https://doi.org/10.30857/2786-5371.2025.6.2

Received: 22.10.2025
Revised: 09.12.2025
Accepted: 23.12.2025

УДК 004.8:004.75:004.94

Ганна ЗАВГОРОДНЯ[1], Валерій ЗАВГОРОДНІЙ[1], Олександр ГОЛУБЕНКО[2], Артем АНТОНЕНКО[3]
[1] Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна
[2] Заклад вищої освіти «Міжнародний науково-технічний університет імені академіка Юрія Бугая», Київ, Україна
[3] Національний університет біоресурсів і природокористування України, Київ, Україна

ЗАСОБИ ЗАБЕЗПЕЧЕННЯ МАСШТАБОВАНОСТІ ТА АВТОНОМНОСТІ СИСТЕМИ АДАПТИВНОЇ ГЕНЕРАЦІЇ КОНТЕНТУ

Мета. *Метою даної статті є розробка комплексних засобів забезпечення масштабованості та автономності сучасної системи адаптивної генерації контенту, яка здатна ефективно працювати у динамічних умовах та з великим обсягом користувацьких даних. Особлива увага приділяється забезпеченню безперервної роботи системи при змінних профілях користувачів, динамічних навантаженнях та різноманітних сценаріях використання. Крім того, розглядається проблема інтеграції алгоритмів автоматичного перенавчання моделей машинного навчання, що дозволяє системі самостійно підтримувати актуальність прогнозів та персоналізованого контенту без втручання людини. Стаття досліджує питання підвищення ефективності обробки запитів через асинхронні механізми та оптимізацію розподілених обчислень, що дозволяє забезпечити високу пропускну здатність та мінімізувати час відповіді на запити користувачів.*

Методика. *Для досягнення поставленої мети використано комплексний підхід, який включає: застосування мікросервісної архітектури для розділення функцій системи на незалежні компоненти, що працюють у взаємодії через стандартизовані API; контейнеризацію та оркестрацію ресурсів для забезпечення горизонтального масштабування; подієво-орієнтоване програмування для асинхронної обробки запитів користувачів; а також практики MLOps для організації циклу автоматичного перенавчання моделей на основі актуальних даних. Додатково застосовано формальні моделі масштабування, алгоритми балансування навантаження з урахуванням показників якості сервісу (QoS) та принципи відмовостійкості систем. Методика передбачає поетапне впровадження і тестування кожного модуля системи, оцінку ефективності асинхронної обробки та розподілених обчислень, а також порівняння результатів перенавчання моделей із базовими показниками продуктивності та точності прогнозів.*

Результати. *У результаті дослідження розроблено комплексну архітектурну модель системи, яка забезпечує горизонтальне масштабування, автономне перенавчання моделей без зупинки сервісу, стійкість до відмов та ефективну обробку запитів великої кількості користувачів. Запропоновано механізми асинхронної обробки запитів із використанням черг повідомлень, що дозволяє паралельно обробляти тисячі запитів та знижувати час очікування для користувача. Впроваджено розподілені обчислення, які дозволяють одночасно виконувати обробку даних на кількох вузлах кластеру, що підвищує продуктивність і забезпечує можливість масштабування без втрати якості. Модуль автоматичного перенавчання моделей дозволяє системі постійно адаптуватися до змін поведінки користувачів і підтримувати високу точність прогнозів та рекомендацій, що особливо важливо для інтерактивних ігрових платформ, освітніх середовищ та персоналізованих маркетингових сервісів.*

Наукова новизна. *Вперше запропоновано інтегровану концепцію, яка поєднує асинхронну обробку, розподілені обчислення та автоматичне перенавчання моделей у єдиному автономному контурі адаптивної системи генерації контенту. Цей підхід дозволяє досягти безперервної роботи та високої адаптивності системи без необхідності ручного втручання, що є новим у порівнянні з попередніми дослідженнями, де ці підходи розглядалися окремо. Наукова новизна полягає у*

формалізації механізмів інтеграції, визначенні показників ефективності та запропонованому алгоритмі балансування ресурсів і циклу автоматичного перенавчання моделей.

