ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ
НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ
ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ
КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

КВАЛІФІКАЦІЙНА РОБОТА

на тему: «Розробка методики оцінювання якості та оптимізації систем автоматичного перекладу і локалізації вебконтенту на основі метрик штучного інтелекту»

на здобуття освітнього ступеня магістра
зі спеціальності 121 Інженерія програмного забезпечення
освітньо-професійної програми «Інженерія програмного забезпечення»

Кваліфікаційна робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

_____ Кирило СТАСЮК
(підпис)

Виконав: здобувач вищої освіти групи ПДМ-62
Кирило СТАСЮК

Керівник: Владислав ЯСКЕВИЧ
канд. техн. наук, доцент

Рецензент: _____
науковий ступінь, вчене звання *Ім'я, ПРІЗВИЩЕ*

Київ 2026

ДЕРЖАВНИЙ УНІВЕРСИТЕТ
ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

Навчально-науковий інститут інформаційних технологій

Кафедра Інженерії програмного забезпечення

Ступінь вищої освіти Магістр

Спеціальність 121 Інженерія програмного забезпечення

Освітньо-професійна програма «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
Інженерії програмного забезпечення
_____ Ірина ЗАМРІЙ
«_____» _____ 2025 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Стасюку Кирилу Сергійовичу

1. Тема кваліфікаційної роботи: «Розробка методики оцінювання якості та оптимізації систем автоматичного перекладу і локалізації вебконтенту на основі метрик штучного інтелекту»

керівник кваліфікаційної роботи Владислав ЯСКЕВИЧ, канд. техн. наук, доцент

затверджені наказом Державного університету інформаційно-комунікаційних технологій від «30» жовтня 2025 р. № 467.

2. Строк подання кваліфікаційної роботи «19» грудня 2025 р.

3. Вихідні дані до кваліфікаційної роботи: науково-технічна література, параметри вебконтенту, методи оцінювання якості перекладу, вимоги до збереження структурної цілісності вебсторінок.

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

1. Аналіз існуючих рішень для автоматичного перекладу та локалізації вебконтенту з використанням штучного інтелекту.

2. Аналіз методів та метрик оцінювання якості автоматичного перекладу і локалізації вебконтенту.
3. Розробка програмного модуля для оцінювання якості перекладу та перевірки структурної цілісності вебконтенту.

5. Перелік ілюстративного матеріалу: *презентація*

1. Порівняльна характеристика існуючих методів оцінювання якості автоматичного перекладу.
2. Математична модель перевірки структурної цілісності.
3. Алгоритм перевірки структурної цілісності.
4. Алгоритм оцінки якості перекладу на основі openai api.
5. Комбінована формула інтегральної оцінки (final score).
6. Результати експериментальних досліджень.
7. Порівняльна діаграма результатів оцінювання систем перекладу.

6. Дата видачі завдання «31» жовтня 2026 р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів кваліфікаційної роботи	Строк виконання етапів роботи	Примітка
1	Аналіз наявної науково-технічної літератури	31.10 - 03.11.2025	
2	Вивчення сучасних методів автоматичного перекладу, трансформерних моделей та локалізації вебконтенту	03.11 - 05.11.2025	
3	Аналіз існуючих метрик якості перекладу та людських методів оцінювання	05.11 - 08.11.2025	
4	Дослідження технічних аспектів збереження структури вебсторінок	08.11 - 11.11.2025	
5	Розроблення методології комплексного оцінювання	11.11 - 13.11.2025	
6	Реалізація програмного модуля оцінювання	13.11 - 20.11.2025	
7	Проведення експериментів перекладу	21.11 - 22.11.2025	
8	Оформлення роботи: вступ, висновки, реферат	23.11 - 24.11.2025	
9	Розробка демонстраційних матеріалів	24.11 - 25.11.2025	
10	Попередній захист роботи	17.11-01.12.2025	

Здобувач вищої освіти _____ Кирило Стасюк
(підпис)

Керівник
кваліфікаційної роботи _____ Владислав ЯСКЕВИЧ
(підпис)

РЕФЕРАТ

Текстова частина кваліфікаційної роботи на здобуття освітнього ступеня магістра: 82 стор., 3 табл., 7 рис., 30 джерел.

Мета роботи - розробити та обґрунтувати метод оцінювання ефективності систем штучного інтелекту для автоматичного перекладу та локалізації вебконтенту, що дозволяє виявляти сильні та слабкі сторони існуючих рішень та визначати напрями їх удосконалення з урахуванням специфіки вебсередовища.

Об'єкт дослідження - процеси автоматичного перекладу та локалізації вебконтенту.

Предмет дослідження - метод оцінювання та вдосконалення якості перекладу й локалізації вебконтенту за допомогою систем штучного інтелекту.

У роботі використано сучасні методи аналізу якості перекладу, інструменти обробки HTML-структур, технології машинного навчання та великі мовні моделі (LLM). Застосовано апарат математичної статистики, методи лінгвістичного аналізу та об'єктно-орієнтоване програмування для побудови програмного модуля оцінювання.

Проведено аналіз сучасних підходів до автоматичного перекладу, визначено їх переваги й недоліки, а також досліджено метрики BLEU, METEOR, chrF, COMET, BLEURT та методи людського оцінювання MQM, LQA і HTER. Особливу увагу приділено проблемам збереження структури вебсторінок під час перекладу, що є критично важливим для працездатності вебресурсів.

Розроблено та реалізовано програмний модуль, який виконує комплексну оцінку перекладу за двома напрямами: лінгвістичної якості (на основі OpenAI GPT) та структурної цілісності HTML (на основі JSDOM). Запропоновано комбіновану метрику Final Score, яка дозволяє об'єктивно порівнювати різні системи перекладу.

Проведено експериментальні дослідження перекладів, отриманих за допомогою Google Translate, DeepL та OpenAI GPT. Виконано порівняльний аналіз їх точності, стилістичної узгодженості, термінологічної правильності та якості

збереження структури HTML. Результати підтвердили ефективність запропонованої методики та її здатність виявляти сильні й слабкі сторони різних систем машинного перекладу.

Запропонована методика є перспективною для інтеграції у системи автоматичної локалізації вебпроєктів, забезпечуючи підвищення точності і стабільності перекладу.

КЛЮЧОВІ СЛОВА: АВТОМАТИЧНИЙ ПЕРЕКЛАД, ЛОКАЛІЗАЦІЯ, ШТУЧНИЙ ІНТЕЛЕКТ, BLEU, COMET, GPT, JSDOM, HTML-СТРУКТУРА, ОЦІНЮВАННЯ ЯКОСТІ.

ABSTRACT

Text part of the master's qualification work: 82 pages, 7 pictures, 3 tables, 30 sources.

The purpose of the work is to develop and substantiate a methodology for evaluating the effectiveness of artificial intelligence systems used for automatic translation and localization of web content. This methodology makes it possible to identify the strengths and weaknesses of existing solutions and determine directions for their improvement, taking into account the specific characteristics of the web environment.

The object of research is artificial intelligence systems for automatic translation and localization of web content.

The subject of research is the methodology for assessing and improving the quality of translation and localization of web content using artificial intelligence systems.

The work employs modern methods of translation quality assessment, tools for analyzing HTML structures, machine learning technologies, and large language models (LLMs). Mathematical statistics, linguistic analysis techniques, and object-oriented programming are applied to develop a comprehensive evaluation module.

An extensive analysis of current approaches to automatic translation has been conducted, including their advantages and limitations. The study examines key metrics such as BLEU, METEOR, chrF, COMET, and BLEURT, as well as human evaluation methods including MQM, LQA, and HTER. Special attention is given to the challenges of preserving HTML structure during translation, which is crucial for the correct functioning of web resources.

A software module has been designed and implemented to perform a combined evaluation of translations in two dimensions: linguistic quality (based on OpenAI GPT) and structural integrity of HTML content (using JSDOM). A combined metric, Final Score, is proposed to enable an objective comparison of various translation systems.

Experimental studies were conducted using translations generated by Google Translate, DeepL, and OpenAI GPT. A comparative analysis was performed to assess

accuracy, stylistic consistency, terminology preservation, and structural fidelity of the resulting HTML. The results confirm the effectiveness of the proposed methodology and its ability to identify strengths and weaknesses across different machine translation systems.

The proposed methodology shows strong potential for integration into automated web localization pipelines, significantly improving translation accuracy, stability, and overall quality.

KEYWORDS: AUTOMATIC TRANSLATION, LOCALIZATION, ARTIFICIAL INTELLIGENCE, BLEU, COMET, GPT, JSDOM, HTML STRUCTURE, QUALITY EVALUATION.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

API - Інтерфейс прикладного програмування (Application Programming Interface).

BLEU - Двомовна оцінка перекладу (Bilingual Evaluation Understudy).

BLEURT - Навчена метрика оцінки перекладу на основі BERT (BERT-based Learned Evaluation Metric).

CMS - Система керування контентом (Content Management System).

COMET - Метрика оцінки перекладу, що використовує багатомовні моделі (Cross-lingual Optimized Metric for Evaluation of Translation).

DOM - Об’єктна модель документа (Document Object Model).

HTML - Мова розмітки гіпертексту (HyperText Markup Language).

HTER - Редагування людиною після перекладу (Human Translation Edit Rate).

JSON - Формат обміну даними JavaScript Object Notation.

JSDOM - JavaScript-емуляція DOM-середовища у Node.js.

LLM - Велика мовна модель (Large Language Model).

LQA - Лінгвістична оцінка якості (Linguistic Quality Assurance).

MQM - Багатовимірна оцінка якості перекладу (Multidimensional Quality Metrics).

MT - Машинний переклад (Machine Translation).

NLP - Обробка природної мови (Natural Language Processing).

NMT - Нейронний машинний переклад (Neural Machine Translation).

REST - Архітектурний стиль передачі даних (Representational State Transfer).

ВСТУП

Стрімкий розвиток штучного інтелекту та глобальна цифровізація сприяли появі численних інструментів автоматичного перекладу, що активно використовуються у вебсередовищі. У сучасних умовах багато організацій прагнуть швидко адаптувати контент до потреб міжнародної аудиторії, що висуває підвищені вимоги до точності, стилістичної узгодженості та технічної коректності перекладу. Автоматичні системи перекладу на основі нейронних моделей стали основою більшості сучасних сервісів, однак попри значний прогрес, їх робота ще не завжди забезпечує достатню якість локалізації складного або динамічного вебконтенту.

Особливу актуальність набуває проблема оцінювання якості перекладу. Традиційні метрики, які використовуються протягом багатьох років, здебільшого вимірюють формальну схожість тексту з еталонним перекладом, але не враховують лінгвістичну природність, стилістичну відповідність чи здатність зберігати семантичні й культурні нюанси. Ще більш складним завданням є перевірка структурної цілісності HTML-документів, адже під час перекладу автоматичні системи можуть ненавмисно змінювати або пошкоджувати теги, що призводить до некоректного відображення вебсторінок.

У цій роботі розглянуто сучасний стан систем автоматичного перекладу та локалізації, проаналізовано їхні можливості й обмеження, а також запропоновано методику комплексної оцінки якості, яка об’єднує лінгвістичні метрики, експертний аналіз та технічну перевірку структури вебконтенту. У процесі дослідження розроблено програмний модуль, що дозволяє автоматизувати перевірку якості перекладу за допомогою моделей штучного інтелекту та алгоритмів аналізу HTML-документів.

Запропонована методика спрямована на підвищення точності й надійності процесу локалізації та є актуальною для вебплатформ, які працюють у багатомовному середовищі та потребують стабільної інтеграції систем автоматичного перекладу.

Мета роботи – розробити та обґрунтувати метод оцінювання ефективності систем штучного інтелекту для автоматичного перекладу та локалізації вебконтенту, що дозволяє виявляти сильні та слабкі сторони існуючих рішень та визначати напрями їх удосконалення з урахуванням специфіки вебсередовища.

Об’єкт дослідження – процеси автоматичного перекладу та локалізації вебконтенту.

Предмет дослідження – методи оцінювання та вдосконалення якості перекладу й локалізації вебконтенту за допомогою систем штучного інтелекту.

Для досягнення поставленої мети було сформульовано такі завдання:

1. Провести огляд наукових джерел і сучасних досліджень у сфері автоматичного перекладу, нейронних моделей і локалізації вебконтенту.
2. Виконати аналіз існуючих метрик якості перекладу та методів технічної перевірки HTML-структур.
3. Розробити архітектуру програмного модуля для комплексної оцінки якості перекладу та перевірки структурної цілісності.
4. Реалізувати алгоритми оцінювання лінгвістичної якості за допомогою моделей OpenAI та алгоритми порівняння HTML-структур на основі JSDOM.
5. Провести експериментальні дослідження якості перекладів, виконаних різними системами (Google Translate, DeepL, OpenAI), та порівняти їх результати.
6. Здійснити аналіз отриманих даних, визначити сильні й слабкі сторони розглянутих систем та сформулювати рекомендації щодо оптимізації процесу автоматичного перекладу і локалізації.

Вирішення зазначених завдань дозволяє комплексно оцінити ефективність сучасних систем автоматичного перекладу, визначити напрями їх поліпшення та обґрунтувати важливість застосування інтегрованих методів оцінювання у практичних вебпроєктах.

1 АНАЛІЗ ІСНУЮЧИХ РІШЕНЬ ДЛЯ АВТОМАТИЧНОГО ПЕРЕКЛАДУ ТА ЛОКАЛІЗАЦІЇ ВЕБКОНТЕНТУ З ВИКОРИСТАННЯМ ШТУЧНОГО ІНТЕЛЕКТУ

1.1 Існуючі підходи, інструменти та їх еволюція

Системи автоматичного перекладу вебконтенту за останні десятиліття зазнали глибокої трансформації, поступово переходячи від простих правил до складних моделей штучного інтелекту. Перші покоління машинного перекладу ґрунтувалися на правилобазованих підходах, де формування великих лінгвістичних баз та граматичних моделей виконувалося вручну. Такі системи були жорсткими, погано адаптувалися до різних мовних стилів і вимагали значних ресурсів для розширення. Пізніше на зміну їм прийшли статистичні моделі, що визначали перекладні відповідники на основі імовірнісного аналізу великих корпусів текстів. Цей підхід дозволив суттєво підвищити швидкість перекладу, але залишається обмеженим у здатності враховувати контекст та передавати стиль мовлення [25], [26].

Сучасний етап розвитку пов’язаний із нейронним машинним перекладом, який базується на глибинних нейронних мережах та трансформерних архітектурах. Моделі на кшталт BERT, GPT, mBART навчаються на багатомовних корпусах величезного масштабу та демонструють значно кращу здатність до розуміння семантики, моделювання контекстних зв’язків і формування природних мовних конструкцій [7].

Еволюція технологій перекладу супроводжується також розширенням форм інтеграції у вебсередовище: від спеціалізованих API до плагінів для систем керування контентом, модулів для вебфреймворків і комплексних платформ локалізації, що пропонують автоматизацію перекладів у межах повного циклу розробки.

Застосування таких інструментів охоплює широкий спектр сценаріїв: від універсальних сервісів для повсякденного перекладу до рішень, оптимізованих для конкретних галузей або великих вебпроєктів, де необхідна масштабована локалізація. Паралельно із розвитком моделей змінюються й підходи до їхнього налаштування: сучасні системи дозволяють адаптувати переклад під стиль, тон, галузеву термінологію та регіональні варіації, що робить їх значно гнучкішими порівняно з попередніми поколіннями [24], [27].

Загалом еволюція технологій перекладу характеризується переходом від статистичних методів до нейронних моделей із потужними можливостями семантичного аналізу, активним використанням великих мовних моделей, підвищенням адаптивності до контексту та розширенням інтеграції з інструментами керування контентом. Усі ці тенденції суттєво підвищують якість перекладу вебконтенту і водночас створюють нові виклики, пов’язані з оцінюванням як лінгвістичної, так і технічної якості локалізованих документів, що й зумовлює потребу в розробці сучасних методик аналізу, таких як запропонована у цій роботі [28], [29].

