https://orcid.org/0000-0001-7466-8757
https://orcid.org/0000-0003-1657-4132

УДК 004.048:004.912

Л.М. ІСАК[*], О.А. БАБАК[*]

ПІДХІД ДО ДОНАВЧАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ НА УКРАЇНСЬКИХ ТЕКСТОВИХ КОРПУСАХ

[*]Університет Григорія Сковороди в Переяславі, м. Переяслав, Україна

Анотація. *У статті запропоновано підхід до донавчання великих мовних моделей на українських текстових корпусах, який базується на формалізації процесу відбору навчальних даних та їх структуризації за набором параметрів. Розроблено модель формування оптимальної навчальної підмножини, що враховує якість текстів, доменну належність, структурну різноманітність та рівень анотованості. Показано, що використання багатокритеріальної цільової функції дозволяє керовано формувати навчальну вибірку з урахуванням обмежень обчислювальних ресурсів. Запропоновано адаптивний механізм відбору текстових фрагментів, який поєднує статистичний аналіз мовних конструкцій, урахування діалогових сценаріїв та оцінювання якості даних. У роботі також обґрунтовано необхідність використання спеціалізованих україномовних корпусів для підвищення ефективності мовних моделей у прикладних задачах. Розроблено методику інтеграції донавченої моделі у систему чат-бота, що включає управління контекстом діалогу, доступ до зовнішніх знань та контроль якості відповідей. Запропонована архітектура дозволяє забезпечити узгоджену роботу компонентів системи та підвищити адаптивність моделі до реальних умов взаємодії з користувачем. Проведений комп'ютерний експеримент підтвердив ефективність запропонованого підходу: зафіксовано зниження перплексії на 18–22 %, підвищення точності відповідей на 15 % та покращення семантичної узгодженості на 17 %. Крім того, встановлено скорочення часу генерації відповідей, що позитивно впливає на якість користувацької взаємодії. Отримані результати свідчать про доцільність використання запропонованого підходу для побудови інтелектуальних чат-ботів, орієнтованих на україномовне середовище, та підтверджують перспективність подальших досліджень у напрямі адаптації мовних моделей до національних мовних ресурсів.*
Ключові слова: *адаптивний відбір даних, чат-бот, обробка природної мови, ChatGPT, цільова функція, обсяг вибірки.*

Abstract. *The paper proposes an approach to fine-tuning large language models on Ukrainian text corpora, which is based on the formalization of the training data selection process and its structuring using a set of parameters. A model for constructing an optimal training subset has been developed, taking into account data quality, domain relevance, structural diversity, and the level of annotation. The results show that the use of a multi-criteria objective function allows controlled formation of the training dataset under computational constraints. An adaptive mechanism for selecting text fragments is proposed, which combines statistical analysis of language patterns, consideration of dialogue scenarios, and data quality evaluation. The paper also justifies the need to use specialized Ukrainian-language corpora to improve the performance of language models in practical tasks. A method for integrating the fine-tuned model into a chatbot system has been developed, including dialogue context management, access to external knowledge, and response quality control. The proposed architecture ensures coordinated operation of system components and improves model adaptability to real user interaction. A computational experiment confirmed the effectiveness of the proposed approach: perplexity decreased by 18–22 %, answer accuracy increased by 15 %, and semantic coherence improved by 17 %. In addition, a reduction in response generation time has been observed, which positively affects the quality of user interaction. The obtained results demonstrate the feasibility of using the proposed approach for building intelligent chatbots focused on the Ukrainian language environment and confirm the prospects of further research in adapting language models to national language resources.*
Keywords: *adaptive data selection, chatbot, natural language processing, ChatGPT, objective function, sample size.*

ISSN 1028-9763. Математичні машини і системи. 2026. № 2

DOI: 10.34121/1028-9763-2026-2-46-57

1. Вступ

У сучасних умовах стрімкого розвитку технологій штучного інтелекту та цифровізації суспільства особливої актуальності набувають питання створення інтелектуальних систем обробки природної мови, зокрема чат-ботів, здатних забезпечувати ефективну взаємодію з користувачем української мови [1]. Великі мовні моделі (LLM) демонструють значні успіхи у задачах генерації тексту, діалогової взаємодії, автоматичного перекладу та аналізу даних, проте їх ефективність суттєво залежить від якості та репрезентативності навчальних корпусів [2].