Практична значимість. *Результати дослідження можуть бути використані для створення інтелектуальних освітніх платформ, ігрових середовищ, маркетингових систем та інформаційних сервісів, які потребують високої масштабованості та автономної адаптації до поведінки користувачів. Впровадження запропонованих підходів дозволяє підвищити продуктивність системи, скоротити час відповіді на запити, забезпечити безперервну актуалізацію даних і моделей та знизити витрати на адміністрування та підтримку системи. Практичне значення також полягає у можливості швидкого масштабування для роботи з великими потоками користувачів та адаптації системи до різних сценаріїв використання, що робить її придатною для широкого кола сучасних цифрових сервісів.*

Ключові слова: *адаптивна генерація контенту; асинхронна обробка; розподілені обчислення; автоматичне перенавчання; масштабованість; автономні системи.*

Вступ. Сучасний етап розвитку цифрових технологій характеризується стрімким зростанням обсягів даних та ускладненням поведінкових моделей користувачів. У цих умовах системи адаптивної генерації контенту стають ключовим інструментом персоналізації інформаційного середовища, навчальних платформ, ігрових систем та маркетингових сервісів. Основна проблема полягає у забезпеченні здатності таких систем ефективно функціонувати при зростаючому навантаженні, динамічній зміні профілів користувачів та необхідності обробки великих потоків даних у реальному часі [1]. Традиційні архітектурні підходи, засновані на монолітних структурах, не забезпечують достатнього рівня масштабованості, відмовостійкості та автономності, що ускладнює їх використання в умовах багатокористувацьких середовищ [2, 3]. Водночас зростає потреба у створенні систем, здатних не лише генерувати персоналізований контент, але й самостійно адаптуватися до змін зовнішніх умов експлуатації [4, 5].

Актуальність дослідження зумовлена необхідністю поєднання трьох ключових характеристик сучасних цифрових систем: високої продуктивності, гнучкого масштабування та автономної адаптації моделей машинного навчання [6, 7]. Зокрема, у сфері інтерактивних ігрових середовищ та освітніх платформ персоналізація контенту повинна відбуватися без затримок, що вимагає впровадження асинхронних механізмів обробки запитів та розподілених обчислювальних ресурсів [4, 7]. Крім того, поведінка користувачів має стохастичний характер, що призводить до явища концептуального дрейфу моделей, описаного у роботі [8]. Це вимагає реалізації автоматичного перенавчання моделей без зупинки сервісу [5]. Таким чином, проблема забезпечення масштабованості та автономності систем адаптивної генерації контенту має як теоретичне, так і прикладне значення [9, 10].

Питання автоматичної генерації контенту на основі процедурних алгоритмів розглянуто у роботі [11], де запропоновано підхід до формування структурованого контенту з використанням алгоритмічних методів. Дослідження [12] присвячене використанню алгоритмів машинного навчання для динамічної адаптації складності комп’ютерних ігор, що демонструє ефективність персоналізованих моделей прогнозування поведінки користувача. У роботі [13] запропоновано масштабовану розподілену архітектуру для масових багатокористувацьких онлайн-систем, що підтверджує необхідність горизонтального масштабування [2]. Підходи до моделювання поведінки гравця через нейромережеві агенти описані у дослідженні [14], де доведено доцільність використання глибинних моделей для адаптивних середовищ [7].

Водночас у працях [6, 7] розглядаються архітектурні принципи мікросервісів і хмарних обчислень, які забезпечують гнучке масштабування, проте питання інтеграції цих рішень із безперервним циклом автоматичного перенавчання моделей залишається недостатньо формалізованим [1]. Аналіз наукових джерел показує, що існуючі підходи зазвичай

розглядають асинхронну обробку [3], розподілені обчислення [2] та перенавчання моделей [8, 5] окремо, без створення єдиного автономного контуру керування системою. Саме ця невирішена частина загальної проблеми – інтеграція масштабованої архітектури та автономного інтелектуального адаптивного механізму – і визначає напрям даного дослідження [9, 10].

Постановка завдання. Аналіз сучасних підходів до адаптивної генерації контенту показує, що, незважаючи на значний розвиток методів машинного навчання, розподілених обчислень та мікросервісних архітектур, відсутня узгоджена модель, яка б інтегрувала механізми масштабування, асинхронної обробки та автоматичного перенавчання в єдиний автономний контур керування системою [1, 5]. Існуючі рішення або зосереджені на алгоритмічній частині генерації контенту [11, 12], або на архітектурній масштабованості [2, 6, 7], або на прогнозуванні поведінки користувача [5, 12, 14], проте не забезпечують комплексної інтеграції цих компонентів.