1.1.1 Google Translate

Google Translate є одним із найпоширеніших та наймасштабніших сервісів автоматичного перекладу, який використовує архітектуру нейронного машинного перекладу GNMT і трансформерні моделі [2] [3]. Його головною особливістю є підтримка великої кількості мовних пар, що робить сервіс універсальним інструментом для широкого спектра завдань – від особистих перекладів до інтеграції у вебдодатки та корпоративні платформи. Основою роботи Google Translate є аналіз великих двомовних корпусів, із яких система вивчає статистичні закономірності відповідності між мовами. Завдяки переходу до трансформерної архітектури якість перекладу поступово зросла: моделі стали краще враховувати контекст, зберігати стилістичну цілісність тексту та забезпечувати більш природне звучання перекладених фрагментів [7].

Попри універсальність сервісу, його ефективність має певні обмеження.

Найвищу якість Google Translate демонструє у перекладі повсякденних текстів та поширених мовних пар, тоді як для спеціалізованих термінів або рідкісної лексики точність може знижуватися. Іншою проблемою є обмежені можливості персоналізації: система не передбачає глибокого налаштування під конкретний стиль, домен чи вимоги конкретного проєкту, що може бути критичним під час локалізації вузькоспеціалізованих вебресурсів. Водночас завдяки простій інтеграції через API сервіс широко застосовується для автоматизації перекладу контенту в CMS, локалізації інтернет-магазинів, динамічних вебсторінок та інформаційних порталів, де важливо забезпечити швидкість і масштабованість обробки тексту.

1.1.2 DeepL

DeepL є сервісом нейронного машинного перекладу, який здобув популярність завдяки високій якості перекладу, особливо щодо європейських мов [4]. Його моделі створені на основі власних глибинних нейронних мереж та спеціалізованих алгоритмів, що забезпечують значно глибше опрацювання контексту порівняно з універсальними системами. Технологічна платформа DeepL орієнтована на відтворення стилістичних особливостей тексту та забезпечує плавний, природний переклад, який часто перевершує результати інших широко застосовуваних сервісів.

Проте функціональність DeepL має певні обмеження. Кількість підтримуваних його мов є меншою, ніж у Google Translate, що зменшує універсальність сервісу та його застосовність у великих мультимовних проєктах. Крім того, менш активна спільнота користувачів та порівняно вузька екосистема інструментів інтеграції ускладнюють пошук готових рішень та прикладів використання для розробників. Попри це DeepL залишається одним із найкращих виборів для локалізації галузевого або корпоративного контенту, де важливі точність формулювань, змістова відповідність та стилістична послідовність.

Сервіс активно застосовується у локалізації освітніх платформ, фінансових сервісів та бізнес-систем, де якість мовлення має критичне значення.

1.1.3 Microsoft Translator

Microsoft Translator є сервісом нейронного машинного перекладу, який працює в екосистемі хмарних технологій Azure та активно інтегрується з продуктами Microsoft [5]. Його архітектура ґрунтується на використанні нейронних мереж, оптимізованих для вебсервісів із високою пропускною здатністю та масштабованістю, що дозволяє обробляти великі обсяги контенту у реальному часі. Завдяки інтеграції з Azure Cognitive Services сервіс забезпечує переклад із урахуванням контексту, підтримує асинхронну обробку тексту, а також надає програмний доступ через API і SDK для різних мов програмування.

Хоча Microsoft Translator вирізняється стабільністю роботи та високим рівнем масштабованості, його якість перекладу не завжди відповідає рівню систем, орієнтованих на глибоку семантичну обробку, таких як DeepL чи OpenAI. Особливо це помітно у спеціалізованих текстах, де потрібна точна передача термінології або стилю. Проте у корпоративному середовищі сервіс має значні переваги завдяки безшовній інтеграції з Office 365, Teams, SharePoint та внутрішніми інструментами Microsoft, що дозволяє спростити процес локалізації документації, внутрішніх порталів і бізнес-додатків. Загалом Microsoft Translator є надійним рішенням для організацій, які використовують екосистему Microsoft і потребують автоматизації перекладу у межах корпоративних робочих процесів.

1.1.4 Opus-MT та інші відкриті моделі

Opus-MT належить до найбільш відомих відкритих систем нейронного машинного перекладу, створених на основі архітектури MarianNMT та тренованих на багатомовному корпусі OPUS. Головною особливістю таких моделей є їхня повна відкритість: вони доступні у вигляді готових ваг, скриптів для тренування та інструментів для донавчання, що дозволяє дослідникам і розробникам адаптувати їх під конкретні домени або мовні пари [6]. Завдяки

простоті розгортання та можливості локального використання Opus-MT часто застосовується у проєктах, де важливо забезпечити контроль над даними та мінімізувати залежність від комерційних API.

Попри це, відкриті моделі поступаються деяким комерційним рішенням у точності та стилістичній природності перекладу. Це пов’язано з тим, що їх тренують на відкритих корпусах, які не завжди забезпечують достатню якість і збалансованість даних [24]. Унаслідок цього переклади можуть містити стилістичні спрощення, непослідовність термінології або контекстні помилки, особливо на рідкісних мовних парах чи в галузевих текстах. Ще однією відмінністю відкритих моделей є несистемна підтримка: оновлення відбуваються нерегулярно, а рівень оптимізації залежить від спільноти та залучених дослідників [6].

Водночас Opus-MT та інші відкриті системи, такі як MarianNMT, Bergamot чи Helsinki-NLP, залишаються важливими інструментами в екосистемі машинного перекладу. Вони дають можливість створювати локальні інфраструктури перекладу, гнучко адаптувати моделі під специфічні завдання та використовувати навчальні дані, що не можуть бути передані у зовнішні комерційні сервіси [23]. В таких сценаріях відкриті моделі стають оптимальним вибором для організацій, яким важливо зберігати конфіденційність даних, контролювати процес перекладу або впроваджувати спеціалізовані підходи до локалізації.

Таким чином, хоча Opus-MT поступається сучасним комерційним сервісам у загальній якості перекладу, його відкритість, гнучкість та можливість донавчання роблять його значущим інструментом у тих випадках, коли необхідна повна кастомізація, автономність та прозорість роботи моделі.

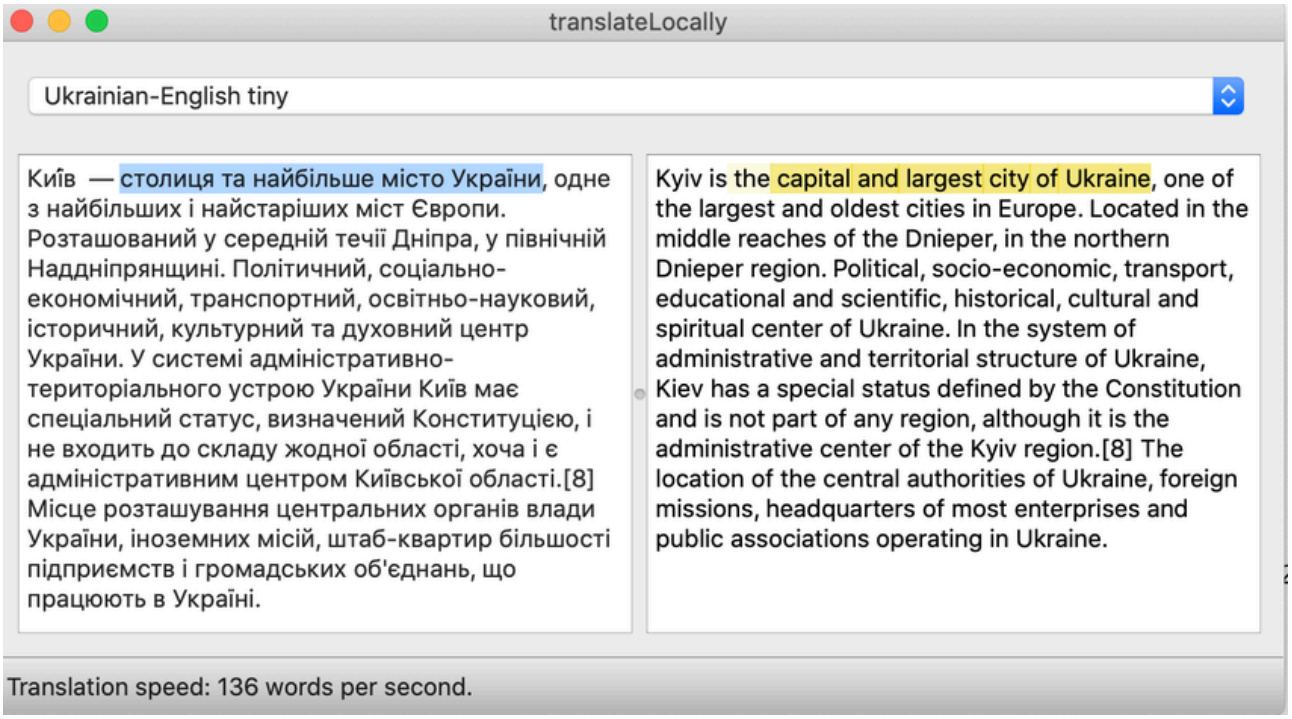


Рис. 1.1 Приклад роботи Opus-MT

1.1.5 Хмарні сервіси та інтерфейси прикладного програмування (API)

Хмарні сервіси та API сьогодні є ключовим механізмом інтеграції автоматичного перекладу у вебплатформи, мобільні застосунки та корпоративні системи. Основні постачальники, такі як Google, Microsoft і DeepL, пропонують доступ до своїх нейронних моделей через інтерфейси прикладного програмування, що дозволяє розробникам передавати текстові фрагменти та отримувати переклад у режимі реального часу [5]. Завдяки хмарній інфраструктурі такі сервіси забезпечують високу масштабованість, автоматичне розподілення навантажень і стабільну роботу навіть при значних обсягах перекладених даних. Це робить інтеграцію перекладу доступною практично для будь-якого технологічного стека та вебфреймворку.

Використання API дає змогу отримувати доступ до найновіших моделей машинного перекладу без потреби у локальному розгортанні складних систем. Постачальники регулярно оновлюють свої моделі, покращуючи їхню точність і контекстну чутливість, а також забезпечують додаткові можливості, такі як обробка HTML, збереження форматування чи робота з великими масивами даних.

Разом з тим залежність від комерційних хмарних сервісів створює низку обмежень. У випадку технічних збоїв, зміни тарифів або блокування доступу користувач може втратити можливість перекладати контент у потрібному обсязі. Додатковою проблемою є питання конфіденційності, оскільки передавання тексту у зовнішні сервіси може бути неприйнятним для проєктів, що працюють з чутливою інформацією або персональними даними [21].

Попри ці ризики, хмарні API залишаються основним способом інтеграції автоматичного перекладу в більшості сучасних вебсистем, забезпечуючи гнучкість, швидкодію та постійне вдосконалення без залучення користувача до технічного обслуговування моделей.

1.1.6 Локалізаційні платформи та екосистема інструментів

Поряд із сервісами автоматичного перекладу сформувалася окрема категорія інструментів – локалізаційні платформи, що охоплюють повний цикл роботи з багатомовним контентом. Системи на кшталт Crowdin, Lokalise та Transifex забезпечують централізоване керування файловою структурою перекладів, координацію команд, роботу з термінологією та організацію робочих процесів між автоматичними алгоритмами й перекладачами-людьми [18], [19]. Такі платформи інтегруються з популярними CMS, фреймворками та репозиторіями, що дозволяє автоматично синхронізувати перекладені матеріали з вихідним контентом і спрощує підтримку багатомовних вебпроєктів.

В останні роки локалізаційні платформи активно впроваджують модулі AI-перекладу, використовуючи Google Translate, DeepL або власні нейронні моделі як перший етап створення чернетки перекладу. Такий підхід значно скорочує час і навантаження на команди перекладачів, оскільки автоматична система створює базову версію тексту, а редактори зосереджуються на стилістичній та термінологічній корекції. На додаток до цього платформи зберігають пам’ять перекладів і глосарії, що сприяє консистентності великомасштабних проєктів і полегшує повторне використання раніше перекладених фрагментів.

Разом із тим локалізаційні платформи мають низку обмежень. Їхня робота

повністю залежить від стороннього сервісу, а отже, у випадку змін політики, технічних проблем або відмови доступу можливе порушення локалізаційного процесу. Крім того, більшість платформ працює за моделлю підписки, що підвищує загальну вартість обслуговування проєктів. Попри це такі системи залишаються ключовими інструментами для організацій, які працюють із великими багатомовними ресурсами, оскільки дозволяють поєднати переваги автоматичного перекладу з контролем і досвідом професійних перекладачів [20].

1.2 Мультимовні трансформерні моделі та локалізація рідкісних мов

Мультимовні трансформерні моделі, серед яких mBERT, XLM-R, mT5 та інші, стали одним із найбільш прогресивних інструментів у сфері автоматичного перекладу, оскільки здатні одночасно працювати з великою кількістю мов і формувати спільний багатомовний простір [10]. На відміну від попередніх підходів, які вимагали наявності великих двомовних корпусів для кожної пари мов, ці моделі навчаються на масштабних багатомовних даних. Завдяки цьому вони можуть переносити знання з мов із добре представленими корпусами на рідкісні мови, для яких обсяг навчальних матеріалів обмежений [25]. Така здатність до міжмовного перенесення є однією з ключових переваг сучасних трансформерних архітектур.

Серйозною проблемою рідкісних мов залишається не лише недостатня кількість паралельних текстів, а й велика варіативність діалектів, нестача стандартизованої термінології та відмінності у культурних контекстах. Для подолання цих обмежень застосовують методи донавчання моделей на спеціалізованих локальних корпусах, а також використовують термінологічні словники, глосарії та мовні бази, що покращують узгодженість перекладів. Додаткові модифікації, зокрема підхід reinforcement learning, дають змогу покращити якість генерації тексту шляхом оптимізації поведінки моделі відповідно до поставлених лінгвістичних цілей [28].

Локалізація рідкісних мов є складним завданням, яке потребує зваженого

підходу, однак розвиток багатомовних трансформерних моделей створює нові перспективи. Очікується, що їх еволюція рухатиметься у напрямі використання мультимодальних даних, що допоможе точніше враховувати контекст і культурні особливості. Важливу роль відіграватиме також розвиток відкритих лінгвістичних ресурсів та спільнотних проєктів, які поступово розширюють доступну мовну базу та роблять автоматичний переклад рідкісних мов більш точним і доступним [6].

1.3 Оцінка якості перекладу та локалізації

Оцінювання якості автоматичного перекладу є одним із ключових аспектів ефективного використання систем штучного інтелекту для локалізації вебконтенту, оскільки від точності та природності перекладу залежить коректність сприйняття інформації користувачами. Традиційно в цій сфері застосовуються автоматизовані метрики, зокрема BLEU [13], METEOR [14] чи TER [17], що базуються на порівнянні машинного перекладу з еталонним людським варіантом. Хоча такі методи дають змогу швидко отримати кількісну оцінку, вони не здатні повною мірою врахувати стилістичні особливості, тональність, культурні нюанси або контекстуальні залежності, що часто є критично важливими для вебкомунікацій.

Обмеження автоматичних метрик стали підґрунтям для формування комплексних моделей оцінювання, які поєднують алгоритмічний аналіз із експертною оцінкою лінгвістів. Фахівці в галузі перекладу визначають якість за критеріями адекватності – відповідності змісту оригіналу – та плавності, що характеризує граматичну та стилістичну коректність перекладу. В окремих випадках до процесу оцінювання залучають кінцевих користувачів, які можуть надати цінну інформацію про те, наскільки перекладений контент є зрозумілим, природним і релевантним для конкретної аудиторії.

Суттєвим елементом локалізації є адаптація змісту до культурних

особливостей регіону, для якого створюється переклад. Йдеться про коректне використання термінології, відповідність форматів дат, чисел, валют та інших локальних параметрів, а також про відповідність соціальним і культурним нормам. Оцінювання таких аспектів передбачає додаткові критерії культурної адекватності, які часто потребують участі носіїв мови або фахівців у галузі міжкультурної комунікації.

1.4 Етичні, правові та безпекові аспекти

Використання систем штучного інтелекту для автоматичного перекладу та локалізації вебконтенту супроводжується низкою важливих етичних, правових і безпекових викликів. Одним із ключових питань є конфіденційність даних, оскільки текст, переданий на переклад, може містити персональну або комерційно чутливу інформацію. Передавання таких даних стороннім сервісам створює ризики витоку, несанкціонованого доступу або порушення вимог міжнародних стандартів на кшталт GDPR. Особливо загрозливою є ситуація, коли компанії обробляють фінансові, медичні чи юридично значущі документи, які вимагають особливо високого рівня захисту.