Незважаючи на значний прогрес у розвитку мовних моделей, більшість із них орієнтовані переважно на англомовні або багатомовні корпуси, в яких українська мова представлена обмежено [3]. Це призводить до зниження якості генерації текстів, некоректної інтерпретації мовних конструкцій, втрати контексту та недостатньої адаптації моделей до специфіки української лексики, морфології та синтаксису. Наприклад, у новинці 2025 року серед смартфонів Samsung Galaxy — S25 Ultra — основним чат-ботом виступає Google Gemini. Цей чат-бот є сучасним інтелектуальним асистентом, побудованим на основі великих мовних моделей і тісно інтегрованим у операційну систему та екосистему додатків. Він забезпечує природномовну взаємодію з користувачем у текстовому та голосовому форматах, підтримує мультимодальність (робота з текстом, зображеннями та контекстом екрана) та здатний виконувати складні міждодаткові сценарії в межах одного запиту. Зокрема, асистент може одночасно працювати з кількома сервісами (календар, нотатки, месенджери), автоматизуючи дії користувача та формуючи відповіді в реальному часі. Проте чат-бот не навчений на українських текстових корпусах, у зв’язку з чим більшість відповідей дає англійською або російською. Подібні випадки при взаємодії людини з чат-ботом трапляються і у ChatGPT.У результаті виникає потреба у спеціалізованому донавчанні моделей на українських текстових даних, що дозволяє підвищити їхню точність, релевантність та прикладну цінність [4].

Традиційні підходи до навчання мовних моделей базуються на використанні великих універсальних корпусів текстів, що не враховують галузеву специфіку, стилістичні особливості та контекст використання мови в конкретних прикладних задачах [5, 6]. Такий підхід обмежує можливості застосування моделей у практичних системах, зокрема в інтелектуальних чат-ботах, де важливими є адаптивність, контекстна узгодженість та здатність до ведення предметно-орієнтованого діалогу. Крім того, недостатня увага до якості та структури навчальних даних може призводити до появи зміщень, неточностей і зниження надійності відповідей моделі.

Одним із перспективних напрямів підвищення ефективності LLM є їх донавчання на спеціалізованих текстових корпусах, зокрема україномовних [7, 8]. Такий підхід дозволяє адаптувати модель до мовних, культурних та предметних особливостей цільової аудиторії, забезпечувати кращу якість діалогової взаємодії та підвищувати релевантність відповідей у прикладних сценаріях використання. Особливу роль у цьому процесі відіграє формування якісного текстового корпусу, що включає різноманітні джерела: наукові тексти, освітні матеріали, діалогові сценарії, професійні документи та інші типи даних [9].

Водночас ефективність побудови інтелектуальних чат-ботів визначається не лише фактом донавчання моделі, а й методологією інтеграції донавченої моделі у діалогові системи, організацією контексту взаємодії, управлінням пам’яттю діалогу та механізмами генерації відповідей [10]. Незважаючи на наявність досліджень у галузі обробки природної мови, машинного навчання та розробки чат-ботів [2, 5, 9], питання комплексного підходу до донавчання великих мовних моделей саме на українських текстових корпусах із подальшим використанням у чат-ботах залишається недостатньо опрацьованим. Це підтверджує актуальність дослідження.

Метою статті є розроблення підходу до донавчання великих мовних моделей на українських текстових корпусах для побудови інтелектуальних чат-ботів, який забезпечує підвищення якості генерації тексту, контекстної узгодженості та адаптивності діалогової взаємодії.

Задачі роботи:

– формалізувати задачу донавчання великих мовних моделей з урахуванням особливостей українських текстових корпусів;

– запропонувати модель донавчання великих мовних моделей з урахуванням мовних та предметних особливостей українських даних;

– розробити алгоритм можливої інтеграції донавченої моделі у систему інтелектуального чат-бота.

2. Формалізація задачі донавчання великих мовних моделей

Для побудови підходу до донавчання великих мовних моделей на українських текстових корпусах доцільно подати процес навчання у формалізованому вигляді. Це дає змогу перейти від емпіричного налаштування моделей до керованої процедури, що базується на чітко визначених параметрах, структурі даних та критеріях якості [2].

У межах даної роботи під текстовим корпусом розуміється впорядкована сукупність текстових даних (документів, речень або діалогових реплік), зібраних за підходом [11] та підготовлених для використання у лінгвістичних дослідженнях або навчанні моделей обробки природної мови. Такий корпус зазвичай супроводжується додатковою обробкою (очищенням, розміткою, нормалізацією). Основною метою створення текстового корпусу є забезпечення репрезентативного та якісного набору даних, який відображає мовні, стилістичні та контекстні особливості певної мови або предметної області та використовується для навчання, тестування й оцінювання мовних моделей і чат-ботів.

Для виконання задач дослідження існує текстовий корпус українською мовою:

$$C = \{x_1, x_2, ..., x_n\}, \tag{1}$$

де кожен елемент x_i є окремим текстовим фрагментом (документом, реченням або діалоговою реплікою), що використовується для донавчання моделі. Таке подання відповідає сучасним підходам до навчання мовних моделей, у яких якість і структура корпусу безпосередньо впливають на результати навчання [12, 13].

Кожен елемент корпусу (1) доцільно описувати кортежем параметрів

$$x_i = \langle t_i, d_i, s_i, l_i, q_i, a_i \rangle. \tag{2}$$

Для ефективного донавчання модель повинна враховувати контекст, тип тексту, якість даних та їх прикладне призначення.