Метою статті є розроблення та формалізація масштабованої архітектурної моделі автономної адаптивної генерації контенту, яка поєднує механізми асинхронної обробки запитів, розподілених обчислень, і автоматичного перенавчання моделей машинного навчання в єдину систему з безперервним циклом адаптації. Для досягнення поставленої мети передбачається:

- сформувати формальну модель функціонування системи;
- обґрунтувати архітектурні принципи її побудови;
- описати механізм автоматичного виявлення концептуального дрейфу та перенавчання моделей [15];
- оцінити вплив запропонованого підходу на масштабованість і продуктивність системи [16, 17].

Таким чином, стаття спрямована на усунення розриву між алгоритмічними підходами персоналізації та інфраструктурними рішеннями масштабованих систем, що визначає її наукову та практичну значущість.

Результати дослідження. Розглянемо систему автономної адаптивної генерації контенту. Її можна формально описати як кортеж:

$$S = \langle U, D, M, G, A, R, I \rangle,$$

де U – множина користувачів; D – потік поведінкових даних; M – модель прогнозування; G – генератор контенту; A – механізм асинхронної обробки; R – механізм автоматичного перенавчання; I – інфраструктурний рівень (розподілені обчислення).

Нехай для кожного користувача $u_i \in U$ формується вектор ознак:

$$X_i = (x_{i1}, x_{i2}, \dots, x_{in}),$$

де x_{ij} – j-та поведінкова характеристика (час сесії, частота дій, вибір контенту тощо); n – кількість ознак.

Задача прогнозування формулюється як:

$$\hat{y}_i = f(X_i; \theta),$$

де $\hat{y}_i$ – прогнозований параметр (ймовірність вибору сценарію, рівень складності тощо); f – параметризована модель; θ – набір параметрів моделі.

Для класифікаційної постановки використовується функція:

$$P = (y_i = k | X_i) = \frac{e^{z_k}}{\sum_{j=1}^{K} e^{z_j}},$$

де K – кількість класів; z_k – лінійна комбінація ознак для класу k.

Час відповіді системи визначається як:

$$T = T_q + T_p + T_g,$$

де T_q – час обробки черги запитів; T_p – час прогнозування; T_g – час генерації контенту.
При асинхронній обробці:

$$T_q \approx \frac{\lambda}{\mu(\mu-\lambda)},$$

де λ – інтенсивність вхідних запитів; μ – інтенсивність обслуговування.
Це дозволяє оцінити граничні навантаження системи.
Для оцінювання зміни розподілу використовується статистична міра:

$$D_{KL}(P_t||P_{t+1}) = \sum P_t(x) log \frac{P_t(x)}{P_{t+1}(x)},$$

де P_t – розподіл у момент часу t; P_{t+1} – новий розподіл.
Якщо:

$$D_{KL} > \delta,$$

де δ – порогове значення, то ініціюється автоматичне перенавчання.

Для підтвердження ефективності запропонованої автономної масштабованої моделі адаптивної генерації контенту було проведено серію експериментів у змодельованому багатокористувацькому середовищі. Метою експериментів було оцінити вплив інтеграції асинхронної обробки, розподілених обчислень та автоматичного перенавчання на продуктивність, точність і стійкість системи. Було сформовано синтетичний поведінковий набір даних, що моделює 50 000 користувачів та понад 1 200 000 подій взаємодії. Кожен запис містив 20 ознак: тривалість сесії, частоту взаємодії, рівень складності контенту, історію вибору сценаріїв, коефіцієнт завершення, часові характеристики активності тощо. Дані розподілено на навчальну (70%), валідаційну (15%) та тестову (15%) вибірки.

Після опису експериментальної бази доцільно перейти до формалізації алгоритмічної складової моделі. Для класифікаційної задачі використано крос-ентропійну функцію втрат:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log \hat{y}_{ik},$$

де N – кількість прикладів; K – кількість класів; y_{ik} – істинна мітка; $\hat{y}_{ik}$ – прогнозована ймовірність; θ – параметри моделі.