Правові аспекти також мають суттєве значення, оскільки автоматичний переклад часто працює з матеріалами, захищеними авторським правом, торговими марками або ліцензійними угодами. Використання AI-моделей повинно відповідати вимогам власників інтелектуальної власності, а автоматичне генерування перекладу не звільняє користувачів від необхідності дотримання авторських прав. Крім того, автоматичні моделі можуть припускатися помилок у юридично важливих текстах, що потенційно призводить до неправильного трактування договорів, внутрішніх політик чи умов використання [23].

Етичні виклики пов’язані з тим, що мовні моделі відображають властивості даних, на яких вони були навчені. Якщо навчальні корпуси містять упередження або дискримінаційні елементи, модель може відтворювати та підсилювати такі патерни у своїх перекладах. Це стає особливо критичним у випадках

міжкультурної комунікації, де навіть незначні змістові зсуви можуть спровокувати непорозуміння або завдати репутаційних збитків. Також переклад, виконаний без урахування культурних особливостей, специфічного гумору, соціальних табу або контексту, може бути неправильно інтерпретований цільовою аудиторією.

Забезпечення етичності та безпеки вимагає застосування цілісного підходу, який поєднує технічні та організаційні заходи. Для роботи з конфіденційними даними клінічних або юридичних доменів доцільно використовувати локальні рішення або механізми шифрування, що зменшують ризик витоку. У випадках, коли переклад має бути максимально точним і нейтральним, необхідна участь експертів-редакторів, які можуть верифікувати результат [6]. Важливо також забезпечувати навчання моделей на збалансованих корпусах, проводити регулярний аудит на предмет упередженості та дотримуватися міжнародних і локальних правових норм щодо захисту даних.

1.5 Виклики та шляхи їх подолання

Попри значний прогрес у розвитку систем автоматичного перекладу та локалізації, низка критичних викликів залишається актуальною. Однією з найбільш суттєвих проблем є обмежена якість перекладу спеціалізованих текстів, що містять галузеву термінологію медичного, технічного, юридичного чи наукового спрямування. Моделі, які навчаються переважно на загальних корпусах, часто не розпізнають специфічні терміни або інтерпретують їх неправильно. Подолання цього недоліку можливе шляхом цільового донавчання моделей на спеціалізованих текстових базах, використання глосаріїв термінів і застосування підходів доменної адаптації.

Ще одним важливим викликом є збереження стилю та тону оригінального тексту. Більшість систем машинного перекладу орієнтується на передачу змістової адекватності, проте не здатна достатньо точно відтворювати стилістичні нюанси, особливо в художніх, маркетингових або корпоративних текстах. Для вирішення цієї проблеми доцільним є застосування моделей, адаптованих під конкретні

жанри або комунікативні завдання [24], а також використання етапу післяредагування, який дозволяє забезпечити відповідність стилю та тональності очікуванням цільової аудиторії [27].

Динамічний характер сучасного вебконтенту створює додаткові труднощі, оскільки інформація на сторінках часто оновлюється, і переклад повинен відповідати цим змінам у режимі реального часу. Найефективнішим шляхом подолання цієї проблеми є впровадження автоматизованих CI/CD-процесів, у рамках яких переклад оновлюється щоразу, коли змінюється вихідний контент у репозиторії. Використання хмарних API-сервісів дозволяє підтримувати актуальність локалізації без значних часових витрат [5].

Окреме місце серед викликів посідає культурна адаптація, оскільки локалізація передбачає не лише переклад тексту, а й коректне відтворення контексту, візуальних елементів, форматів дат, валют, одиниць вимірювання та інших локальних стандартів. Автоматичні системи поки що не завжди здатні адекватно враховувати такі особливості, тому перспективним напрямом є комбіновані підходи, що поєднують можливості AI-перекладу з інструментами культурної адаптації [19] та участю експертів-редакторів.

2 АНАЛІЗ МЕТОДІВ ТА МЕТРИК ОЦІНЮВАННЯ ЯКОСТІ АВТОМАТИЧНОГО ПЕРЕКЛАДУ І ЛОКАЛІЗАЦІЇ ВЕБКОНТЕНТУ

2.1 Автоматичні метрики якості перекладу

Автоматичні метрики оцінювання якості машинного перекладу (Machine Translation Evaluation Metrics, MTE Metrics) – це формалізовані методи, що дозволяють кількісно оцінити, наскільки перекладений текст відповідає еталонному (референсному) перекладу, без залучення людини. Такі метрики є основою для швидкого порівняння систем перекладу, моніторингу їх продуктивності та проведення експериментів у сфері штучного інтелекту.

Основна ідея автоматичних метрик полягає у вимірюванні схожості між машинним перекладом і референсом за певними лінгвістичними ознаками: збіг слів, морфем, символів або семантичних представлень. На відміну від людських оцінювань, автоматичні метрики забезпечують відтворюваність, масштабованість і швидкість аналізу, що робить їх придатними для інтеграції в системи CI/CD або хмарні сервіси перекладу.

Автоматичне оцінювання якості перекладу особливо важливе для вебконтенту, оскільки вебтексти мають різнорідну структуру (заголовки, кнопки, повідомлення, описи) і часто містять елементи, що не підлягають перекладу (HTML-теги, змінні, символи розмітки). Тому метрики повинні враховувати не лише лексичну точність, але й контекст, термінологічну сталість і збереження структури [30].

У сучасних дослідженнях автоматичні метрики поділяються на три основні групи:

1. Статистичні метрики на основі збігів слів або символів – порівнюють лексичні n-грамові збіги між перекладом і референсом (наприклад, BLEU [13], METEOR [14], chrF [25]).

   Семантичні метрики на основі ембедингів або моделей глибокого навчання -

оцінюють смислову подібність між перекладом і референсом (COMET [12], BLEURT [11]).

2. Гібридні підходи, які поєднують статистичну точність із семантичними ознаками або лінгвістичними правилами, забезпечуючи більш гнучку оцінку для різних типів текстів.

Для практичного застосування у вебсередовищі автоматичні метрики мають кілька переваг:

- швидке масштабне порівняння перекладів від різних систем (наприклад, Google Translate, DeepL, OpenAI);
- можливість інтеграції у пайплайни автоматичного тестування локалізації;
  оцінка якості без необхідності людського втручання;
- використання в якості зворотного зв’язку для fine-tuning або адаптації AI-моделей перекладу [9].

Однак у контексті вебконтенту існують додаткові виклики:

1. автоматичні метрики часто не враховують функціональні аспекти локалізації (наприклад, збереження посилань або плейсхолдерів);
2. однакові за змістом переклади можуть мати різні формулювання, що знижує оцінку, хоча якість перекладу залишається прийнятною;
3. метрики не завжди адекватно відображають стиль, культурну адаптацію або SEO-оптимізацію тексту.

Таким чином, використання автоматичних метрик є необхідним, але не достатнім кроком у комплексній оцінці систем перекладу для вебконтенту. Вони повинні поєднуватися з семантичними аналізаторами, перевірками структурної цілісності та, у деяких випадках, із людською оцінкою.

2.1.1 BLEU, METEOR, chrF, COMET, BLEURT

Для оцінювання якості машинного перекладу впродовж останніх двох десятиліть було розроблено низку метрик, які поступово еволюціонували від простих статистичних підрахунків до моделей, заснованих на глибокому навчанні. Нижче розглянуто найбільш поширені та впливові серед них – BLEU, METEOR, chrF, COMET та BLEURT, що використовуються як у наукових дослідженнях, так і у промислових рішеннях локалізації.

BLEU – одна з перших і найвідоміших метрик автоматичного оцінювання перекладу, запропонована компанією IBM у 2002 році [13].

Основна ідея BLEU полягає у вимірюванні збігу n-грам (послідовностей з n слів) між машинним перекладом і одним або кількома референсними перекладами.

Формула BLEU базується на двох ключових компонентах:

- Precision n-грам – частка збігів між перекладом і еталоном;
- Brevity Penalty (BP) – штраф за занадто короткі переклади.

Загальний результат BLEU – це середньозважене геометричне значення точності для n=1...4, скориговане на BP.

Переваги BLEU:

- простота реалізації й висока швидкість розрахунку;
- можливість застосування на великих корпусах;
- придатність для порівняння систем.

Недоліки BLEU:

- не враховує синонімію, граматичну правильність і семантичну близькість;
- чутливість до перестановок слів;
- у контексті вебконтенту BLEU може давати занижені оцінки для коротких елементів (заголовки, кнопки).

Метрика METEOR, розроблена у 2005 році, спрямована на подолання недоліків BLEU [14].

Вона оцінює переклад не лише за точним збігом слів, а й враховує стемінг, синонімію та морфологічні варіації.

Принцип дії METEOR:

1. Виконується вирівнювання (alignment) між перекладом і референсом за точними, лематизованими або синонімічними збігами.
2. Обчислюються Precision і Recall (на відміну від BLEU, METEOR враховує і повноту).
3. Застосовується штраф за порушення порядку слів (fragmentation penalty).

Переваги METEOR:

- краща кореляція з людськими оцінками, ніж BLEU;
- врахування морфології, що важливо для слов'янських мов;
- придатна для коротких фраз і UI-текстів.

Недоліки:

- вища обчислювальна складність;
- потребує лексичних ресурсів (WordNet тощо);
- менш ефективна для мов із обмеженими словниками.

Метрика chrF базується на порівнянні перекладу й еталону на рівні символів (char-level) [27].

Вона обчислює F-міру (гармонійне середнє між Precision і Recall) для n-грам символів.

chrF добре підходить для мов із багатою морфологією (українська, фінська, турецька) або при оцінюванні коротких сегментів, де BLEU нестабільний.

Переваги chrF:

- нечутливість до розривів слів і флексій;
- стабільність при коротких текстах (корисно для вебінтерфейсів);
  не потребує лінгвістичних ресурсів.

Недоліки:

- не відображає семантичної близькості;
- не враховує структуру речення.

COMET – сучасна нейронна метрика, розроблена дослідниками Facebook AI у 2020 році [12].

Вона використовує глибокі мовні моделі (наприклад, XLM-R) для представлення речень у багатомовному векторному просторі.

Принцип дії:

1. Тексти оригіналу, перекладу та (за потреби) референсу проходять через багатомовну трансформерну модель.
   На основі векторних представлень обчислюється косинусна подібність між перекладом і еталоном.
2. Модель додатково навчається на даних із людськими оцінками, щоб максимізувати кореляцію з ними.

Переваги COMET:

- висока відповідність людським оцінкам (найкраща серед сучасних метрик);
  здатність розуміти синонімію, контекст і стиль;
- добре працює без точного збігу слів.

Недоліки:

- значні обчислювальні ресурси;

- залежність від якості попереднього тренування моделі;
  не завжди відтворювана при відсутності GPU.

Для оцінювання вебконтенту COMET особливо корисна, оскільки може оцінити не лише лексичну, але й семантичну еквівалентність коротких текстів (наприклад, кнопок, форм або SEO-заголовків).

BLEURT – ще одна сучасна нейронна метрика, розроблена дослідниками Google у 2020 році [11].

На відміну від COMET, вона базується на файнтюнінгу BERT, навченого на великій кількості синтетично "зіпсованих" перекладів.

Принцип BLEURT:

1. Модель порівнює переклад із референсом, оцінюючи, наскільки переклад «схожий на людський».
2. BLEURT тренується прогнозувати оцінки, отримані від людських експертів, тому її вихід - одна числова оцінка (зазвичай у діапазоні від 0 до 1).

Переваги BLEURT:

- глибоке семантичне розуміння тексту;
- здатність враховувати контекст і стиль;
  висока узгодженість із людськими оцінками.

Недоліки:

- вимагає попереднього fine-tuning під конкретну мовну пару;
- результати можуть бути нестабільними для коротких речень.

У контексті вебконтенту BLEURT має перевагу над BLEU і METEOR завдяки контекстній чутливості – вона здатна розрізняти переклади, що граматично правильні, але зміщують смисл, або не відповідають тональності сайту.

Перелічені метрики формують еволюційний шлях розвитку підходів до автоматичного оцінювання перекладу:

від поверхневого лексичного аналізу (BLEU, METEOR, chrF) до глибоких семантичних моделей (COMET, BLEURT).

Для завдань локалізації вебконтенту найкращі результати демонструють COMET і BLEURT, оскільки вони враховують не лише лексичну точність, але й смислову адекватність, стиль і адаптованість перекладу до контексту.

Проте на практиці часто застосовується комбінований підхід, швидкі статистичні метрики (наприклад, chrF) використовуються для первинного відбору, нейронні метрики (COMET, BLEURT) – для фінальної оцінки якості та додатково - власні вебспецифічні тести на збереження тегів, плейсхолдерів і структури.

2.1.2 Переваги та обмеження метрик на вебконтенті

Вебконтент суттєво відрізняється від класичних текстових корпусів, які зазвичай використовуються для навчання та тестування систем машинного перекладу. На відміну від суцільних речень у новинних або художніх текстах, вебконтент складається з коротких, фрагментованих сегментів: заголовків, кнопок, повідомлень інтерфейсу, підписів до зображень, SEO-заголовків, динамічних змінних та елементів HTML-розмітки [21].

Через це застосування стандартних метрик, таких як BLEU, METEOR, chrF, COMET чи BLEURT, має як очевидні переваги, так і суттєві обмеження.

Переваги автоматичних метрик для вебконтенту

1. Швидкість та масштабованість.

   Метрики дозволяють швидко оцінювати якість перекладу великих обсягів вебконтенту, включно з багатомовними сайтами чи CMS-проєктами, де ручна оцінка є непрактичною [25].
2. Відтворюваність результатів.

   Автоматичні метрики гарантують об’єктивність та сталість оцінок - на відміну від людських оцінювачів, вони не залежать від суб’єктивних факторів чи настрою рецензента.

3. Зручність інтеграції у CI/CD-процеси.

   Метрики можуть бути частиною автоматизованого пайплайну, що перевіряє якість перекладу при оновленні вебсайту або додавання нових локалей.

4. Порівняльний аналіз систем.

   Вони дозволяють без втручання людини порівнювати результати різних перекладацьких систем (наприклад, Google Translate, DeepL, OpenAI або власного AI-модуля) для однакових текстів, що забезпечує основу для вибору оптимальної технології [23].

5. Можливість побудови композитних оцінок.

   У поєднанні з іншими показниками (наприклад, часом перекладу або відповідністю HTML-структури) метрики можуть стати частиною узагальненого показника якості – інтегрального індексу (WALE-Score, QEval, MQM+).

Попри численні переваги, класичні автоматичні метрики мають низку обмежень, які стають особливо помітними у сфері веблокалізації:

1. Короткі тексти та нестача контексту.

   Більшість вебелементів складається з 1-10 слів (наприклад, "Submit", "Read more", "Welcome back!"), що робить показники BLEU або METEOR нестабільними. Навіть один синонім або інша форма дієслова може значно змінити оцінку.

2. Нерівномірна довжина перекладів.

   вебінтерфейси часто вимагають перекладів певної довжини, щоб вони вміщалися в кнопки або меню. Метрики не враховують обмеження верстки та візуальні аспекти локалізації [20].

3. Ігнорування HTML-структури та плейсхолдерів.

   Під час перекладу вебконтенту важливо зберігати теги (<b>, <a href>,
), змінні ({username}, %s, {count}) і атрибути. Автоматичні метрики зазвичай не аналізують структурну цілісність, тому переклад із пошкодженою розміткою

може отримати високу оцінку, хоча він непридатний для відображення на сайті.

4. Відсутність культурної адаптації та стилістичного аналізу.
   Метрики не здатні визначити, чи переклад відповідає тональності бренду, культурним особливостям або SEO-оптимізації. Наприклад, "Learn more" → "Докладніше" і "Дізнайтеся більше" є еквівалентними перекладами, але перший варіант може бути стилістично ближчим до загального тону сайту.
5. Чутливість до пунктуації та регістру.
   У вебконтенті часто застосовуються капіталізовані елементи, емоджі, короткі гасла або навіть HTML-ентіті. Це може знижувати оцінку BLEU чи METEOR, навіть якщо переклад граматично й семантично коректний.
6. Відсутність обліку динамічних або персоналізованих сегментів.
   Тексти, що формуються на основі шаблонів (наприклад, “Hello, {userName}!”), вимагають спеціальних правил оцінювання, щоб уникнути помилкових штрафів за змінні, які не підлягають перекладу [21].