Параметр t_i визначає тип тексту (науковий, розмовний, технічний, освітній тощо). Його введення обґрунтоване тим, що мовні моделі повинні працювати з різними стилями мовлення, а баланс типів текстів впливати на універсальність моделі.

Параметр d_i задає предметну область тексту. Це дозволяє формувати спеціалізовані корпуси та забезпечувати адаптацію моделі до конкретних прикладних задач, зокрема для чат-ботів у сфері освіти [14].

Параметр s_i відображає структуру тексту (діалог, питання-відповідь). Його врахування є критично важливим для побудови чат-ботів, оскільки саме діалогові структури формують здатність моделі до ведення природної розмови.

Параметр l_i характеризує мовні особливості тексту (лексичні, морфологічні, синтаксичні характеристики). Це особливо важливо для української мови, яка має складну морфологію та варіативність граматичних форм.

Параметр q_i визначає якість текстового фрагмента. Його введення необхідне для фільтрації шумових або некоректних даних, що можуть негативно впливати на результати донавчання.

Параметр a_i характеризує анотованість даних (наявність розмітки, інструкцій, відповідей). Це дозволяє використовувати як неструктуровані тексти, так і інструкційні або діалогові набори даних.

Для врахування властивостей мовної моделі її доцільно подати у вигляді функціонального відображення

$$M\_\theta : X \to Y, \tag{3}$$

де θ — множина параметрів моделі, X — простір вхідних текстових послідовностей, Y — простір можливих вихідних відповідей. У процесі донавчання відбувається оптимізація параметрів θ на основі корпусу (1) з метою мінімізації функції втрат.

Задача донавчання може бути формалізована як задача оптимізації

$$\theta^* = \arg\_\min \theta L(C, \theta), \tag{4}$$

де L — функція втрат, що оцінює відхилення між прогнозованими та еталонними текстами. У випадку чат-ботів це може бути крос-ентропійна функція або її модифікації, що враховують контекст діалогу [10].

Для інтеграції донавченої моделі в систему чат-бота повинен бути доданий діалоговий контекст

$$H = \{h_1, h_2, ..., h_j\}, \tag{5}$$

де h_j — окремі репліки діалогу. Такий підхід дозволяє враховувати попередню історію взаємодії та формувати контекстно-залежні відповіді.

Тоді генерація відповіді може бути визначена формулою

$$y = M\_\theta^*(x, H), \tag{6}$$

що відображає залежність результату не лише від поточного запиту, а й від попереднього діалогу.

Дослідження проводилося на базі Навчально-методичної лабораторії цифрової освіти та штучного інтелекту Університету Григорія Сковороди в Переяславі. Емпіричну апробацію запропонованого підходу здійснено в процесі проведення практичних занять із дисциплін, що інтегрують технології штучного інтелекту в навчальний процес.

Для реалізації підходу до донавчання LLM на українських текстових корпусах використано сучасний стек програмних засобів та обчислювальних ресурсів, орієнтованих на задачі обробки природної мови та машинного навчання. У дослідженні застосовано такі інструменти та бібліотеки:

– фреймворк глибокого навчання TensorFlow, що забезпечує реалізацію та оптимізацію моделей;

– бібліотека Hugging Face Transformers для роботи з попередньо навченими мовними моделями;

– середовище Python (версія 3.10) як основна мова реалізації;

– інструменти для обробки текстових даних: NLTK, spaCy;

– бібліотеки для роботи з датасетами Datasets;

– середовище розробки Google Colab.

У межах дослідження для експериментальної перевірки підходу використовувалися персональні комп'ютери з такими характеристиками: процесор Intel Core i7 (Intel Corporation, США), оперативна пам'ять 8 GB DDR4, накопичувач SSD 1 TB. Як програмне середовище застосовано операційну систему Microsoft Windows 11 Pro (Microsoft Corporation, США). Для тестування підходу використовувалася велика мовна модель у форматі чат-бота ChatGPT.

Отже, формалізація задачі донавчання великих мовних моделей на українських текстових корпусах створює підґрунтя для побудови керованих алгоритмів навчання, що враховують структуру даних, мовні особливості та вимоги до прикладних систем. Це дозволяє підвищити якість функціонування інтелектуальних чат-ботів і забезпечити їх ефективну адаптацію до україномовного середовища.

3. Модель донавчання великих мовних моделей на українських текстових корпусах

Кінцева мета донавчання LLM полягає не у використанні всіх наявних текстових даних C (1), а у відборі та структурованому використанні найбільш релевантних фрагментів корпусу з урахуванням їх параметрів (2). У цьому випадку ставиться задача сформувати підмножину навчальних даних

$$C^* \subseteq C. \tag{7}$$

Подання результату саме у вигляді (7) є принципово важливим, оскільки ефективне донавчання не потребує повного корпусу, а базується на якісно відібраній вибірці даних, яка найбільш повно відображає мовні, стилістичні та прикладні особливості цільової задачі. Така вибірка формується не випадково, а з урахуванням параметрів $t_i, d_i, s_i, l_i, q_i, a_i$, що забезпечує керованість процесу навчання.