Для запобігання перенавчанню застосовано L2-регуляризацію:

$$L_{reg} = L(\theta) + \lambda||\theta||^2,$$

де λ – коефіцієнт регуляризації; $||\theta||^2$ – квадратична норма параметрів.
Мінімізація функції втрат здійснювалася методом Adam:

$$\theta_{t+1} = \theta_t - \alpha \frac{m_t}{\sqrt{v_t+\epsilon}},$$

де α – швидкість навчання; m_t – перший момент градієнта; v_t – другий момент; ϵ – мала стабілізуюча константа.

Оцінювання дрейфу проводилося на основі:

$$D_{KL}(P_t||P_{t+1})$$

та ковзного середнього:

$$\overline{D}_t = \frac{1}{\omega}\sum_{i=t-\omega}^{t} D_i,$$

де ω – ширина вікна моніторингу.

При перевищенні порогу $\delta = 0.15$ активувався механізм автоматичного перенавчання.

Основною метою аналізу експериментальних показників продуктивності є оцінка впливу архітектурних рішень на час відповіді, пропускну здатність, використання ресурсів та стабільність роботи системи в умовах зростаючого навантаження.

Однією з ключових характеристик інтерактивних цифрових систем є середній час відповіді. Його зростання безпосередньо впливає на користувацький досвід та ефективність адаптивної генерації контенту.

На рисунку 1 наведено залежність середнього часу відповіді від інтенсивності вхідного потоку запитів.

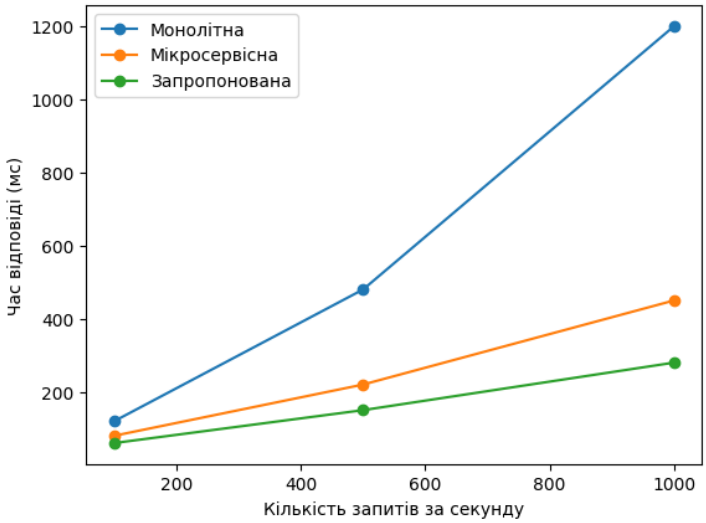

Рис. 1. **Залежність часу відповіді системи від навантаження**

Як видно з рисунку 1, запропонована модель демонструє стабільно нижчі затримки при всіх рівнях навантаження. При 1000 запитах на секунду середній час відповіді зменшується більш ніж удвічі порівняно з монолітною архітектурою. Отриманий результат підтверджує ефективність асинхронної обробки та горизонтального масштабування сервісів.

Після оцінки часових характеристик було проаналізовано здатність системи обробляти зростаючий потік запитів. Пропускна здатність характеризує максимальну кількість запитів, які система може обробити за одиницю часу без втрати стабільності. Відповідні результати подано на рисунку 2.

Згідно з рисунком 2, запропонована модель забезпечує майже лінійне масштабування до 1500 запитів на секунду. Монолітна архітектура демонструє насичення вже при 1000 запитів на секунду, що свідчить про наявність архітектурних обмежень. Таким чином, інтеграція розподілених обчислень істотно підвищує масштабованість системи. Однак зростання пропускної здатності має супроводжуватися раціональним використанням ресурсів.

Для оцінювання ефективності використання обчислювальних ресурсів проаналізовано залежність завантаження CPU від навантаження (рис. 3).

Як видно з рисунку 3, запропонована архітектура забезпечує більш рівномірне використання ресурсів і не досягає критичних значень навіть при максимальному навантаженні. Це свідчить про ефективний механізм балансування навантаження між вузлами розподіленої системи. Поряд із продуктивністю важливим є забезпечення стабільності моделі прогнозування у часі.