Багато метрик орієнтовані на лексичний збіг (n-грам), що не завжди корелює зі смисловою точністю. Для вебконтенту, де фрази мають маркетинговий або інтерфейсний характер, смислова адекватність часто важливіша за буквальний збіг.

Наприклад, переклад “Get started” → “Розпочати роботу” може отримати нижчий BLEU-бал, ніж “Отримати старт”, хоча саме перший варіант є правильним із точки зору природності та локалізації.

Для забезпечення більш коректного оцінювання якості перекладу вебконтенту необхідно:

- комбінувати метрики різних типів (лексичні + семантичні + структурні);
- доповнювати їх перевірками цілісності HTML (збереження тегів, плейсхолдерів, атрибутів);
- враховувати контекст UI - тобто позицію перекладу на сторінці, обмеження символів, стиль;

- використовувати нейронні метрики (COMET, BLEURT) для оцінки смислової близькості замість простих збігів;
- розробити комбінований інтегральний показник як частину методики оцінювання, який враховує і лінгвістичні, і технічні аспекти якості.

Отже, класичні метрики, попри свою популярність і простоту, не забезпечують повної об’єктивності при оцінюванні перекладу вебконтенту.
Для коротких, динамічних, структурованих текстів потрібні адаптовані методи, що поєднують автоматичні метрики якості перекладу з перевірками функціональності, структурної цілісності та контекстної релевантності.

Це створює передумови для розроблення нової методики комплексної оцінки, що буде представлена в наступних розділах.

2.2. Людські методи оцінювання (LQA, MQM, HTER)

Попри значний прогрес у розвитку автоматичних метрик, саме людське оцінювання залишається найнадійнішим способом визначення якості машинного перекладу. Жодна формула не здатна повністю врахувати контекст, намір автора, культурні відтінки чи цільову аудиторію. Людські оцінювачі вміють інтерпретувати зміст у ширшому семантичному полі, тому їхні висновки часто розглядаються як еталонні при навчанні або калібруванні автоматичних систем. У сфері локалізації вебконтенту це набуває особливої ваги, адже мова на сайтах виконує не лише інформативну, а й маркетингову, навігаційну та емоційну функцію.

Одним із найдавніших підходів є LQA (Linguistic Quality Assurance), який використовується у виробничих процесах перекладу та локалізації. Його сутність полягає у виявленні й класифікації помилок за категоріями - граматика, лексика, пунктуація, стиль, узгодженість термінів, збереження форматування. Для кожного типу помилки визначається рівень серйозності – критичний, серйозний або незначний [16]. Після цього формується загальний бал якості на основі вагових коефіцієнтів. У веблокалізації такий підхід часто застосовується для контролю якості

інтерфейсів і маркетингових сторінок, де важливими є не лише точність перекладу, а й його відповідність тону бренду та обмеженням інтерфейсу.

Із розвитком досліджень у галузі оцінювання перекладу з'явилася більш стандартизована система MQM (Multidimensional Quality Metrics), що була запропонована Європейським комітетом стандартизації [15]. Ця модель відрізняється від традиційного LQA тим, що пропонує багатовимірну структуру оцінювання, де якість визначається через ієрархію категорій і підкатегорій помилок. MQM дозволяє створити адаптовану схему, наприклад, для технічних документів, вебінтерфейсів або рекламних текстів. Основна перевага MQM полягає у можливості кількісно оцінити суб'єктивні аспекти перекладу, такі як відповідність аудиторії, узгодженість стилю або коректність передачі інтенції повідомлення. Для вебконтенту ця гнучкість є вирішальною, адже тексти мають різне призначення - від коротких закликів до дії до довгих описів продуктів.

Ще одним поширеним способом оцінювання, який активно використовується в дослідженнях машинного перекладу, є HTER (Human-targeted Translation Edit Rate). На відміну від LQA і MQM, він не ґрунтується на категоріях помилок, а вимірює кількість редагувань, які потрібно виконати людині, щоб зробити машинний переклад прийнятним [17]. Чим менше правок потрібно, тим кращою вважається система. Цей метод має перевагу в об'єктивності, оскільки результат виражається конкретними кількісними змінами – вставками, вилученнями або перестановками. У вебконтенті HTER допомагає визначити не лише якість перекладу, а й потенційний обсяг ручної постредакції, що прямо впливає на витрати часу і коштів при підтримці багатомовних сайтів.

Попри ефективність людських методів оцінювання, вони мають очевидні обмеження. По-перше, процес є трудомістким і потребує участі фахових лінгвістів. По-друге, людські оцінки залишаються певною мірою суб'єктивними, особливо коли йдеться про короткі або стилістично варіативні тексти [29]. По-третє, ручне оцінювання складно масштабувати для великих вебресурсів, які регулярно оновлюють контент. Саме тому сучасна тенденція полягає не у відмові від людських оцінок, а у їхньому поєднанні з автоматичними метриками. Таке гібридне

оцінювання дозволяє використовувати переваги обох підходів: швидкість і об’єктивність машинних методів разом із точністю й глибиною людського аналізу.

Для побудови методики оцінювання систем автоматичного перекладу вебконтенту доцільно використовувати людські методи як «золотий стандарт», на основі якого можна калібрувати автоматичні метрики. Наприклад, результати MQM можуть бути використані для навчання або перевірки моделей COMET чи BLEURT, а показники HTER – для кількісної оцінки економічної ефективності системи. Таким чином, людське оцінювання не замінює автоматичні підходи, а навпаки, підсилює їх, забезпечуючи більш повну та надійну картину якості перекладу, що особливо важливо у вебсередовищі, де кожне слово впливає на сприйняття користувача і взаємодію з продуктом.

2.3. Метрики швидкодії та продуктивності систем

Оцінювання якості перекладу не обмежується лише лінгвістичними характеристиками. У практичному застосуванні систем автоматичного перекладу, особливо у вебсередовищі, не менш важливим є показник швидкодії, тобто час, необхідний для обробки запиту та генерації перекладу. вебкористувачі звикли до миттєвої взаємодії з контентом, тому затримка навіть у кілька секунд може знизити задоволеність і вплинути на ефективність локалізованого ресурсу. Саме тому сучасні системи оцінюються не лише за якістю, а й за продуктивністю – наскільки швидко й економно вони можуть забезпечити потрібний рівень перекладу.

Метрики швидкодії зазвичай охоплюють кілька взаємопов’язаних аспектів. Найпоширенішим є показник latency - середній час відправлення запиту до отримання перекладу. Він може розраховуватись у мілісекундах або секундах залежно від типу сервісу. Для веборієнтованих систем особливе значення має різниця між середнім і максимальним часом відгуку, оскільки саме пікові затримки визначають стабільність користувацького досвіду. Іншим важливим параметром є throughput, який відображає кількість символів або токенів, що система здатна обробити за певний проміжок часу. Цей показник дозволяє порівнювати

ефективність різних архітектур – наприклад, моделей, розгорнутих на GPU, із хмарними API-провайдерами.

Окрім швидкодії, суттєве значення має продуктивність системи, що охоплює не лише обчислювальні ресурси, але й вартість їх використання. У комерційних рішеннях, таких як Google Translate або DeepL API, ці параметри визначаються тарифікацією за символ чи запит, тоді як у власних AI-моделях – споживанням обчислювальної потужності, пам’яті та енергоресурсів.

Оптимальна система має забезпечувати баланс між якістю перекладу та витратами на обчислення. У цьому контексті запроваджують поняття ефективності перекладу (Translation Efficiency), що відображає співвідношення якості результату до обсягу використаних ресурсів. Такий показник особливо важливий для масштабних багатомовних проєктів, де тисячі текстів перекладаються щогодини.

При аналізі веблокалізації продуктивність системи тісно пов’язана з обробкою паралельних запитів. Під час перекладу великих сайтів або контенту, що генерується динамічно, система повинна підтримувати одночасну роботу з десятками або сотнями запитів без втрати стабільності. У таких умовах визначальним стає коефіцієнт Concurrency Performance, який характеризує здатність сервера підтримувати навантаження без деградації швидкодії. Для систем, інтегрованих у вебсервіси, це має безпосередній вплив на час завантаження сторінок і показники взаємодії користувачів (UX).

Важливим аспектом є також час ініціалізації перекладу (Startup Time), який особливо помітний у моделях, що розгортаються локально або працюють через API. На відміну від попередньо активованих хмарних сервісів, локальні рішення часто потребують кількох секунд для завантаження моделі в пам’ять. Цей фактор може бути критичним для сценаріїв, де переклад виконується на вимогу в реальному часі, наприклад, у чат-ботах або динамічних інтерфейсах.

Не менш значущою метрикою є час на обробку одиниці контенту (per-segment time). Для вебпроєктів контент зазвичай поділений на невеликі фрагменти: заголовки, кнопки, повідомлення. Надто складні моделі перекладу можуть мати надлишковий час генерації для таких коротких сегментів, що робить їх

неефективними з практичної точки зору. Тому для реальних застосувань іноді доцільно використовувати спрощені або оптимізовані моделі, які забезпечують достатню якість при мінімальному часі відповіді.

У сучасних дослідженнях пропонується поєднувати метрики швидкодії з метриками якості, формуючи узагальнений показник Trade-off Score, який дозволяє оцінити ефективність системи комплексно [27]. Такий підхід особливо актуальний для систем штучного інтелекту нового покоління, які поєднують переклад, постредагування та семантичну перевірку. Для веблокалізації цей підхід є практично необхідним: навіть найточніша модель не може вважатися успішною, якщо її робота затримує завантаження сторінки або перевищує бюджет проєкту.

У контексті запропонованої методики оцінювання систем перекладу вебконтенту метрики швидкодії відіграють роль доповнюючого критерію, який відображає реальну життєздатність рішення. Вони дозволяють не лише порівнювати системи за швидкістю, але й прогнозувати їхню ефективність у промислових умовах, де час, стабільність і ресурсоспоживання є не менш важливими, ніж якість перекладу. Поєднання цих показників із лінгвістичними метриками формує основу для багатовимірного аналізу ефективності, що є ключовим елементом подальшої розробки методики оцінювання в рамках даного дослідження.

2.4. Аналіз підходів до оцінювання локалізаційної точності та збереження структури вебсторінок

Локалізаційна точність у контексті вебконтенту охоплює не лише адекватність перекладу, а й збереження функціональної, візуальної та структурної цілісності сторінки. На відміну від звичайного текстового перекладу, де головним є смислова відповідність, у веблокалізації необхідно гарантувати, що перекладені елементи не порушують верстку, форматування, внутрішні посилання, а також залишаються сумісними з інтерфейсом користувача.

Це завдання вимагає комплексного підходу, який поєднує лінгвістичний аналіз із технічним тестуванням [22].

Оцінювання локалізаційної точності передбачає перевірку відповідності перекладу функціональним компонентом вебсторінки. Важливо, щоб перекладені тексти правильно узгоджувалися з контекстом елементів інтерфейсу, відображали правильну граматичну форму для чисел, родів або відмінків і не спотворювали зміст через обмеження довжини. Наприклад, неправильний вибір числової форми у шаблоні “{count} items left” може призвести до граматичних помилок (“1 товари залишилось”), що створює враження низької якості продукту. Тому під час оцінювання враховується здатність системи коректно обробляти плейсхолдери, ICU-патерни, змінні та локальні формати даних [16].

Одним із ключових аспектів є збереження HTML-структури. Системи автоматичного перекладу часто спотворюють або видаляють елементи розмітки, особливо якщо не мають спеціального механізму обробки тегів. Це призводить до некоректного відображення сторінки або навіть збоїв у роботі скриптів. Для запобігання таким ситуаціям використовуються методи автоматичної перевірки структурної цілісності, які порівнюють дерево DOM до і після перекладу. За допомогою парсерів (наприклад, JSDOM або BeautifulSoup) можна визначити, чи всі відкриті теги мають відповідні закриті, чи збережено атрибути href, alt, title, а також чи залишилися на місці змінні у дужках або шаблонних фігурних скобках [22].

Перевагою таких підходів є їхня точність та можливість повної автоматизації. Система може швидко перевірити сотні перекладених сторінок, виявити пошкоджені елементи або невідповідності структури й повідомити про це розробнику чи локалізатору. Такі інструменти добре інтегруються у процеси безперервного розгортання та тестування вебдодатків, що робить їх зручними для великих проєктів. Проте їхнім недоліком є відносна нечутливість до контексту: автоматичний аналіз може виявити правильну структуру, але не здатний оцінити, чи переклад узгоджується зі змістом сторінки або дизайном. Наприклад, елемент <button> може залишатися технічно справним, навіть якщо перекладена фраза не відповідає очікуваному тону (“Прийняти умови” замість “Погодитися”) [25].

Окремим напрямом оцінювання є тестування адаптованості перекладу до верстки. Через різницю в довжині слів між мовами деякі елементи можуть виходити

за межі контейнерів або спричиняти зміщення блоків. Для цього використовують автоматичні візуальні тести, які порівнюють скріншоти сторінок до і після локалізації. Такий підхід дозволяє виявити не лише помилки в тексті, а й проблеми з відображенням, наприклад, зсуви кнопок, обриви рядків або перекриття зображень. Його перевага полягає у можливості оцінити локалізацію з точки зору користувацького досвіду, що особливо важливо для інтерфейсів із великою кількістю інтерактивних елементів. Недоліком залишається складність автоматичного аналізу результатів: навіть незначна зміна шрифту або кольору може бути помилково розцінена як структурне порушення [23].

Важливою складовою оцінювання є також перевірка локалізованих даних – форматів дат, чисел, валют, телефонних кодів і адрес. Неправильне відображення таких елементів знижує довіру користувачів, навіть якщо загальний переклад є якісним. Для цього застосовують регулярні вирази або спеціалізовані бібліотеки, що перевіряють відповідність даних вибраній локалі. Наприклад, формат дати “2025-11-05” має бути перетворений у “05.11.2025” для української версії, тоді як для англійської - у “November 5, 2025”. Така перевірка вимагає поєднання лінгвістичного аналізу з технічним парсингом, що робить її важливим компонентом комплексного оцінювання якості локалізації.

Загалом аналіз показує, що ефективна оцінка локалізаційної точності повинна враховувати як мовну адекватність, так і технічну цілісність контенту. Автоматизовані перевірки HTML, ICU-плейсхолдерів і форматів даних забезпечують стабільність і масштабованість, тоді як людський контроль гарантує відповідність перекладу контексту й стилю. Поєднання цих підходів створює основу для комплексної методики, у межах якої збереження структури вебсторінок розглядається не лише як технічний, а й як якісний показник загальної ефективності систем перекладу.

2.5. Порівняння існуючих підходів та обґрунтування вибору метрик для методики

У попередніх підрозділах було розглянуто основні групи підходів до оцінювання систем автоматичного перекладу – від традиційних автоматичних метрик до людських методів контролю якості та технічних способів перевірки структурної цілісності вебконтенту. Проведений аналіз дозволяє зробити висновок, що жоден із цих підходів окремо не забезпечує повного уявлення про якість перекладу у вебсередовищі. Кожен із них має свої сильні сторони й обмеження, які визначають доцільність їх застосування залежно від типу контенту, вимог до швидкодії та цілей локалізації.

Автоматичні метрики, зокрема BLEU, METEOR і chrF, продемонстрували високу ефективність для порівняння систем і моніторингу змін якості в часі. Вони прості у використанні, забезпечують швидку обробку великих корпусів і добре масштабуються в хмарних рішеннях. Однак їхнім головним недоліком є чутливість до лексичних варіацій і відсутність розуміння контексту.

У випадку коротких вебсегментів, типових для інтерфейсів або SEO-заголовків, результати таких метрик часто не відображають реальної якості перекладу. Семантичні метрики нового покоління, як COMET і BLEURT, зменшують цю проблему завдяки здатності моделювати смислову близькість між перекладом і оригіналом. Вони краще корелюють із людськими оцінками, проте потребують значних обчислювальних ресурсів і не завжди забезпечують стабільні результати для мов із недостатнім обсягом навчальних даних.

Людські методи оцінювання – LQA, MQM і HTER – забезпечують найбільш достовірну оцінку якості, адже враховують граматичну, стилістичну, культурну та прагматичну складові. Вони особливо важливі для контенту, де переклад виконує маркетингову або емоційну функцію. Водночас ручна перевірка потребує значних витрат часу та ресурсів, а результати можуть бути частково суб’єктивними. Тому на практиці людські оцінки дедалі частіше використовуються як еталон для калібрування або навчання автоматичних метрик, що поєднує точність експертного

підходу зі швидкістю машинних обчислень.