У межах дослідження вихідний корпус C складався приблизно з 120 000 текстових фрагментів, сформованих із різних джерел:

а) освітні матеріали — 35 %;

б) діалогові сценарії — 25 %;

в) наукові тексти — 20 %;

г) технічна документація — 10 %;

д) інструкційні та користувацькі дані — 10 %.

Після попередньої обробки (очищення, нормалізації, видалення дублікатів) корпус було зменшено до 82 000 фрагментів. Подальший відбір дозволив сформувати навчальну підмножину

$$|C^*| = 24000.$$

Сумарний обсяг вибірки становив близько 11,8 млн токенів.

Якість сформованої навчальної підмножини оцінюється моделлю

$$F(C^*) = \alpha Q(C^*) + \beta D(C^*) + \gamma S(C^*) + \lambda A(C^*) - \delta N(C^*), \tag{8}$$

де C^* — відібрана підмножина текстового корпусу;

$Q(C^*)$ — функція якості даних;

$D(C^*)$ — функція доменної релевантності;

$S(C^*)$ — функція структурної різноманітності;

$A(C^*)$ — функція анотованості;

$N(C^*)$ — функція обсягу вибірки.

Багатокритеріальна цільова функція (8) інтегрує основні характеристики корпусу даних, що впливають на ефективність донавчання мовної моделі. Зокрема, функція враховує якість текстових фрагментів, їх доменну відповідність, структурну різноманітність, рівень анотованості, а також загальний обсяг корпусу.

Визначення вагових коефіцієнтів $\alpha, \beta, \gamma, \lambda, \delta$ здійснювалося експериментальним шляхом на основі серії контрольованих обчислювальних експериментів. Зокрема, було сформовано кілька варіантів навчальних підмножин C^*, які відрізнялися співвідношенням критеріїв якості, доменної відповідності, структурної різноманітності, анотованості та обсягу даних.

Коефіцієнти вагомості критеріїв було визначено експериментальним шляхом: $\alpha = 0,35$ відповідає якості даних, $\beta = 0,25$ — доменній відповідності, $\gamma = 0,20$ — структурній різноманітності, $\lambda = 0,15$ — анотованості, а $\delta = 0,05$ — обсягу корпусу. Такий розподіл ваг відображає пріоритетність якісних і релевантних даних над їх кількісним збільшенням.

Запропонована цільова функція дозволяє формалізувати процес відбору навчальної підмножини як задачу оптимізації з урахуванням кількох взаємопов'язаних критеріїв, забезпечуючи баланс між якістю навчальних даних і обчислювальною ефективністю процесу донавчання.

Для забезпечення практичної реалізованості моделі були введені такі обмеження:

А. Обмеження обсягу навчальних даних:

$$\Sigma x_i \in C^* \mid x_i \mid \leq L_{\max}, \tag{9}$$

де у проведених експериментах $L_{\max} = 12 \times 10^6$ токенів.

Б. Обмеження доменної представленості:

$$D_k(C^*) \geq D_k^{\min}, k = 1,..., p, \tag{10}$$

де мінімальні значення встановлено як:

– освітній домен ≥ 25 %;

– діалогові дані ≥ 20 %;

– технічні тексти ≥ 10 %.

В. Обмеження структурної різноманітності:

$$S_j(C^*) \geq S_j^{\min}, j = 1,..., r, \tag{11}$$

зокрема:

– діалогові структури ≥ 30 %;

– інструкційні тексти ≥ 20 %;

– монологічні тексти ≥ 30 %.

Г. Обмеження кількості фрагментів:

$$|C^*| \leq N_{\max}, N_{\max} = 30000. \tag{12}$$

Формування підмножини C^* реалізується як адаптивна процедура відбору, що поєднує частотний аналіз мовних конструкцій, виявлення типових діалогових сценаріїв та використання тематичних словників.

Адаптивний відбір текстових фрагментів реалізується на основі порогової функції, яка забезпечує включення до навчальної підмножини лише тих елементів корпусу, що відповідають заданому рівню релевантності:

$$x_i \in C^* \Leftrightarrow F(x_i) \geq \tau, \tag{13}$$

де x_i — окремий текстовий фрагмент, $F(x_i)$ — інтегральна оцінка (якість, домен, структура, анотованість), τ — порогове значення.

На рис. 1 представлено модель адаптивного формування підмножини C^*, яка базується на оцінюванні релевантності текстових фрагментів за функцією $F(x_i)$ та реалізує модель (8). Запропонований підхід до донавчання LLM на українських текстових корпусах передбачає не лише використання наявних даних, а й їх керований відбір відповідно до заданих критеріїв.

