33. Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S. and Bossan, B., (2022). *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods* [online]. GitHub. Available at: https://github.com/huggingface/peft.
34. Ngrok, (2026) *ngrok: AI & API Gateway | Secure Tunnels & Traffic* [online]. Available at: https://ngrok.com/.
35. Paniv, Y., Chaplynskyi, D., Trynus, N. and Kyrylov, V., (2024). *Setting up the data printer with improved English to Ukrainian machine translation*. In Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024 (pp. 41-50). ELRA and ICCL, Torino, Italia.
36. Papineni, K., Roukos, S., Ward, T. and Zhu, W. J., (2002). *Bleu: a Method for Automatic Evaluation of Machine Translation*. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311-318).
37. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. and Chintala, S., (2019). *PyTorch: An imperative style, high-performance deep learning library*. Advances in Neural Information Processing Systems, 32.
38. Pellitteri, M., (2024). "*Lettering (comics)"*, *EBSCO Information Services, Inc.* [online]. Available at: https://www.ebsco.com/research-starters/literature-and-writing/lettering-comics.

39. Popović, M., (2015). *chrF: character n-gram F-score for automatic MT evaluation*. In Proceedings of the Tenth Workshop on Statistical Machine Translation (pp. 392-395).
40. Post, M., (2026). *sacreBLEU: Reference BLEU implementation that auto-downloads test sets and reports a version string to facilitate cross-lab comparisons* [online]. GitHub. Available at: https://github.com/mjpost/sacrebleu.
41. Rei, R., Stewart, C., Farinha, A. C. and Lavie, A., (2020). *COMET: A Neural Framework for MT Evaluation*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 2685-2702).
42. Rei, R., Stewart, C., Farinha, A. C. and Lavie, A., (2025). *COMET: A Neural Framework for MT Evaluation* [online]. GitHub. Available at: https://github.com/Unbabel/COMET.
43. Romanyshyn, M., Romanyshyn, N., Hlybovets, A. and Ignatenko, O., (2024). *Proceedings of the 3rd Workshop on Ukrainian Natural Language Processing (UNLP 2024) @ LREC-COLING 2024*. ELRA and ICCL, Torino, Italia.
44. Saichyshyna, N., Maksymenko, D., Turuta, O., Yerokhin, A., Babii, A. and Turuta, O., (2023). *Extension Multi30K: Multimodal Dataset for Integrated Vision and Language Research in Ukrainian*. In Proceedings of the Second Ukrainian Natural Language Processing Workshop (UNLP) (pp. 54-61).
45. Saunders, D., (2022). *Domain Adaptation and Multi-Domain Adaptation for Neural Machine Translation: A Survey*. Journal of Artificial Intelligence Research, 75, pp.351-424.
46. Schimansky, T., (2023). *CustomTkinter: A modern customization of the standard Tkinter library for Python* [online]. GitHub. Available at: https://github.com/TomSchimansky/CustomTkinter.
47. Snover, M., Dorr, B., Schwartz, R., Micciulla, L. and Makhoul, J., (2006). *A Study of Translation Edit Rate with Targeted Human Annotation*. In Proceedings of the 7th

Conference of the Association for Machine Translation in the Americas: Technical Papers (pp. 223-231).

48. Soykan, G., Yuret, D. and Sezgin, T. M., (2022). *COMICS Text+: A Comprehensive Gold Standard and Benchmark for Comics Text Detection and Recognition* [online]. GitHub. Available at: https://github.com/gsoykan/comics_text_plus.

49. Tabsharani, F., (2024). *"What is Machine Translation?", TechTarget Inc.* [online]. Available at: https://www.techtarget.com/searchenterpriseai/definition/machine-translation.

50. Tiedemann, J. and Thottingal, S., (2020). *OPUS-MT – Building open translation services for the World*. In Proceedings of the 22nd Annual Conference of the 22nd Annual Conference of the European Association for Machine Translation (pp. 479-480).

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

52. von Werra, L., Belkada, Y., Tunstall, L., Beeching, E., Thrush, T., Lambert, N., Huang, S., Rasul, K. and Gallouédec, Q., (2020). *TRL: Transformer Reinforcement Learning* [online]. GitHub. Available at: https://github.com/huggingface/trl.

53. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q. and Rush, A. M., (2020). *Transformers: State-of-the-Art Natural Language Processing*. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).

ДОДАТКИ

Додаток А. Скрипти підготовки навчального та тестового корпусів.

https://colab.research.google.com/drive/1yv4rNSvKuQ4lMGjX11SwGlCEuVRMeKYz?usp=sharing
За вказаним посилання розміщено набір скриптів, які реалізують усі етапи підготовки паралельного корпусу для донавчання та тестування моделі, а саме: формування стратифікованої вибірки за типом і довжиною тексту; відібраних пар у формат інструкцій; змішування доменного і загального корпусів; лінгвістична нормалізація англійського та українського текстів засобами автоматичної обробки природної мови.

Додаток Б. Скрипт донавчання моделі.

https://colab.research.google.com/drive/1D19MAl0ZZz9ecm8GeMmn9JE5zxCYppVB?usp=sharing
За вказаним посилання знаходиться скрипт, який реалізує повний процес донавчання моделі у середовищі Google Colaboratory. Він завантажує необхідні моделі та адаптори з моделі платформи Hugging Face Hub; підготовлює моделі до навчання в режимі PEFT; виконує конфігурацію гіперпараметрів і запуск процесу донавчання; зберігає проміжні та фінальний результати у на хмарному сховищі Google Drive.

Додаток В. Репозиторій донавченої моделі на платформі Hugging Face Hub.

https://huggingface.co/yngillia/comics-translator
За вказаним посилання розміщено публічний репозиторій, у якому зберігаються ваги фінального LoRA-адаптера. Репозиторій містить конфігураційні файли адаптера та може бути використаний для відтворення результатів дослідження або подальшого донавчання.

Додаток Г. Скрипти автоматичного оцінювання якості перекладу.

https://colab.research.google.com/drive/1lM4mJaiteaj6sXf5QvmtuQY63UGjjGL_?usp=ring
За вказаним посиланням розміщено набір скриптів, які реалізують процедуру автоматичного оцінювання, а саме: генерацію перекладів тестового корпусу трьома досліджуваними системами; обчислення класичних метрик BLEU, TER і chrF засобами бібліотеки sacreBLEU; обчислення нейронної метрики COMET із застосуванням спеціального обхідного рішення, що усуває конфлікти залежностей між версіями бібліотек, необхідних для запуску метрики, та середовищем Google Colaboratory.

Додаток Ґ. Репозиторій застосунку на платформі GitHub.

https://github.com/yngillia/comics-translator
За вказаним посилання розміщено публічний репозиторій із повним вихідним кодом застосунку. Репозиторій містить: головний модуль графічного інтерфейсу користувача; модулі клієнтського пайплайну для взаємодії із

серверним компонентом; посилання на серверне середовище Google Colaboratory із розгорнутими моделями OCR та машинного перекладу; допоміжні файли графічного інтерфейсу; виконуваний файл для спрощеного встановлення застосунку.