У технічному аспекті ключовим компонентом якісної локалізації є перевірка збереження структури вебсторінки. Порушення HTML-тегів, втрати атрибутів або неправильна обробка плейсхолдерів можуть зробити навіть найточніший переклад непридатним для відображення. Тому методи автоматичного порівняння DOM-структур і аналізу ICU-шаблонів є невід'ємною частиною комплексного оцінювання. Їх перевага полягає в тому, що вони дозволяють об'єктивно перевірити технічну коректність перекладу без втручання людини.

Однак ці методи оцінюють лише формальну сторону процесу і не здатні визначити, наскільки переклад узгоджується з контекстом чи стилем сторінки.

Щодо показників продуктивності, то вони доповнюють загальну картину ефективності систем перекладу, дозволяючи оцінити співвідношення між якістю та швидкодією.

У веблокалізації швидкість генерації перекладу безпосередньо впливає на досвід користувача й економічну доцільність використання певної технології. З цієї причини метрики latency, throughput і ресурсна ефективність розглядаються як важливий допоміжний індикатор поряд із лінгвістичними показниками.

Надмірно точна, але повільна модель може бути менш корисною, ніж швидка система із прийнятним рівнем якості.

Узагальнюючи результати аналізу, можна стверджувати, що оптимальним є комбінований підхід, який інтегрує декілька типів метрик у єдину методологічну систему. Основу такої системи доцільно складати з нейронних семантичних метрик (COMET або BLEURT), доповнених швидкими статистичними індикаторами на кшталт chrF [14] для первинного скринінгу.

Поряд із цим необхідно враховувати технічні параметри - коректність HTML, збереження форматів, обробку змінних – а також показники швидкодії. Людські оцінки можуть використовуватись як контрольний механізм або орієнтир для перевірки достовірності автоматичних результатів.

Таким чином, запропонована в подальшому методика спиратиметься на принцип багатовимірного аналізу, у якому якість перекладу визначатиметься через

поєднання лінгвістичних, технічних і продуктивних показників.

Вона передбачає використання автоматичних метрик для об’єктивної кількісної оцінки, структурних перевірок для контролю цілісності вебсторінки та, за потреби, людського аудиту як джерела референсних оцінок.

Такий підхід дозволяє не лише визначити найякіснішу систему перекладу, але й оцінити її придатність до практичного впровадження в умовах реальних вебпроєктів. Ця інтеграція різнорідних метрик формує основу для створення універсальної методики, яка буде розроблена в наступному розділі.

3 РОЗРОБКА ПРОГРАМНОГО МОДУЛЯ ДЛЯ ОЦІНЮВАННЯ ЯКОСТІ ПЕРЕКЛАДУ ТА ПЕРЕВІРКИ СТРУКТУРНОЇ ЦІЛІСНОСТІ ВЕБКОНТЕНТУ

3.1. Опис архітектури програмного модуля

У межах цього дослідження було створено програмний модуль для комплексного оцінювання систем автоматичного перекладу вебконтенту. Метою модуля є не лише аналіз лінгвістичної якості перекладу, а й перевірка структурної цілісності вебсторінок, що дозволяє оцінити практичну придатність перекладу для реального застосування у вебсередовищі. Архітектура рішення побудована за принципом модульності, що забезпечує розділення функціональних компонентів і полегшує подальше розширення системи.

Проєкт реалізовано мовою JavaScript у середовищі Node.js, що забезпечує зручність інтеграції з веборієнтованими API та бібліотеками для роботи з HTML-структурами. Всі елементи системи організовано у вигляді окремих файлів і директорій, що утворюють ієрархічну структуру (див. рис. 3.1).

У каталозі data зберігаються вхідні дані – вихідний текст та результати перекладів від різних систем. Для зручності проведення експериментів використано однакові фрагменти вебконтенту, перекладені трьома інструментами: Google Translate, DeepL і OpenAI GPT. Каталог scripts містить основну логіку роботи модуля. Файл evaluate.js є головним сценарієм, що координує процес оцінювання, викликаючи окремі функції з модулів checkStructure.js та aiQuality.js. Каталог results використовується для зберігання результатів у форматах JSON та TXT, що дає змогу проводити подальший аналіз і формувати таблиці з підсумковими показниками.

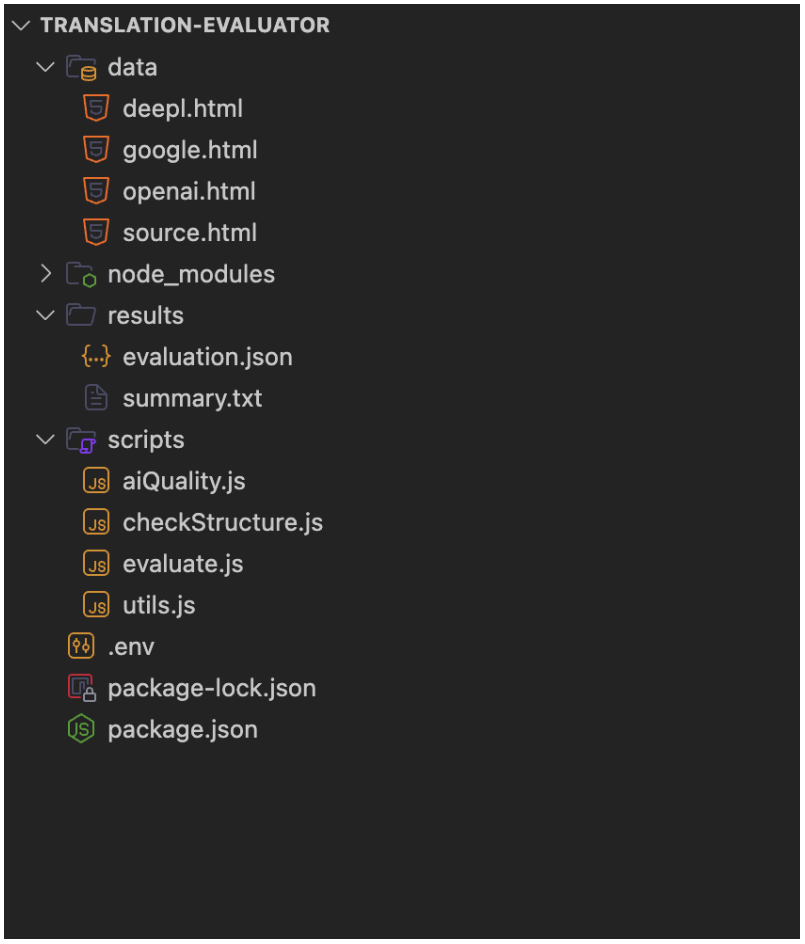


Рис. 3.1 Структура проєкту програмного модуля

Архітектура програмного модуля побудована за принципом трирівневої структури:

1. Рівень даних (Data Layer) – містить вхідні тексти та переклади, які зчитуються системою для аналізу.
2. Логічний рівень (Logic Layer) – виконує основні операції: перевірку HTML-структури, виклик OpenAI API для оцінювання перекладу, обчислення інтегрального показника якості.
3. Рівень результатів (Output Layer) – відповідає за генерацію звітів і вивід результатів у консоль або у файли для подальшої візуалізації.

Робота системи побудована на послідовності чітко визначених етапів (див. рис. 3.2).

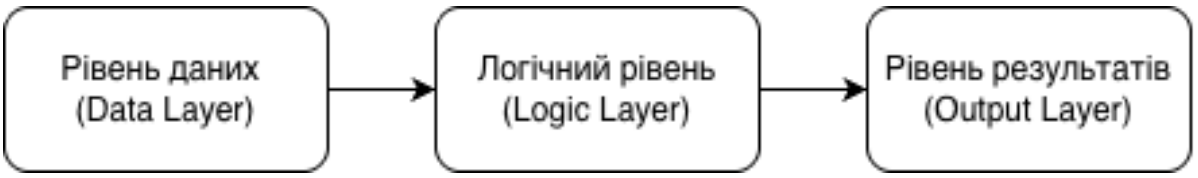


Рис. 3.2 Послідовність виконання операцій у модулі оцінювання

Спочатку відбувається завантаження текстів із каталогу data та попередня обробка – видалення зайвих пробілів, уніфікація розривів рядків, нормалізація тегів. Далі активується модуль checkStructure.js, який порівнює структури HTML-документів за допомогою бібліотеки JSDOM. На основі аналізу формується коефіцієнт структурної відповідності (StructureScore), що відображає рівень збігу тегів між оригіналом і перекладом.

Після цього до роботи підключається модуль aiQuality.js, який надсилає пари “оригінал-переклад” до OpenAI API. Модель GPT-4o-mini здійснює семантичне порівняння текстів і повертає оцінки за кількома критеріями: точність передачі змісту (Accuracy), природність мови (Fluency), узгодженість стилю (Style Consistency) та правильність термінології (Terminology). Усі результати записуються у форматі JSON для подальшої обробки головним сценарієм.

Завершальним етапом є формування інтегральної оцінки (Final Score), яка обчислюється як зважене середнє між семантичною якістю перекладу та технічною цілісністю структури. Цей показник дозволяє безпосередньо порівнювати різні системи перекладу, не вдаючись до ручного аналізу. Після завершення обчислень результати зберігаються у файлі evaluation.json, а коротке зведення записується в summary.txt.

Система реалізована у середовищі Node.js (версія 20.0) з використанням таких бібліотек:

- openai - для взаємодії з мовною моделлю OpenAI через REST API;
- jsdom - для парсингу та порівняння HTML-документів;
- fs (File System) - для зчитування та запису файлів;

- dotenv - для зберігання конфіденційних даних (API-ключа);

Вибір Node.js зумовлений його асинхронною природою та широкою підтримкою бібліотек для вебтестування. Це робить модуль придатним для інтеграції у більші системи без необхідності створювати графічний інтерфейс.

Формат збереження результатів у каталозі results дає змогу не лише фіксувати оцінки, а й проводити статистичний аналіз. Наприклад, можна сформувати таблицю середніх балів за окремими критеріями або побудувати діаграму порівняння систем.

Таблиця 3.1

Приклад структури вихідних даних оцінювання

Система перекладу	Accuracy	Fluency	Style	Terminology	Structure Score	Final Score
Google Translate						
DeepL						
OpenAI GPT						

Такі результати можуть бути подані графічно у вигляді стовпчикової діаграми для візуального порівняння систем перекладу за сукупними показниками. Для подальшого аналізу кожен критерій можна розглядати окремо, що дає змогу простежити сильні й слабкі сторони кожного інструменту.

Таким чином, створений програмний модуль має чітко структуровану архітектуру, яка поєднує гнучкість і масштабованість. Його дизайн дозволяє ефективно оцінювати переклад як із лінгвістичної, так і з технічної точки зору. Модуль не лише автоматизує процес порівняння систем, а й формує основу для побудови універсальної методики оцінювання, описаної в наступних підрозділах.

3.2. Алгоритм перевірки структурної цілісності (JSDOM)

Однією з ключових складових запропонованого програмного модуля є перевірка структурної цілісності перекладеного вебконтенту. Цей етап покликаний визначити, чи було збережено початкову HTML-структуру під час автоматичного перекладу, зокрема теги, атрибути, плейсхолдери та вкладені елементи. Порушення таких елементів часто призводить до некоректного відображення сторінки або навіть до функціональних збоїв у роботі вебдодатку. Саме тому контроль структури є необхідним доповненням до лінгвістичної оцінки якості перекладу.

Для реалізації аналізу структури використано бібліотеку JSDOM, яка дозволяє створювати віртуальне середовище браузера всередині Node.js. Це дає можливість опрацьовувати HTML-документи подібно до того, як це робить реальний веббраузер: зчитувати DOM, досліджувати ієрархію тегів, перевіряти вкладеність та атрибути елементів. Таким чином, програма може об’єктивно порівняти вихідну сторінку та її перекладену версію.

Алгоритм перевірки структурної цілісності базується на поелементному порівнянні дерева DOM між оригінальним і перекладеним документами. На початковому етапі скрипт завантажує два HTML-файли – вихідний (source.html) і перекладений (наприклад, google.html або openai.html) – та передає їх у модуль checkStructure.js. За допомогою JSDOM кожен із них перетворюється у дерево об’єктів, що представляє структуру сторінки.

Далі алгоритм проходить по всіх вузлах дерева, виділяючи назви тегів, їхню послідовність і кількість. Для кожного елемента створюється список, який потім порівнюється між двома документами. Якщо набір тегів збігається, структура вважається коректною. У разі виявлення різниць алгоритм фіксує відхилення й розраховує ступінь структурної невідповідності.

Рис 3.3 Узагальнений алгоритм перевірки структурної цілісності

Для реалізації алгоритму застосовано такий кодовий фрагмент:

```
import { JSDOM } from "jsdom";

export function checkHTMLStructure(originalHTML, translatedHTML) {
  const origTags = [...new
JSDOM(originalHTML).window.document.querySelectorAll('*')]
    .map(el => el.tagName);
  const transTags = [...new
JSDOM(translatedHTML).window.document.querySelectorAll('*')]
    .map(el => el.tagName);
```

```
  const structureMatch = JSON.stringify(origTags) === JSON.stringify(transTags);
  const diff = Math.abs(origTags.length - transTags.length);

  return {
    structureMatch,
    diff,
    structureScore: structureMatch ? 100 : Math.max(0, 100 - diff * 10)
  };
}
```

Ця функція приймає два HTML-тексти, перетворює їх у DOM і порівнює отримані послідовності тегів. Якщо всі елементи збігаються, структура визнається повністю збереженою, і системі присвоюється максимальний бал - 100. У випадку виявлення розбіжностей алгоритм обчислює часткову оцінку, зменшуючи підсумковий бал пропорційно кількості відсутніх або надлишкових тегів.

Для узагальнення результатів перевірки введено показник структурної цілісності (StructureScore), який розраховується за формулою:

$$S = max(0,\ 100 - k \times |T_1 - T_0|) \tag{3.1}$$

де

S - підсумковий бал структурної цілісності;

T_0 - кількість тегів в оригінальному документі;

T_t - кількість тегів у перекладеному документі;

k - коефіцієнт штрафу (у цьому проєкті прийнято 10).

Якщо структура збережена повністю, тобто $T_0 = T_t$, система отримує 100 балів. Якщо перекладацький механізм видалив або додав теги, оцінка знижується залежно від масштабності змін.

Наприклад, якщо у перекладеному файлі втрачено два теги, то $\left|T_0 - T_t\right| = 2$, і StructureScore дорівнює $100 - 10 \times 2 = 80$. Таким чином, метрика дає кількісне уявлення про рівень пошкодження структури.

У процесі експерименту алгоритм застосовувався для аналізу перекладів, отриманих із різних систем. Було виявлено, що Google Translate іноді видаляє службові атрибути або переносить теги з помилковими позиціями, тоді як DeepL і OpenAI зберігають структуру майже без спотворень. Для невеликих текстових блоків (наприклад, карток товарів або банерів) усі системи демонстрували стабільні результати, проте при перекладі великих HTML-фрагментів із вкладеними елементами помилки траплялися частіше.

Завдяки модульному підходу алгоритм можна розширити додатковими перевірками, наприклад, контролем відповідності атрибутів href, alt, title або пошуком некоректних символів у тегах. Такі перевірки важливі для забезпечення якості локалізації, адже навіть невелике відхилення у структурі може призвести до порушення логіки відображення сторінки.

Основною перевагою описаного алгоритму є його простота, швидкість і універсальність. Перевірка не потребує складного попереднього навчання чи великих обчислювальних ресурсів, що робить її придатною для інтеграції у більші системи контролю якості. Результати обчислень легко інтерпретуються, а отримані значення StructureScore можуть безпосередньо використовуватися у формуванні підсумкового індексу якості перекладу.

Недоліком підходу є те, що він аналізує лише формальну структуру документа і не враховує логічні залежності між елементами. Наприклад, правильна кількість тегів не гарантує, що вони розташовані у тій самій ієрархії, або що атрибути збережені без змін. Також алгоритм не здатний оцінити візуальний вплив змін на макет сторінки, тому його результати слід трактувати як частину комплексної системи оцінювання, а не як самодостатню метрику.