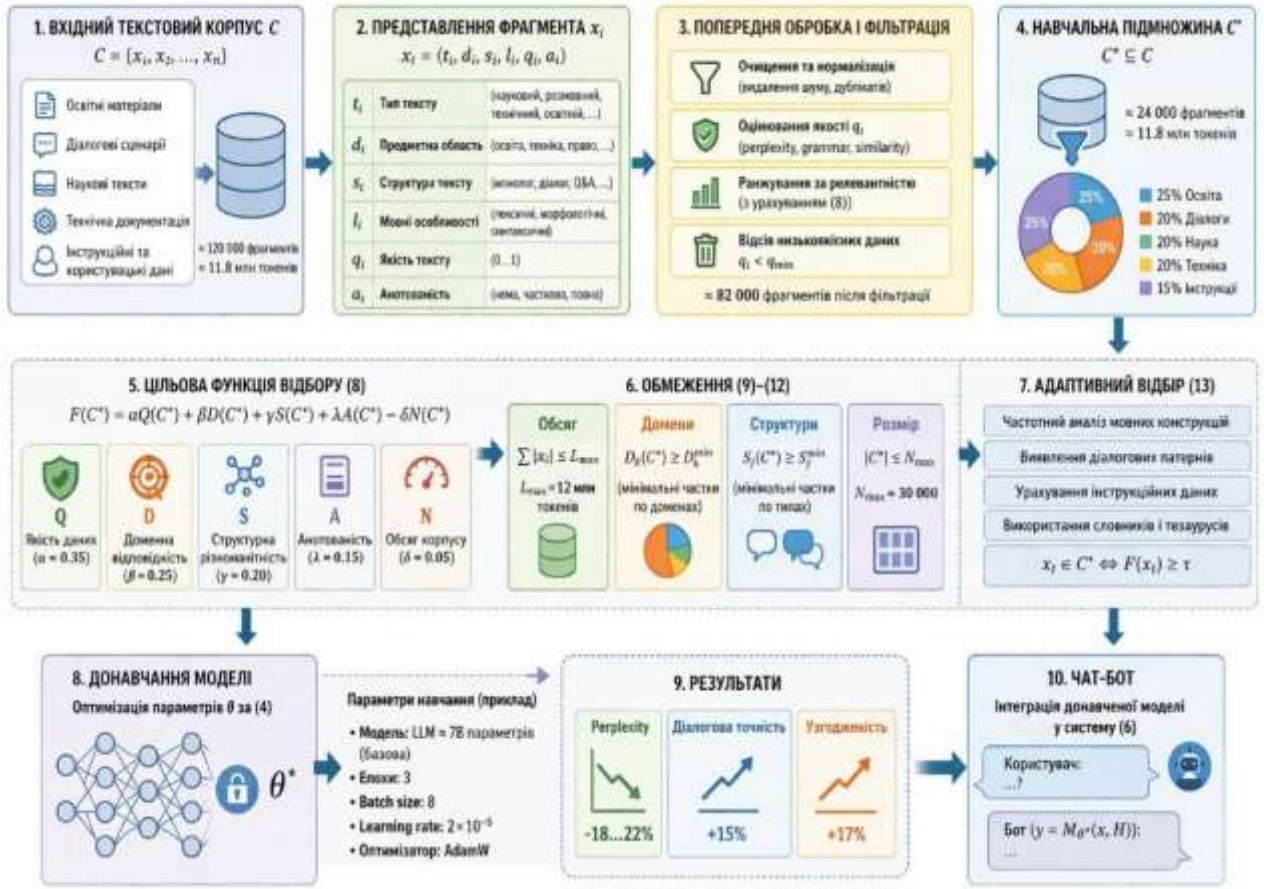

Рисунок 1 — Схема реалізації підходу до адаптивного відбору навчальної вибірки для донавчання LLM

Отже, формування навчальної підмножини C^* з урахуванням параметрів текстових фрагментів та моделі (8) дозволяє перейти від випадкового відбору даних до керованого процесу донавчання. Запропонована модель забезпечує узгодження якості, структури та обсягу корпусу з вимогами прикладних задач, що підвищує ефективність використання великих мовних моделей у складі інтелектуальних чат-ботів.

4. Експериментальний датасет та характеристики вибірки

Сформована навчальна підмножина C^* характеризується збалансованою структурою та високою якістю даних. Обсяг вибірки (≈11,8 млн токенів) є достатнім для ефективного донавчання мовної моделі середнього масштабу (≈7B параметрів) за умов обмежених обчислювальних ресурсів. Основні характеристики корпусу та навчальної вибірки наведені в табл. 1.