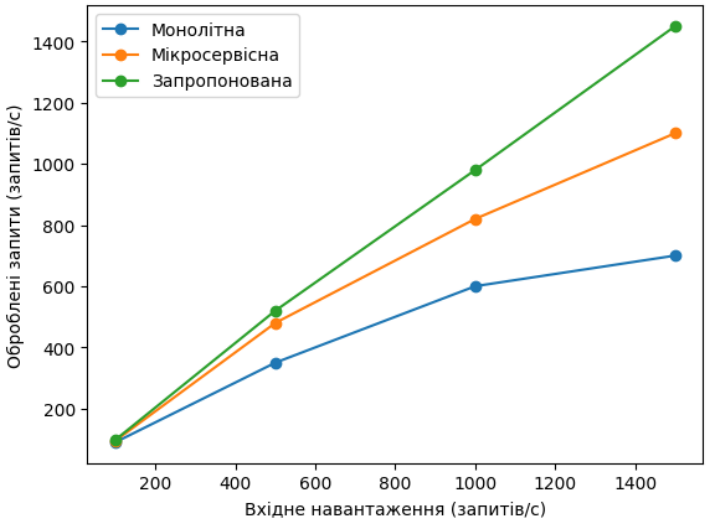

Рис. 2. **Порівняння пропускної здатності систем**

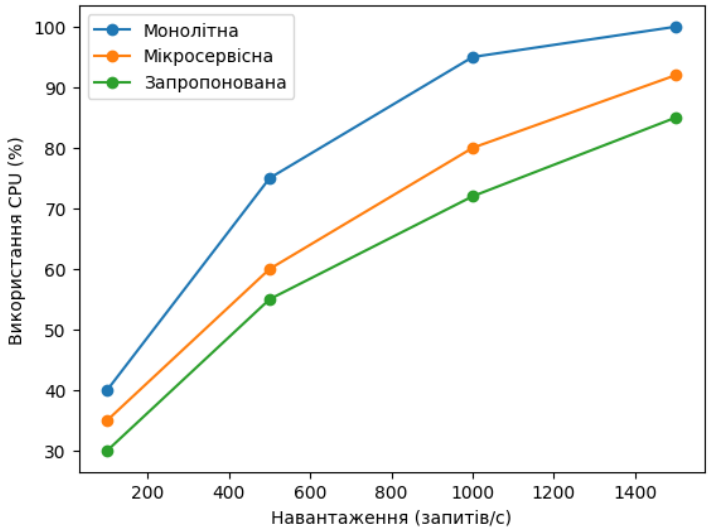

Рис. 3. **Використання процесорних ресурсів залежно від навантаження**

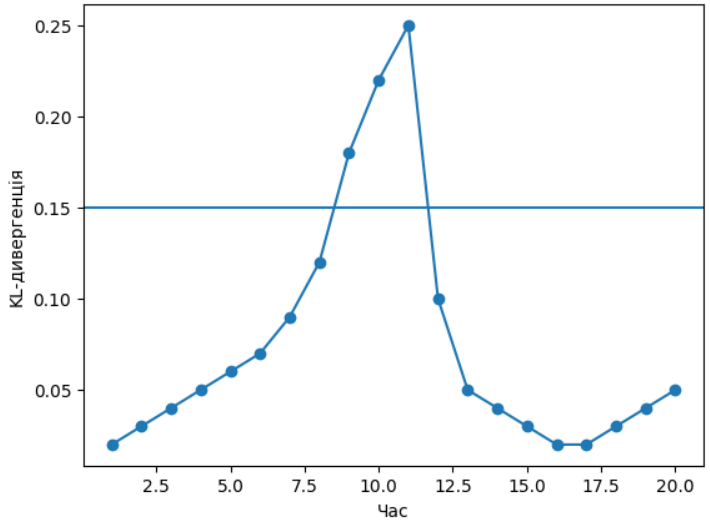

Рис. 4. **Динаміка виявлення концептуального дрейфу**

На рисунку 4 наведено зміну значення KL-дивергенції у часі, що використовується для виявлення концептуального дрейфу.