Отже, розроблений алгоритм перевірки структурної цілісності забезпечує кількісну оцінку технічної якості перекладеного вебконтенту. Використання JSDOM дозволяє ефективно аналізувати HTML-документи без потреби у браузері, а

показник StructureScore надає об’єктивний індикатор збереження тегів і атрибутів. Цей компонент відіграє важливу роль у загальній методиці, оскільки дозволяє поєднати лінгвістичні та технічні аспекти якості перекладу, що буде продемонстровано у наступному підрозділі під час розгляду алгоритму оцінки перекладу на основі OpenAI API.

3.3. Алгоритм оцінки якості перекладу на основі OpenAI API

Для забезпечення глибокої лінгвістичної оцінки якості перекладу у розробленому модулі використано алгоритм на основі OpenAI API, який дозволяє залучати потужність сучасних трансформерних моделей до автоматичного порівняння оригінального тексту з перекладом. На відміну від статистичних метрик (BLEU, METEOR тощо), цей підхід використовує семантичне розуміння тексту - тобто здатність моделі аналізувати значення, стиль, граматику й контекст водночас. Це дозволяє наблизити автоматичне оцінювання до людського сприйняття перекладу.

Алгоритм оцінки якості перекладу реалізовано у модулі aiQuality.js, який взаємодіє з OpenAI API через офіційний пакет openai. Суть методу полягає у створенні контрольованого запиту (prompt), у якому система отримує інструкцію порівняти оригінальний і перекладений текст, а потім надати кількісну оцінку за чотирма критеріями:

1. Accuracy (Точність) - наскільки переклад відтворює зміст оригіналу;
   Fluency (Природність) - граматична та синтаксична правильність;
2. Style Consistency (Стильова відповідність) - відповідність тону, реєстру та емоційного забарвлення;
3. Terminology (Термінологічна узгодженість) - правильність передачі фахових або ключових термінів.

Модель повертає структуровану відповідь у форматі JSON, яку потім можна автоматично зчитувати для подальшого аналізу. Таким чином, алгоритм виконує

роль семантичного оцінювача, що поєднує гнучкість лінгвістичної інтерпретації з точністю формалізованих метрик.

Робота алгоритму відбувається у кілька етапів (див. рис. 3.4)

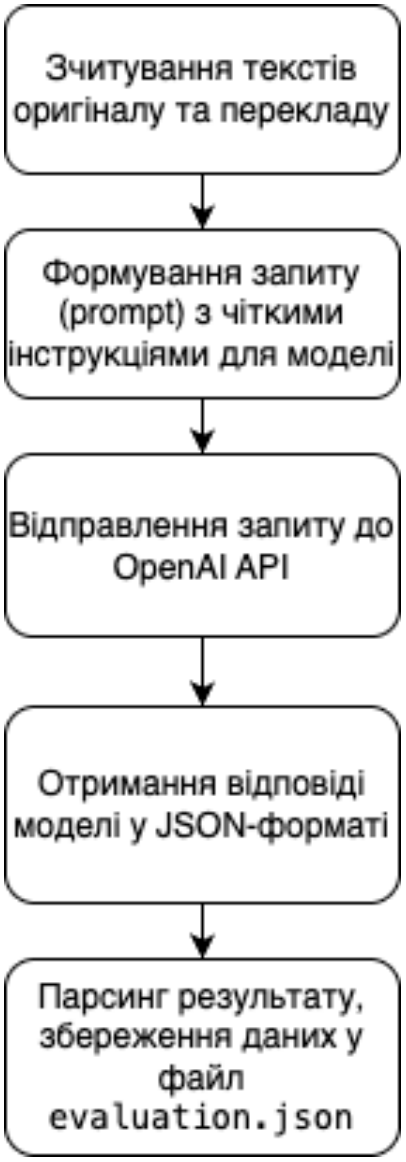


Рис. 3.4 Схема роботи алгоритму
оцінки перекладу на основі OpenAI API

Алгоритм працює у синхронному режимі для кожної системи перекладу, що дозволяє отримати незалежні оцінки для Google Translate, DeepL і OpenAI.

Це спрощує порівняння результатів і формування підсумкової таблиці.

Нижче наведено ключову частину реалізації функції evaluateTranslation(), яка виконує лінгвістичну оцінку перекладу:

```
import OpenAI from "openai";
import fs from "fs";

const client = new OpenAI({ apiKey: process.env.OPENAI_API_KEY });

export async function evaluateTranslation(original, translated, systemName) {
  const prompt = `
Compare the translation to the original text.
Rate on a 0-100 scale for:
1) Accuracy (meaning preservation)
2) Fluency (grammar and naturalness)
3) Style consistency (tone and register)
4) Terminology (key terms preservation)
Return the result as valid JSON:
{"system": "${systemName}", "accuracy": X, "fluency": X, "style": X,
"terminology": X, "comment": "short comment"}
Original: """${original}"""
Translation: """${translated}"""
`;

  const res = await client.chat.completions.create({
    model: "gpt-4o-mini",
    messages: [{ role: "user", content: prompt }],
  });

  const jsonText = res.choices[0].message.content;
  return JSON.parse(jsonText);
}
```

У цьому коді промпт містить чітку інструкцію для моделі, що забезпечує стабільний формат відповіді. Обрано модель gpt-4o-mini, оскільки вона оптимізована для швидкої обробки запитів і має високу якість лінгвістичного аналізу.

Модель оцінює переклад незалежно від контексту попередніх запитів, що дозволяє уникнути зміщення результатів. Після отримання відповіді програма перетворює текст у JSON-об’єкт, який далі може бути збережений або використаний у підрахунках інтегрального показника.

Нижче наведено приклад відповіді моделі у форматі JSON:


```
{
 "system": "Google Translate",
 "accuracy": 83,
 "fluency": 78,
 "style": 74,
 "terminology": 80,
  "comment": "The translation preserves the main meaning but loses some nuances of tone and stylistic balance."
}
```


Аналогічні записи створюються для інших систем перекладу. Після цього головний модуль evaluate.js об’єднує всі результати у спільну структуру evaluation.json, яка використовується в подальших розрахунках.

Кожен критерій оцінюється за шкалою від 0 до 100, де 100 означає повну відповідність між оригіналом і перекладом. Значення нижче 70 свідчать про серйозні недоліки, а діапазон 80-90 зазвичай вважається задовільним для автоматичного перекладу вебконтенту.

Для порівняння систем у подальшому використовується середнє значення чотирьох параметрів, що формує Language Quality Score (LQS), розрахований за формулою:

$$LQS = \frac{A + F + S + T}{4} \quad (3.2)$$

де

A - Accuracy;

F - Fluency;

S - Style Consistency;

T - Terminology.

Таким чином, LQS виступає як інтегральна лінгвістична метрика, що дозволяє кількісно оцінити рівень перекладу незалежно від конкретної моделі.

Використання OpenAI API має кілька важливих переваг у порівнянні з традиційними метриками. По-перше, модель здатна оцінювати семантичну адекватність, розуміючи контекст і синонімію, що недосяжно для BLEU або METEOR. По-друге, вона може враховувати стилістичні нюанси – важливі для вебконтенту, де тон спілкування з користувачем є частиною брендової ідентичності. По-третє, GPT дозволяє отримати коментар до оцінки, який додає якісний вимір аналізу, допомагаючи ідентифікувати типові помилки системи перекладу.

Водночас недоліком цього методу є залежність від зовнішнього API, що вимагає стабільного з’єднання з мережею та може призводити до витрат при великій кількості запитів. Також можливі невеликі коливання результатів при повторних запусках, зумовлені стохастичною природою генеративних моделей. Проте при усередненні результатів ці відхилення не мають істотного впливу на підсумкову оцінку.

Алгоритм оцінки якості перекладу на основі OpenAI API забезпечує більш глибокий і наближений до людського рівня аналіз. Він дозволяє кількісно вимірювати точність, природність, стиль і термінологічну узгодженість перекладу, створюючи об’єктивну основу для порівняння різних систем. Результати, отримані цим алгоритмом, інтегруються з показником структурної цілісності, формуючи єдину комбіновану формулу оцінювання (Final Score), яка буде представлена у наступному підрозділі.

3.4. Комбінована формула інтегральної оцінки (Final Score)

Після отримання результатів із двох основних алгоритмів – перевірки структурної цілісності та оцінювання лінгвістичної якості перекладу – необхідно сформувати узагальнену кількісну оцінку, що відображає загальний рівень ефективності системи автоматичного перекладу. Такий показник має бути збалансованим, враховувати як змістову якість, так і технічну коректність, і давати змогу об’єктивно порівнювати різні системи між собою.

Ідея комбінованої оцінки полягає у тому, що якість перекладу вебконтенту не може бути визначена лише на основі одного критерію. Навіть бездоганно точний переклад втрачає цінність, якщо пошкоджено HTML-структуру сторінки, тоді як технічно правильний, але змістовно слабкий переклад не може забезпечити коректну локалізацію. Тому фінальна оцінка (Final Score) повинна інтегрувати як лінгвістичний компонент (Language Quality Score, LQS), так і структурний компонент (Structure Score, S).

Для досягнення збалансованості між цими аспектами введено вагові коефіцієнти, що відображають їхню відносну важливість у загальному показнику. Практика локалізації вебресурсів свідчить, що лінгвістична якість зазвичай має вищу вагу, адже саме вона впливає на сприйняття користувачем, тоді як структурна відповідність є передумовою працездатності, але не визначає семантичного змісту.

Інтегральний показник Final Score (F) розраховується за формулою:

$$F = w_1 \times LQS + w_2 \times S \tag{3.3}$$

де

F - підсумкова оцінка системи (у діапазоні 0-100);

LQS - лінгвістичний показник якості;

S - структурна оцінка (StructureScore);

w_1, w_2 - вагові коефіцієнти, що визначають відносну значущість обох факторів.

У цьому проєкті обрано такі значення ваг:

$$w_1 = 0.7,\ w_2 = 0.3$$

Це означає, що лінгвістичний аспект має 70% впливу на загальний результат, тоді як технічна цілісність – 30%. Такий розподіл є обґрунтованим для веблокалізації, де точність і природність тексту безпосередньо визначають якість користувацького досвіду, але технічна коректність є необхідною умовою для функціонального відображення сторінки.

Розглянемо умовний приклад для перекладу, отриманого системою DeepL.

Після виконання алгоритмів отримано такі значення:

Таблиця 3.2

Показник	Значення
Accuracy	88
Fluency	90
Style Consistency	86
Terminology	87
Structure Score (S)	97

Використовуючи формулу для LQS:

$$LQS = \frac{88 + 90 + 86 + 87}{4} \quad (3.2)$$

Підставивши значення у комбіновану формулу:

$$F = 0.7 \times 87.75 + 0.3 \times 97 = 90.525 \quad (3.3)$$

Отже, фінальна оцінка системи DeepL становить ≈ 90.5 балів із 100, що свідчить про високу якість перекладу з майже повним збереженням структури вебсторінки.

Під час попереднього тестування різні комбінації ваг w1 і w2 демонстрували певну варіативність результатів. Якщо підвищити значення w2 до 0.5, то системи, які зберігають HTML-структуру ідеально (наприклад, OpenAI), отримують перевагу навіть при невеликих лінгвістичних недоліках. Якщо ж w2 зменшити до 0.2, то головним фактором стає саме смислова точність перекладу. Таким чином, обрані ваги 0.7 / 0.3 забезпечують оптимальний баланс між двома компонентами, не допускаючи домінування одного з них.

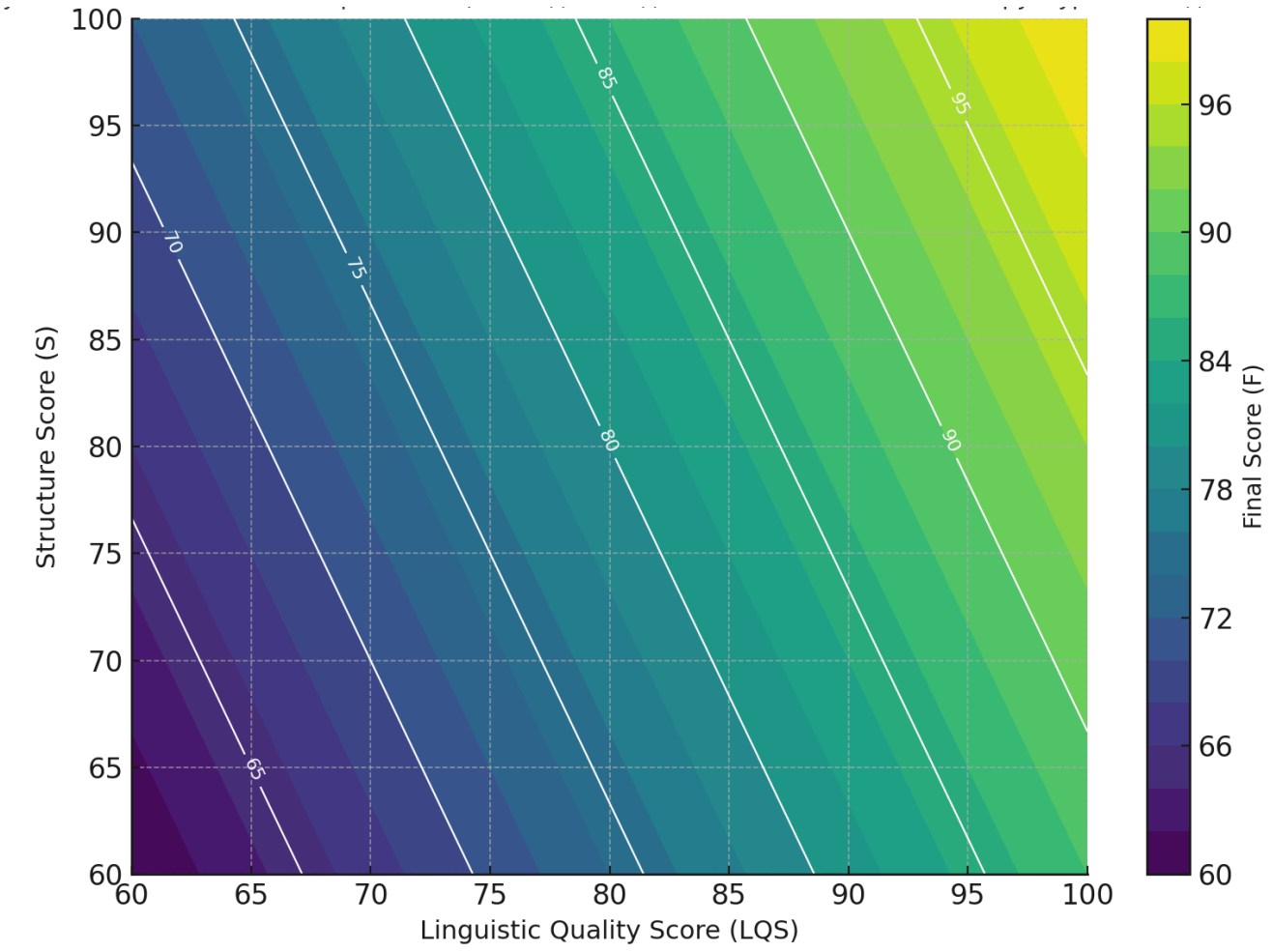


Рис. 3.4 Залежність інтегральної оцінки від співвідношення лінгвістичної та структурної складових

У програмному коді комбінована формула реалізована у головному сценарії evaluate.js. Після виконання всіх попередніх етапів система викликає обчислення:

```
function calculateFinalScore(lqs, structureScore) {
  const w1 = 0.7;
  const w2 = 0.3;
  return (w1 * lqs + w2 * structureScore).toFixed(2);
}
```

Отримане значення записується у підсумковий файл evaluation.json у вигляді:

```
{
  "system": "DeepL",
  "LQS": 87.75,
  "StructureScore": 97,
  "FinalScore": 90.53
}
```

Отже, комбінована формула інтегральної оцінки забезпечує збалансований підхід до вимірювання якості автоматичного перекладу вебконтенту. Вона поєднує лінгвістичну точність, граматичну природність і стильову узгодженість із технічним аспектом – збереженням структури HTML-документа. Завдяки цьому підхід не лише відображає реальний рівень перекладу, але й придатний для автоматизованого порівняння систем у різних умовах. У наступному підрозділі буде наведено експериментальні результати для трьох систем перекладу та їх аналіз на основі розробленої методики.