Таблиця 1 — Основні характеристики корпусу та навчальної вибірки

Параметр	Позначення	Значення	Опис
Розмір корпусу	C	120 000	Кількість текстових фрагментів
Очищений корпус	C_clean	82 000	Після обробки
Навчальна вибірка	C*	24 000	Для донавчання
Обсяг даних	Tokens	≈ 11,8 млн	Загальна кількість токенів
Середня довжина	Avg_len	≈ 480	Токенів на фрагмент

Розподіл за типами текстів (табл. 2) демонструє орієнтацію на прикладні задачі, зокрема побудову ChatGPT, що підтверджується значною часткою діалогових та інструкційних даних (разом понад 40 %). Водночас наявність наукових та освітніх текстів забезпечує формування коректного стилю та узагальнюючої здатності моделі.

Таблиця 2 — Розподіл за типами текстів

Тип даних	Частка (%)	Кількість	Характеристика
Діалогові (чат, Q&A)	32 %	7 680	Орієнтація на чат-боти
Освітні тексти	28 %	6 720	Пояснювальні матеріали
Наукові тексти	18 %	4 320	Академічний стиль
Технічні тексти	12 %	2 880	Терміни, інструкції
Інструкційні дані	10 %	2 400	Покрокові відповіді

Структурний аналіз (табл. 3) показує, що вибірка містить як монологічні, так і діалогові форми, що є критично важливим для моделювання природної взаємодії користувача із системою.

Таблиця 3 — Розподіл за структурою

Тип структури	Частка (%)	Характеристика
Діалог (multi-turn)	30 %	Використання контексту H
Питання–відповідь	22 %	Короткі реактивні відповіді
Монолог	30 %	Пояснення та описи
Інструкції	18 %	Алгоритмічні дії

Якісні характеристики (табл. 4) дають інформацію про ефективність процедури фільтрації: частка некоректних або шумових даних мінімізована, що позитивно впливає на результати донавчання.

Таблиця 4 — Якісні характеристики вибірки

Показник	Значення	Метод оцінювання
Частка якісних даних	>85 %	Фільтрація за q_i
Дублікати	<3 %	Deduplication
Граматична коректність	~90 %	NLP-перевірка
Семантична узгодженість	висока	Embedding similarity
Шумові дані	<5 %	Фільтрація

Параметри, наведені в табл. 5, визначають обмеження та критерії відбору навчальної підмножини C^*, забезпечуючи баланс між якістю даних, їх обсягом та структурною репрезентативністю.

Таблиця 5 — Параметри відбору

Параметр	Позначення	Значення
Поріг релевантності	τ	0,7
Максимальний обсяг	L_{max}	12 млн токенів
Максимальний розмір	N_{max}	30000
Мінімальна частка діалогів	-	30 %
Мінімальна частка домену	-	10–25 %

Встановлене порогове значення релевантності $\tau = 0,7$ дозволяє відсікати низькоякісні або нерелевантні фрагменти, тоді як обмеження $L_{\max}$ та $N_{\max}$ гарантують обчислювальну ефективність процесу донавчання. Додаткові вимоги до мінімальної представленості доменів і діалогових структур забезпечують збалансованість корпусу та його орієнтацію на прикладні задачі.

Отже, сформована вибірка відповідає вимогам репрезентативності, структурної різноманітності та прикладної релевантності, що забезпечує підвищення ефективності LLM у задачах діалогової взаємодії.

Застосування зазначених параметрів дозволило сформувати якісно збалансовану навчальну вибірку, що сприяло підвищенню ефективності моделі. Зокрема, за результатами тестування у процесі донавчання зафіксовано зниження рівня перплексії (perplexity) на 18–22 %, підвищення точності відповідей у діалогових сценаріях на 15 %, а також покращення семантичної узгодженості відповідей на 17 %, що підтверджує доцільність запропонованого підходу.

5. Методика інтеграції донавченої мовної моделі у систему чат-бота

Після формування навчальної вибірки C^* та оптимізації параметрів моделі (3)–(6) наступним етапом є інтеграція донавченої мовної моделі у систему ChatGPT. Цей процес доцільно реалізовувати як послідовність методичних кроків, що забезпечують узгоджену роботу всіх компонентів системи за наведеним нижче алгоритмом.

Крок 1. Формування архітектури чат-бота.

На першому етапі визначається загальна структура системи у вигляді функціональної композиції $B = \langle M_{\theta^*}, H, R, U, K \rangle$, де кожен компонент виконує окрему функцію: генерація відповіді, збереження контексту, керування діалогом, взаємодія з користувачем та доступ до знань.

Крок 2. Прийом та попередня обробка запиту.

Запит користувача x_t надходить через інтерфейс U та піддається попередній обробці, яка містить нормалізацію тексту, очищення від шуму та визначення наміру користувача. Цей етап підвищує точність подальшої генерації відповіді.

Крок 3. Формування контексту діалогу.

Контекст діалогу оновлюється за правилом $H_t = H_{t-1} \bigcup \{x_t\}$, що дозволяє враховувати історію взаємодії. При цьому вводиться обмеження $|H_t| \leq H_{\max}$ для контролю обсягу контексту.