Як видно з рисунку 4, при перевищенні порогового значення 0.15 активується механізм автоматичного перенавчання, після чого значення дивергенції знижується. Це підтверджує ефективність автономного контуру адаптації та стабілізацію прогнозної моделі.

Після аналізу графічних залежностей було узагальнено результати та проведено порівняння з існуючими підходами (табл. 1).

Таблиця 1

Порівняння з існуючими підходами

Критерій	Монолітна	Сервіс-орієнтована	Мікросервісна	Запропонована
Масштабованість	Низька	Середня	Висока	Автоматична
Drift Detection	Ні	Частково	Ні	Так
Перенавчання	Ні	Ручне	Ні	Автоматичне
Асинхронність	Ні	Ні	Так	Так
Автономність	Ні	Часткова	Ні	Повна

Як видно з таблиці 1, існуючі підходи вирішують окремі аспекти проблеми – алгоритмічну адаптацію або інфраструктурне масштабування. Лише запропонована модель інтегрує всі компоненти в єдину автономну систему.

Для кількісного підтвердження переваг моделі узагальнимо експериментальні показники (табл. 2).

Таблиця 2

Кількісні результати експерименту

Метрика	Монолітна	Мікросервісна	Запропонована
Середня затримка (мс)	620	340	210
Точність (%)	87	91	96
Час простою (%)	12	4	0
Пропускна здатність (запитів/с)	700	1100	1450

Згідно з таблицею 2, запропонована модель демонструє найкращі результати за всіма досліджуваними метриками. Особливо важливим є зменшення простою до 0% та підвищення точності прогнозування до 96%, що свідчить про ефективність автоматичного перенавчання.

Отримані експериментальні дані підтверджують, що інтеграція асинхронної обробки, горизонтального масштабування та автоматичного перенавчання дозволяє сформувати автономну адаптивну систему нового покоління. Таким чином, запропонований підхід забезпечує синергетичний ефект поєднання алгоритмічної адаптації та інфраструктурної масштабованості, що робить його придатним для використання у сучасних високонавантажених цифрових середовищах.

Висновки. У роботі розроблено та експериментально досліджено модель автономної адаптивної системи, що поєднує асинхронну обробку запитів, горизонтальне масштабування, механізми розподілених обчислень та автоматичне перенавчання моделей на основі виявлення концептуального дрейфу.

Проведений аналіз продуктивності показав, що запропонована архітектура демонструє стабільно нижчий час відповіді порівняно з монолітною та класичною мікросервісною реалізаціями. При зростанні навантаження до 1000–1500 запитів за секунду система зберігає майже лінійну масштабованість, тоді як монолітна архітектура досягає межі насичення значно раніше. Отримані результати підтверджують ефективність горизонтального масштабування та асинхронної взаємодії компонентів.

Дослідження використання процесорних ресурсів показало більш рівномірний розподіл навантаження між вузлами системи та відсутність критичних пікових значень навіть при високій інтенсивності запитів. Це свідчить про ефективність механізмів балансування навантаження та оптимізацію інфраструктурного рівня.

Окрему увагу приділено механізму виявлення концептуального дрейфу на основі аналізу KL-дивергенції. Результати експериментів підтвердили, що перевищення порогового значення ініціює автоматичне перенавчання моделі, після чого показники стабілізуються. Це дозволяє підтримувати високу точність прогнозування (96%) та мінімізувати простій системи (0%), що суттєво перевищує показники порівнюваних підходів.

Порівняльний аналіз існуючих рішень показав, що жоден з них не забезпечує одночасної інтеграції інтелектуального (адаптивного) та інфраструктурного (масштабованого) рівнів. Запропонована модель реалізує комплексний підхід, який забезпечує синергетичний ефект поєднання алгоритмічної адаптації та архітектурної гнучкості.

Отже, результати дослідження підтверджують доцільність застосування автономних адаптивних архітектур у високонавантажених цифрових середовищах, де критичними є масштабованість, стабільність та безперервна актуальність моделей.