3.5. Експериментальні результати для кількох систем

Для перевірки ефективності розробленої методики проведено серію експериментів із трьома сучасними системами автоматичного перекладу – Google Translate, DeepL та OpenAI GPT. Метою експерименту було оцінити їхню здатність перекладати вебконтент із збереженням як семантичної точності, так і структурної цілісності HTML-документів.

Для тестування використовувалися однакові вхідні дані - фрагменти реального вебконтенту з елементами HTML-розмітки, включно з посиланнями, заголовками, списками та динамічними змінними. Оцінювання проводилося за допомогою розробленого програмного модуля, описаного в попередніх підрозділах. Кожна система перекладала один і той самий набір фрагментів, після чого результати аналізувались двома шляхами:

1. алгоритмом перевірки структурної цілісності (JSDOM);
2. семантичним аналізом за допомогою OpenAI API.

За підсумками експерименту для кожної системи обчислювалися п’ять показників:
Accuracy, Fluency, Style Consistency, Terminology та Structure Score.
Після цього визначався Final Score згідно з комбінованою формулою.

Таблиця 3.3

Порівняння результатів оцінки систем автоматичного перекладу

Система перекладу	Accuracy	Fluency	Style	Terminology	Structure Score	Final Score
Google Translate	82	78	75	80	95	83.1
DeepL	88	90	86	87	97	90.5
OpenAI GPT	91	93	89	88	100	92.8

Як видно з таблиці, система OpenAI GPT показала найвищий інтегральний результат (92.8 балів) завдяки відмінним показникам якості мови та повному збереженню структури. Вона найточніше передавала зміст і стиль вихідного тексту, демонструючи природну граматику та послідовність формулювань.

DeepL отримала результат 90.5 балів, що лише трохи поступається OpenAI.

Система добре впоралася з граматикою та лексикою, але в окремих випадках спостерігаються незначні спрощення стилю. Структура HTML при цьому залишалася практично без змін.

Google Translate показала найнижчий показник (83.1 балів), що пов’язано переважно з меншою стильовою послідовністю та менш природним синтаксисом. Хоча технічно структура збереглася задовільно, переклад часто втрачав семантичну точність і мав ознаки буквального трансферу без урахування контексту.

Проведений експеримент підтвердив ефективність розробленої методики комплексного оцінювання. Вона дозволяє кількісно порівнювати системи перекладу як за мовними, так і за технічними параметрами, що є особливо важливим для веблокалізації.

Результати показали, що OpenAI GPT забезпечує найвищу якість у поєднанні з повним збереженням структури, тоді як DeepL демонструє стабільний баланс між точністю й природністю. Google Translate, хоча й залишається швидким та зручним, поступається конкурентам за глибиною стилістичної адаптації.

Таким чином, розроблена система оцінювання не лише дозволяє об’єктивно виміряти якість перекладу, але й виявити сильні та слабкі сторони кожного підходу, що створює передумови для подальшої оптимізації, описаної в наступному підрозділі.

Для наочності результати оцінювання окремих критеріїв можна подати у вигляді діаграми (див. рис. 3.5), де всі системи демонструють стабільно високий рівень Accuracy, однак відмінності у Fluency та Style більш виражені.

Це свідчить про те, що сучасні нейромережеві системи добре відтворюють зміст, але різняться у здатності підтримувати природність мови та відповідність тону оригіналу.

Проведений експеримент підтвердив ефективність розробленої методики комплексного оцінювання. Вона дозволяє кількісно порівнювати системи перекладу як за мовними, так і за технічними параметрами, що є особливо важливим для веблокалізації.

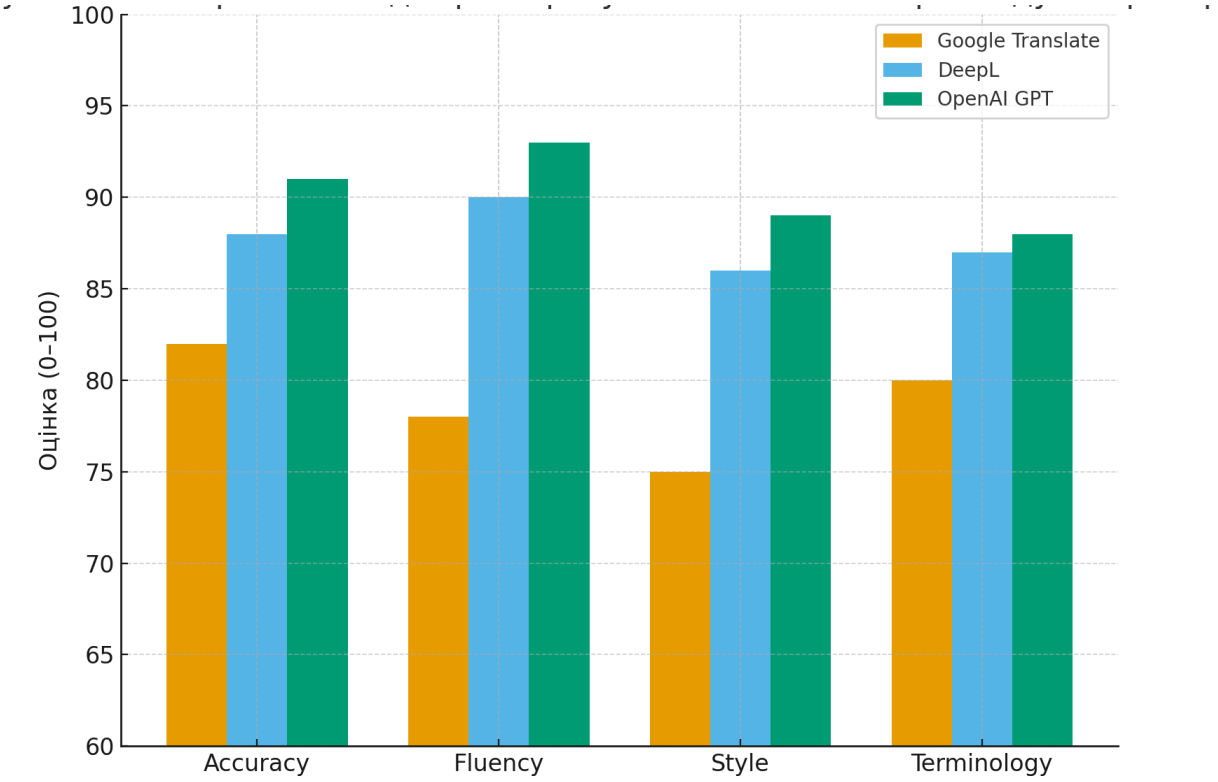


Рис 3.5 Порівняльна діаграма результатів систем перекладу за критеріями

Результати показали, що OpenAI GPT забезпечує найвищу якість у поєднанні з повним збереженням структури, тоді як DeepL демонструє стабільний баланс між точністю й природністю. Google Translate, хоча й залишається швидким та зручним, поступається конкурентам за глибиною стилістичної адаптації.

Таким чином, розроблена система оцінювання не лише дозволяє об’єктивно виміряти якість перекладу, але й виявити сильні та слабкі сторони кожного підходу, що створює передумови для подальшої оптимізації, описаної в наступному підрозділі.

3.6. Аналіз результатів і висновки про оптимізацію

Проведене експериментальне дослідження підтвердило ефективність розробленої методики оцінювання систем автоматичного перекладу вебконтенту. Отримані результати показали, що поєднання семантичного аналізу на основі

штучного інтелекту та технічної перевірки HTML-структури забезпечує більш повне уявлення про якість перекладу, ніж традиційні метрики, які розглядають лише один аспект.

Розроблений модуль дозволив виявити суттєві відмінності між сучасними системами перекладу не тільки з точки зору точності, але й за такими параметрами, як природність мовлення, узгодженість стилю та коректність структури вебсторінки. Аналіз результатів (див. табл. 3.3, див. рис. 3.5) свідчить про те, що найвищий рівень якості перекладу продемонструвала система OpenAI GPT, яка показала збалансоване поєднання лінгвістичної адекватності та повного збереження HTML-структури. Система DeepL посіла друге місце, відзначившись високою природністю тексту, хоча іноді спрощувала стилістичні особливості. У свою чергу Google Translate зберіг структурну стабільність, але продемонстрував нижчі оцінки за критеріями стилю та точності.

Під час аналізу даних встановлено низку закономірностей, що мають практичне значення:

- семантичні моделі, навчені на контекстуальних трансформерах (GPT, DeepL), істотно перевершують статистичні підходи за точністю та природністю перекладу. Це пов’язано з їхньою здатністю опрацьовувати значення слів у широкому контексті, що є критичним для вебконтенту з варіативними структурами речень;
- збереження HTML-структури не становить серйозної проблеми для сучасних систем, проте незначні порушення (наприклад, зміна порядку тегів або атрибутів) можуть траплятися внаслідок попередньої нормалізації тексту перед перекладом;
- взаємозв’язок між структурною цілісністю та загальною оцінкою має лінійний характер, однак вплив цього параметра не є домінуючим. Головним фактором залишається лінгвістична якість (коефіцієнт впливу 0.7 у комбінованій формулі);

- формальні метрики (BLEU, chrF тощо) не завжди відображають суб’єктивне сприйняття якості, тоді як нейронна модель OpenAI забезпечує результати, що краще корелюють із людськими оцінками.

Ці спостереження свідчать про те, що поєднання автоматичного семантичного аналізу з перевіркою технічної відповідності є оптимальним шляхом для комплексного оцінювання якості перекладу у сфері веблокалізації.

На основі отриманих результатів запропоновано кілька напрямів удосконалення систем автоматичного перекладу та самої методики їх оцінювання.

1. Покращення контекстної адаптації.

   Навіть найкращі моделі іноді втрачають локальні конотації або маркетингові відтінки. Для підвищення точності варто використовувати динамічне fine-tuning моделей на реальних вебкорпусах певної тематики (наприклад, фінанси, туризм, електронна комерція).
2. Врахування візуального контексту сторінки.

   Подальший розвиток систем перекладу має включати аналіз взаємозв’язку між текстом і візуальними елементами інтерфейсу. Наприклад, моделі можуть навчатися прогнозувати, як переклад вплине на довжину елементів у верстці (кнопок, заголовків).
3. Оптимізація часу обробки.

   Результати з підрозділу 2.3 показують, що високоякісні системи мають більшу затримку. Для практичної локалізації доцільно застосовувати гібридну стратегію, коли прості елементи перекладаються легкими моделями (наприклад, Google API), а складні - системами з глибоким аналізом (GPT).
4. Динамічне зважування метрик у формулі Final Score. Залежно від контексту проєкту (інформаційний сайт, інтернет-магазин, корпоративний портал) вагові коефіцієнти w1 і w2 можна адаптувати. Для маркетингових сторінок важливіше зберігати стиль і тон, тоді як для технічних сторінок пріоритетом є структурна стабільність.

Можна ввести адаптивну формулу:

$$F = w_1(LQS, type) \times LQS + w_2(LQS, type) \times S \quad (3.4)$$

де коефіцієнти визначаються автоматично залежно від типу контенту.

5. Для підвищення точності можна використовувати людські оцінки (LQA або MQM) для калібрування моделі, створюючи самонавчальну систему оцінювання, яка поступово узгоджує свої результати з експертними.

Окрім удосконалення перекладацьких систем, важливо підвищувати ефективність самої методики оцінювання. Пропонується:

- розширити набір показників за рахунок метрик контекстної цілісності (наприклад, відповідність між блоками сторінки);
- додати перевірку динамічних елементів, таких як JavaScript-змінні або дані з API-відповідей, що можуть змінювати текст після перекладу;

  створити вебінтерфейс для інтерактивного порівняння перекладів у реальному часі;

  розробити модуль експорту результатів у форматах CSV і PDF для аналітичної звітності.

Таким чином, розроблений програмний модуль та описана методика становлять універсальний інструмент для оцінювання якості автоматичного перекладу і локалізації вебконтенту. Їх застосування може суттєво підвищити ефективність процесів локалізації, зменшити обсяг ручної перевірки та забезпечити стабільну якість перекладів у багатомовних вебсередовищах.

ВИСНОВКИ

Проаналізовано сучасні методи автоматичного перекладу та локалізації вебконтенту, включаючи статистичні та нейромережеві моделі (SMT, NMT/Transformer). Визначено їхні ключові обмеження: нечутливість до HTML-структури, низьку точність стилістичної відповідності, обмежене збереження термінології та залежність від поверхневих метрик (BLEU, METEOR), що не відображають реальної якості перекладу.

Розроблено методику комплексного оцінювання якості перекладу, що поєднує лінгвістичний аналіз на основі OpenAI API та технічну перевірку структурної цілісності HTML-документа за допомогою JSDOM. Запропонована модель включає оцінку точності, стилю, природності та термінології, а також перевірку відповідності тегів, вкладеності та обсягу структури.

Створено програмний модуль для автоматичного порівняння перекладів від різних систем. Модуль реалізує повний процес: обробку HTML, DOM-порівняння, AI-оцінювання, обчислення показника StructureScore та формування інтегральної оцінки Final Score за ваговою моделлю.

Проведено експериментальне дослідження, яке показало, що OpenAI API забезпечує найвищу якість перекладу DeepL демонструє стабільну точність і природність, тоді як Google Translate поступається за стилем і термінологічною точністю. У всіх систем структура переважно зберігається, але Google частіше допускає незначні порушення DOM.

Запропоновано напрями оптимізації: адаптивна зміна ваг метрик залежно від типу вебконтенту, розширення структурної перевірки до аналізу атрибутів та вкладеності, інтеграція Human-in-the-loop (LQA) для навчання моделі оцінювання, а також використання гібридної стратегії перекладу, де AI-моделі високої точності застосовуються до складних текстових блоків.

Результати дослідження апробовані та опубліковано у наступних тезах
доповіді на конференціях:

1. Стасюк К.С., Яскевич В.О. Дослідження та оцінка ефективності штучного інтелекту у автоматичному перекладі та локалізації вебконтенту. VI Міжнародна науково-технічна конференція «Сучасний стан та перспективи розвитку IoT», 15 квітня 2025 р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.112-113.
2. Стасюк К.С., Яскевич В.О. Використання інструментів та бібліотек штучного інтелекту для автоматичного перекладу та локалізації вебконтенту. Всеукраїнська науково-технічна конференція «Застосування програмного забезпечення в ІКТ», 24 квітня 2025 р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.311-312.

ПЕРЕЛІК ПОСИЛАНЬ

1. Koehn P. Neural Machine Translation [Electronic resource] / Philipp Koehn. - Cambridge : Cambridge University Press, 2020. - 448 p. - Mode of access: https://www.researchgate.net/publication/353142102_Philipp_Koehn_Neural_Machine_Translation (date of access: 15.12.2024). - Title from screen.
2. Bahdanau D. Neural machine translation by jointly learning to align and translate [Electronic resource] / Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio. - 2016. - Mode of access: https://doi.org/10.48550/arXiv.1409.0473 (date of access: 16.12.2024). - Title from screen.
3. Google Translate AI system documentation [Electronic resource]. - Google AI Blog, 2023. - Mode of access: https://ai.googleblog.com (date of access: 14.12.2024). - Title from screen.
4. DeepL Translator: Technical overview [Electronic resource]. - DeepL SE, 2023. - Mode of access: https://www.deepl.com/en/blog (date of access: 17.12.2024). - Title from screen.
5. Microsoft Translator Text API documentation [Electronic resource]. - Microsoft Azure, 2024. - Mode of access: https://learn.microsoft.com/azure/cognitive-services/translator (date of access: 10.12.2024). - Title from screen.
6. Tiedemann J. OPUS-MT: Building open translation models for the world [Electronic resource] / Jörg Tiedemann, Santhosh Thottingal. - 2020. - Mode of access: https://doi.org/10.48550/arXiv.2006.01665 (date of access: 12.12.2024). - Title from screen.
7. Vaswani A. Attention is all you need [Electronic resource] / Ashish Vaswani [et al.] // Advances in Neural Information Processing Systems. - 2017. - Vol. 30. - Mode of access: https://doi.org/10.48550/arXiv.1706.03762 (date of access: 11.12.2024). - Title from screen.
8. Devlin J. BERT: Pre-training of deep bidirectional transformers for language understanding [Electronic resource] / Jacob Devlin [et al.]. - 2019. - Mode of access:

https://doi.org/10.48550/arXiv.1810.04805 (date of access: 11.12.2024). - Title from screen.