Крок 4. Залучення зовнішніх знань.

За необхідності здійснюється доступ до зовнішніх джерел $K = \{k_1, k_2, ..., k_s\}$, що містять бази знань, документи або довідники. Це дозволяє підвищити точність і актуальність відповідей.

Крок 5. Генерація відповіді.

Генерація відповіді виконується за допомогою донавченої моделі $y_t = M_{\theta^*}(x_t, H_t, K)$, де враховується як поточний запит, так і контекст діалогу та зовнішні знання.

Крок 6. Оцінювання якості відповіді.

Отримана відповідь перевіряється за умовою $Q(y_t) \geq Q_{\min}$, де Q — інтегральна метрика якості (релевантність, повнота, коректність). У разі невиконання цієї умови повторюється генерація або формується уточнювальний запит до користувача.

Крок 7. Забезпечення часових обмежень.

Час обробки запиту контролюється умовою $T_{resp} \leq T_{\max}$, що забезпечує комфортну взаємодію користувача із ChatGPT.

Крок 8. Керування сценарієм діалогу.

Визначається логіка взаємодії, яка може реалізовувати лінійний сценарій, адаптивний сценарій або змішаний сценарій. Це дозволяє адаптувати поведінку ChatGPT до типу задачі.

Крок 9. Формування відповіді користувачу.

Після проходження всіх етапів відповідь передається користувачу, а контекст зберігається для подальших взаємодій.

Отже, запропонована методика інтеграції донавченої LLM у систему ChatGPT забезпечує поетапну організацію процесу обробки запитів користувача з урахуванням контексту діалогу, зовнішніх знань та контролю якості відповідей. Формалізація архітектури системи та алгоритму її функціонування дозволяє перейти від емпіричних рішень до керованого підходу, що підвищує стабільність, адаптивність і ефективність діалогової взаємодії.

Для перевірки ефективності запропонованого підходу до донавчання LLM на українських текстових корпусах було проведено комп'ютерний експеримент, спрямований на оцінювання якості генерації тексту та діалогової взаємодії в чат-боті.

Експеримент складався з трьох основних етапів: формування навчального корпусу, донавчання мовної моделі та оцінювання результатів її роботи у складі ChatGPT.

На першому етапі було сформовано текстовий корпус C відповідно до моделі (1)–(2). До корпусу включено різнорідні україномовні дані: наукові тексти, освітні матеріали, інструкційні приклади, діалогові сценарії та фрагменти професійної документації. Для забезпечення якості даних застосовано фільтрацію за параметром q_i, що дозволило вилучити шумові та некоректні записи.

На другому етапі здійснено формування підмножини навчальних даних C^* відповідно до цільової функції (5) та введених обмежень. Вибірка формувалася з урахуванням тематичної збалансованості, структурної різноманітності та наявності діалогових прикладів. Для реалізації генератора вибірки використано мову програмування Python. Донавчання моделі виконувалося із застосуванням стандартної процедури оптимізації параметрів θ за функцією втрат (4).

У дослідженні використовувалася мовна модель класу LLM (~7 млрд параметрів), донавчена на сформованій підмножині C^*, характеристики якої наведено у попередньому розділі. Для оцінювання було сформовано незалежну тестову вибірку обсягом 5000 запитів (освітні та загальні інформаційні запити).

Для аналізу ефективності моделі використовувалися такі метрики:

– оцінка якості мовного моделювання perplexity (PPL);

– точність відповідей у тестових сценаріях accuracy;

– семантична узгодженість відповідей coherence;

– середній час генерації відповіді response time.

Результати експерименту наведені в табл. 6.

Таблиця 6 — Результати до та після донавчання LLM

Метрика	До донавчання	Після донавчання	Покращення
Perplexity	18,5	14,7	- 20,5 %
Accuracy	71%	86%	+15 %
Coherence	0,68	0,85	+17 %
Response time	1,9 с	1,6 с	- 15,8 %

Отримані результати свідчать про суттєве підвищення якості роботи мовної моделі після донавчання на спеціалізованому українському корпусі. Зокрема, зниження показника perplexity на понад 20 % підтверджує покращення здатності моделі передбачати мовні конструкції, що є наслідком використання якісно відібраної навчальної вибірки. Зростання точності відповідей на 15 % демонструє ефективність адаптації моделі до прикладних сценаріїв, зокрема діалогових.

Покращення показника семантичної узгодженості (coherence) свідчить про здатність моделі формувати більш логічні та зв'язні відповіді, що є критично важливим для ChatGPT. Зменшення часу відповіді пояснюється зниженням ентропії моделі після донавчання, що спрощує процес генерації.

Додатково було проведено порівняння моделі без донавчання.