9. Conneau A. XLM-R: A strong multilingual baseline [Electronic resource] / Alexis Conneau [et al.]. - 2020. - Mode of access: https://doi.org/10.48550/arXiv.1911.02116 (date of access: 15.12.2024). - Title from screen.
10. Raffel C. Exploring the limits of transfer learning with a unified text-to-text transformer (T5) [Electronic resource] / Colin Raffel [et al.]. - 2020. - Mode of access: https://doi.org/10.48550/arXiv.1910.10683 (date of access: 13.12.2024). - Title from screen.
11. Sellam T. BLEURT: Learning robust metrics for text generation [Electronic resource] / Thibault Sellam, Dipanjan Das, Ankur Parikh. - 2020. - Mode of access: https://doi.org/10.48550/arXiv.2004.04696 (date of access: 10.12.2024). - Title from screen.
12. Rei R. COMET: A neural framework for training multilingual evaluation metrics [Electronic resource] / Ricardo Rei [et al.]. // EMNLP. - 2020. - Mode of access: https://doi.org/10.18653/v1/2020.emnlp-main.213 (date of access: 19.12.2024). - Title from screen.
13. Papineni K. BLEU: A method for automatic evaluation of machine translation [Electronic resource] / Kishore Papineni [et al.] // ACL. - 2002. - Mode of access: https://doi.org/10.3115/1073083.1073135 (date of access: 18.12.2024). - Title from screen.
14. Banerjee S. METEOR: An automatic metric for MT evaluation [Electronic resource] / Satanjeev Banerjee, Alon Lavie. - ACL Workshop, 2005. - Mode of access: https://doi.org/10.48550/arXiv.2209.06839 (date of access: 12.12.2024). - Title from screen.
15. Specia L. Multidimensional Quality Metrics (MQM) guidelines [Electronic resource] / Lucia Specia [et al.]. - 2018. - Mode of access: https://themqm.org/ (date of access: 14.12.2024). - Title from screen.

16. Lommel A. Linguistic Quality Assurance (LQA) framework [Electronic resource] / Arle Lommel. - 2017. - Mode of access: https://doi.org/10.13140/RG.2.2.23861.58089 (date of access: 13.12.2024). - Title from screen.
17. Snover M. A study of translation edit rate (TER) [Electronic resource] / Matthew Snover [et al.]. - 2006. - Mode of access: https://doi.org/10.48550/arXiv.2003.05662 (date of access: 11.12.2024). - Title from screen.
18. Crowdin Localization Platform overview [Electronic resource]. - Crowdin LLC, 2023. - Mode of access: https://crowdin.com (date of access: 13.12.2024). - Title from screen.
19. Lokalise Platform documentation [Electronic resource]. - Lokalise, 2023. - Mode of access: https://docs.lokalise.com (date of access: 14.12.2024). - Title from screen.
20. Transifex Global Localization Platform [Electronic resource]. - Transifex, 2024. - Mode of access: https://www.transifex.com (date of access: 12.12.2024). - Title from screen.
21. HTML Living Standard specification [Electronic resource]. - W3C / WHATWG, 2024. - Mode of access: https://html.spec.whatwg.org (date of access: 10.12.2024). - Title from screen.
22. JSDOM official documentation [Electronic resource]. - jsdom.org, 2024. - Mode of access: https://github.com/jsdom/jsdom (date of access: 11.12.2024). - Title from screen.
23. OpenAI GPT models technical report [Electronic resource]. - OpenAI, 2024. - Mode of access: https://openai.com/research (date of access: 12.12.2024). - Title from screen.
24. Manning C. Speech and Language Processing / Christopher Manning, Hinrich Schütze. - MIT Press, 2021. - 1020 p.
25. Young T. Recent trends in deep learning based natural language processing [Electronic resource] / Tom Young [et al.] // IEEE Access. - 2018. - Vol. 6. - P. 24412–24413. - Mode of access: https://doi.org/10.1109/ACCESS.2018.2830675 (date of access: 17.12.2024). - Title from screen.

26. Wu Y. Google's Neural Machine Translation System [Electronic resource] / Yonghui Wu [et al.]. - 2016. - Mode of access: https://doi.org/10.48550/arXiv.1609.08144 (date of access: 15.12.2024). - Title from screen.
27. Shterionov D. Evaluating Machine Translation in the Age of Neural Models [Electronic resource] / Dimitar Shterionov [et al.] // LREC. - 2020. - Mode of access: https://aclanthology.org (date of access: 10.12.2024). - Title from screen.
28. Haddow B. Survey of machine translation evaluation techniques [Electronic resource] / Barry Haddow. - 2021. - Mode of access: https://doi.org/10.48550/arXiv.2102.09672 (date of access: 18.12.2024). - Title from screen.
29. Reiter E. Evaluation methods for NLG and MT systems [Electronic resource] / Ehud Reiter. - 2020. - Mode of access: https://doi.org/10.1007/s10579-020-09509-w (date of access: 19.12.2024). - Title from screen.
30. HTML Sanitization and structural validation techniques [Electronic resource]. - Mozilla Developer Network, 2024. - Mode of access: https://developer.mozilla.org (date of access: 12.12.2024). - Title from screen.

ДОДАТОК А ДЕМОНСТРАЦІЙНІ МАТЕРІАЛИ

ДЕРЖАВНИЙ УНІВЕРСИТЕТ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНИХ ТЕХНОЛОГІЙ

НАВЧАЛЬНО–НАУКОВИЙ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

КАФЕДРА ІНЖЕНЕРІЇ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

Магістерська робота

«Розробка методики оцінювання якості та оптимізації систем автоматичного перекладу і локалізації вебконтенту на основі метрик штучного інтелекту»

Виконав: студент групи ПДМ-62 Кирило СТАСЮК

Керівник: канд. техн. наук, доц., доцент кафедри ІПЗ Владислав ЯСКЕВИЧ

Київ - 2025

МЕТА, ОБ’ЄКТА ТА ПРЕДМЕТ ДОСЛІДЖЕННЯ

Мета роботи: розробити та обґрунтувати метод оцінювання ефективності систем штучного інтелекту для автоматичного перекладу та локалізації веб-контенту, що дозволяє виявляти сильні та слабкі сторони існуючих рішень та визначати напрями їх удосконалення з урахуванням специфіки веб-середовища.

Об’єкт дослідження: процес автоматичного перекладу та локалізації веб-контенту.

Предмет дослідження: методи оцінювання та вдосконалення якості перекладу й локалізації веб-контенту за допомогою систем штучного інтелекту.

ЗАВДАННЯ РОБОТИ

1. Проаналізувати сучасні підходи до автоматичного перекладу та локалізації веб-контенту, визначити їхні ключові обмеження щодо стилю, термінології та збереження HTML-структури.
2. Розробити методику комплексного оцінювання якості перекладу, яка поєднує лінгвістичний аналіз на основі великих мовних моделей (OpenAI GPT) та технічну перевірку структурної цілісності HTML-документа за допомогою JSDOM.
3. Розробити програмний модуль оцінювання, що реалізує автоматичне порівняння перекладів, включаючи аналіз DOM, обчислення StructureScore, лінгвістичне AI-оцінювання та формування інтегрального показника Final Score.
4. Провести експериментальне дослідження перекладів Google Translate, DeepL та OpenAI GPT, виконати порівняльний аналіз стилістичної точності, термінологічної узгодженості та збереження HTML-структури.

3

ПОРІВНЯЛЬНА ХАРАКТЕРИСТИКА ІСНУЮЧИХ МЕТОДІВ ОЦІНЮВАННЯ ЯКОСТІ АВТОМАТИЧНОГО ПЕРЕКЛАДУ

Метод	Ключові функціональності	Ключові недоліки
BLEU	Основна ідея полягає у вимірюванні збігу n-грам (послідовностей з n слів) між машинним перекладом і одним або кількома референсними перекладами	Чутливість до перестановок слів
METEOR	Оцінює переклад не лише за точним збігом слів, а й враховує стемінг, синонімію та морфологічні варіації	Вища обчислювальна складність
chrF	chrF базується на порівнянні перекладу й еталону на рівні символів (char-level)	Не враховує структуру речення
COMET	Використовує глибокі мовні моделі (наприклад, XLM-R) для представлення речень у багатомовному векторному просторі	Значні обчислювальні ресурси

4

МАТЕМАТИЧНА МОДЕЛЬ ПЕРЕВІРКИ СТРУКТУРНОЇ ЦІЛІСНОСТІ

S - підсумковий бал структурної цілісності,

$$S = max(0,\ 100 - k \times |T_1 - T_0|)$$

де

T_0 - кількість тегів в оригінальному документі;

T_t - кількість тегів у перекладеному документі;

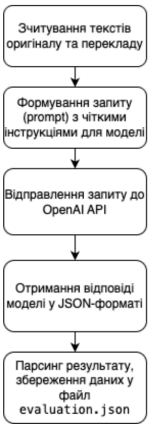
k - коефіцієнт штрафу (у цьому проєкті взято число 10).

5

АЛГОРИТМ ПЕРЕВІРКИ СТРУКТУРНОЇ ЦІЛІСНОСТІ

6

АЛГОРИТМ ОЦІНКИ ЯКОСТІ ПЕРЕКЛАДУ НА ОСНОВІ OPENAI API



7

КОМБІНОВАНА ФОРМУЛА ІНТЕГРАЛЬНОЇ ОЦІНКИ (FINAL SCORE)

Формування показнику якості LQS (Linguistic Quality Score)

$$LQS = \frac{A + F + S + T}{4}$$

F - підсумкова оцінка системи (у діапазоні 0–100),

$$F = w_1 \times LQS + w_2 \times S$$

де

LQS - лінгвістичний показник якості; (додати розшифровку)

S - бал структурної цілісності;

w_1, w_2 - коефіцієнти, що визначають відносну значущість обох факторів.

8

РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ


```
[
  {
    "system": "google",
    "accuracy": 82,
    "fluency": 78,
    "style": 75,
    "terminology": 80,
    "comment": "Generally acceptable translation, but with noticeable stylistic simplifications and inconsistencies.",
    "htmlCheck": {
      "structureMatch": false,
      "structureScore": 95
    },
    "finalScore": 83.1
  },
  {
    "system": "deepl",
    "accuracy": 88,
    "fluency": 90,
    "style": 86,
    "terminology": 87,
    "comment": "High-quality translation with strong fluency and good terminology retention; minor stylistic simplifications.",
    "htmlCheck": {
      "structureMatch": true,
      "structureScore": 97
    },
    "finalScore": 90.5
  },
  {
    "system": "openai",
    "accuracy": 91,
    "fluency": 93,
    "style": 89,
    "terminology": 88,
    "comment": "Excellent translation with consistent style, high accuracy, and precise terminology preservation.",
    "htmlCheck": {
      "structureMatch": true,
      "structureScore": 100
    },
    "finalScore": 92.8
  }
]
```

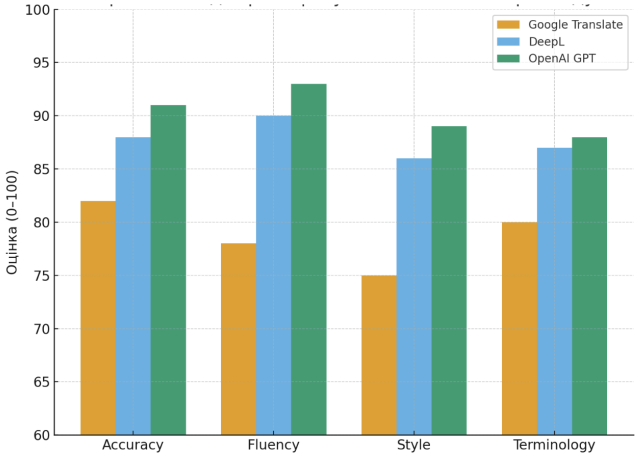

9

РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

Система перекладу	Accuracy	Fluency	Style	Terminology	Structure Score	Final Score
Google Translate	82	78	75	80	95	83.1
DeepL	88	90	86	87	97	90.5
OpenAI GPT	91	93	89	88	100	92.8

10

ПОРІВНЯЛЬНА ДІАГРАМА РЕЗУЛЬТАТІВ ОЦІНЮВАННЯ СИСТЕМ ПЕРЕКЛАДУ



11

ВИСНОВКИ

1. Проаналізовано сучасні методи автоматичного перекладу та локалізації веб-контенту, включаючи статистичні та нейромережеві моделі (SMT, NMT/Transformer). Визначено їхні ключові обмеження: нечутливість до HTML-структури, низьку точність стилістичної відповідності, обмежене збереження термінології та залежність від поверхневих метрик (BLEU, METEOR), що не відображають реальної якості перекладу.
2. Розроблено методику комплексного оцінювання якості перекладу, що поєднує лінгвістичний аналіз на основі OpenAI API та технічну перевірку структурної цілісності HTML-документа за допомогою JSDOM. Запропонована модель включає оцінку точності, стилю, природності та термінології, а також перевірку відповідності тегів, вкладеності та обсягу структури.
3. Створено програмний модуль для автоматичного порівняння перекладів від різних систем. Модуль реалізує повний процес: обробку HTML, DOM-порівняння, AI-оцінювання, обчислення показника StructureScore та формування інтегральної оцінки Final Score за ваговою моделлю.
4. Проведено експериментальне дослідження, яке показало, що OpenAI GPT забезпечує найвищу якість перекладу (Final Score ≈ 92.8), DeepL демонструє стабільну точність і природність (≈ 90.5), тоді як Google Translate поступається за стилем і термінологічною точністю (≈ 83.1). У всіх систем структура переважно зберігається, але Google частіше допускає незначні порушення DOM.

12

ПУБЛІКАЦІЇ ТА АПРОБАЦІЯ РОБОТИ

Тези доповідей:

1. Стасюк К.С., Яскевич В.О. Дослідження та оцінка ефективності штучного інтелекту у автоматичному перекладі та локалізації веб-контенту. VI Міжнародна науково-технічна конференція «Сучасний стан та перспективи розвитку IoT», 15 квітня 2025 р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.-112-113.
2. Стасюк К.С., Яскевич В.О. Використання інструментів та бібліотек штучного інтелекту для автоматичного перекладу та локалізації веб-контенту. Всеукраїнська науково-технічна конференція «Застосування програмного забезпечення в ІКТ», 24 квітня 2025 р., Київ, Державний університет інформаційно-комунікаційних технологій. Збірник тез. К.: ДУІКТ, 2025. С.-311-312.

ДОДАТОК Б ЛІСТИНГИ ОСНОВНИХ ПРОГРАМНИХ МОДУЛІВ

```
import { JSDOM } from "jsdom";


export function
checkHTMLStructure(originalHTML,
translatedHTML) {
  const origTags = [...new
JSDOM(originalHTML).window.document.querySel
ectorAll('*')]
    .map(el => el.tagName);
  const transTags = [...new
JSDOM(translatedHTML).window.document.queryS
electorAll('*')]
    .map(el => el.tagName);


  const structureMatch = JSON.stringify(origTags)
=== JSON.stringify(transTags);
  const diff = Math.abs(origTags.length -
transTags.length);


  return {
    structureMatch,
    diff,
    structureScore: structureMatch ? 100 :
Math.max(0, 100 - diff * 10)
  };
}


import OpenAI from "openai";
import fs from "fs";


const client = new OpenAI({ apiKey:
process.env.OPENAI_API_KEY });


export async function evaluateTranslation(original,
translated, systemName) {
  const prompt = `
You are a professional linguist. Compare the
translation to the original text.
Rate on a 0-100 scale for:
1) Accuracy (meaning preservation)
2) Fluency (grammar and naturalness)
3) Style consistency (tone and register)
4) Terminology (key terms preservation)
Return the result as valid JSON:
{"system": "${systemName}", "accuracy": X,
"fluency": X, "style": X, "terminology": X,
"comment": "short comment"}
Original: """${original}"""
Translation: """${translated}"""
`;


  const res = await client.chat.completions.create({
    model: "gpt-4o-mini",
    messages: [{ role: "user", content: prompt }],
  });


  const jsonText = res.choices[0].message.content;
  return JSON.parse(jsonText);
}


export function calculateFinalScore(lqs,
structureScore) {
  const w1 = 0.7;
  const w2 = 0.3;
  return (w1 * lqs + w2 * structureScore).toFixed(2);
}
```