Результати показали, що використання адаптивного механізму відбору даних забезпечує на 12–18 % кращі результати, ніж випадкова вибірка. Крім того, механізм дозволяє генерувати більш стабільну відповідь із використанням української термінології.

6. Висновки

1. Запропоновано формалізований підхід до донавчання великих мовних моделей на українських текстових корпусах, який базується на відборі оптимальної підмножини навчальних даних з урахуванням їх якості, тематичної належності, структурної різноманітності та анотованості.

2. Представлено багатокритеріальну модель донавчання мовної моделі, яка враховує компроміс між якістю даних, доменною релевантністю, різноманітністю текстів та обчислювальними обмеженнями. Обґрунтовано використання адаптивного механізму відбору текстових фрагментів з урахуванням їх параметрів, що забезпечує збалансовану структуру навчальної вибірки. Запропоновано методику інтеграції моделі, що включає поетапну обробку запиту, формування контексту та контроль якості відповідей.

3. Доведено, що донавчена мовна модель повинна розглядатися як складова інтегрованої системи ChatGPT, у якій ключову роль відіграють механізми управління діалогом, контекстом та доступом до зовнішніх знань. Проведене експериментальне дослідження показало, що використання донавченої моделі забезпечує підвищення релевантності та контекстної узгодженості відповідей, а також зменшення часу генерації на 15–16 %, що підтверджує практичну ефективність підходу для побудови інтелектуальних чат-ботів в україномовному середовищі.

СПИСОК ДЖЕРЕЛ

1. Lloret A. Can natural language processing technologies help the digital transformation of local public administrations? *CEUR Workshop Proceedings*. 2024. Vol. 3797. URL: https://ceur-ws.org/Vol-3797/paper13.pdf.

2. Brown T., Mann B., Ryder N. et al. Language models are few-shot learners / H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (eds.). *Advances in Neural Information Proc. Systems*. 2020. Vol. 33. P. 1877–1901. Curran Associates, Inc. URL: https://dl.acm.org/doi/abs/10.5555/3495724.3495883.

3. Virtanen A., Kanerva J., Ilo R. et al. Multilingual is not enough: BERT for Finnish, 2019. arXiv. DOI: https://doi.org/10.48550/arXiv.1912.07076.

4. Howard J., Ruder S. Universal language model fine-tuning for text classification. 2018. arXiv. DOI: https://doi.org/10.48550/arXiv.1801.06146.
5. Chronopoulou A., Peters M., Dodge J. Efficient hierarchical domain adaptation for pretrained language models. 2021. arXiv. URL: https://arxiv.org/abs/2112.08786.
6. Wang S., Fu Y., Kim J. Toward construction-specialized, small language models: The interplay of domain adaptation, model scale and data volume. *Advanced Engineering Informatics*. 2026. Vol. 69. P. 104035. DOI: https://doi.org/10.1016/j.aei.2025.104035.
7. Syvokon O., Romanyshyn M., Kyslyi R. The UNLP 2024 shared task on fine-tuning large language models for Ukrainian. *Proc. of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING*. 2024. P. 67–74. ELRA. URL: https://aclanthology.org/2024.unlp-1.9/.
8. Kiulian A., Polishko A., Khandoga M. et al. From bytes to borsch: Fine-tuning Gemma and Mistral for the Ukrainian language representation. 2024. arXiv. DOI: https://doi.org/10.48550/arXiv.2404.09138.
9. Brown T., Mann B., Ryder N. et al. / H. Larochelle et al. (eds.). *Advances in Neural Information Processing Systems*. 2020. Vol. 33. P. 1877–1901. Curran Associates, Inc. DOI: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
10. Adiwardana D., Luong M.-T., So D.R. et al. Towards a human-like open-domain chatbot. 2020. arXiv. DOI: https://doi.org/10.48550/arXiv.2001.09977.
11. Kryazhych O., Ivanov I., Isak L., Babak O. Development of an approach to chat-bot personalization with generative artificial intelligence when realize an online assistant. *Technology Audit and Production Reserves*. 2025. Vol. 3 (2 (83)). P. 12–19. DOI: https://doi.org/10.15587/2706-5448.2025.326914.
12. Dodge J., Sap M., Marasović A. et al. / M.-F. Moens, X. Huang, L. Specia, S.W.-t. Yih (eds.). *Proc. of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*. 2021. P. 1286–1305. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2021.emnlp-main.98.
13. Gao L., Biderman S., Black S. et al. The Pile: An 800GB dataset of diverse text for language modeling. 2020. arXiv. DOI: https://doi.org/10.48550/arXiv.2101.00027.
14. Roller S., Dinan E., Goyal N. et al. Recipes for building an open-domain chatbot / P. Merlo, J. Tiedemann, R. Tsarfaty (eds.). *Proc. of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021)*. 2021. P. 300–325. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2021.eacl-main.24.

Стаття надійшла до редакції 11.02.2026 / прийнята до друку 28.04.2026