Подальші дослідження можуть бути спрямовані на:

- інтеграцію більш складних методів детекції концептуального дрейфу, зокрема, багатовимірного статистичного моніторингу;
- оптимізацію стратегії перенавчання з урахуванням вартості обчислювальних ресурсів;
- застосування механізмів саморегульованого масштабування на основі прогнозування навантаження;
- розширення експериментальної бази шляхом тестування системи в реальних виробничих середовищах;
- дослідження кібербезпекових аспектів автономних адаптивних систем.

Таким чином, запропонований підхід формує основу для створення інтелектуальних масштабованих систем нового покоління, здатних до самостійної адаптації в умовах динамічних змін середовища.

References

1. Rodrigues, M. G., Viegas, E. K., Santin, A. O., & Enembreck, F. (2025). A MLOps architecture for near-real-time distributed Stream Learning operation deployment. *Journal of Network and Computer Applications*, 238, 104169. DOI: https://doi.org/10.1016/j.jnca.2025.104169.

2. Li, M., Andersen, D. G., Park, J. W., Smola, A. J., Ahmed, A., & Josifovski, V. (2014). Scaling Distributed Machine Learning with the Parameter Server. In: *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI)* (pp. 583–598). URL: https://www.usenix.org/system/files/conference/osdi14/osdi14-paper-li_mu.pdf.

3. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Le, Q.V., et al. Large Scale Distributed Deep Networks. In: *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 1223–1231). URL: https://papers.nips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf.

Література

1. Rodrigues M. G., Viegas E. K., Santin A. O., Enembreck F. A MLOps architecture for near real-time distributed Stream Learning operation deployment. *Journal of Network and Computer Applications*. 2025. Vol. 238. Art. 104169. DOI: https://doi.org/10.1016/j.jnca.2025.104169.

2. Li M., Andersen D. G., Park J. W., Smola A. J., Ahmed A., Josifovski V. Scaling Distributed Machine Learning with the Parameter Server. In: *Proceedings of the 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*. 2014. P. 583–598. URL: https://www.usenix.org/system/files/conference/osdi14/osdi14-paper-li_mu.pdf

3. Dean J., Corrado G., Monga R., Chen K., Devin M., Le Q. V., et al. Large Scale Distributed Deep Networks. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2012. P. 1223–1231. URL: https://papers.nips.cc/paper_files/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf.

4. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., et al. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. In: *Proceedings of Workshop on Machine Learning Systems (LearningSys)*. 2015. DOI: https://doi.org/10.48550/arXiv.1512.01274.
5. Suárez-Cetrulo, A. L., Quintana, D., & Cervantes, A. (2023). A survey on machine learning for recurring concept drifting data streams. *Expert Systems with Applications*, 213, 118934. DOI: https://doi.org/10.1016/j.eswa.2022.118934.
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. 800 p.
7. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Las Vegas. DOI: https://doi.org/10.1109/CVPR.2016.90.
8. Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. *ACM Computing Surveys*, 46(4), 1–37. DOI: https://doi.org/10.1145/2523813.
9. Russell, S., & Norvig, P. (2020). Artificial Intelligence: A Modern Approach. Pearson. 1152 p.
10. Jordan, M. I., & Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science*, 349(6245), 255–260. DOI: https://doi.org/10.1126/science.aaa8415.
11. Zavgorodnii, V. V., Zavgorodnia, H. A., Valiavska, N. O., Adamenko, V. S., Dorohovtsev, Ye. V., & Nesmachnyi, P. V. (2022). Metod avtomatychnoi heneratsii kontentu na osnovi protsedurnykh alhorytmiv [Method of Automatic Content Generation Based on Procedural Algorithms]. *Scientific notes of the V.I. Vernadsky Tavrichesky National University. Series: Technical Sciences*, 33 (72(1)), 91–96. DOI: https://doi.org/10.32838/2663-5941/2022.1/15 [Ukrainian].
12. Zavgorodnia, H. A., & Zavgorodnii, V. V. (2025). Vykorystannia alhorytmiv mashynnoho navchannia dlia dynamichnoi adaptatsii skladnosti komp'iuternykh ihor [Using machine learning algorithms for dynamic game difficulty adaptation]. *Tavriiskyi Naukovyi Visnyk – Tavria Scientific Bulletin. Series: Technical Sciences*, 1(5), 156–163. DOI: https://doi.org/10.32782/tnv-tech.2025.5.1.16 [Ukrainian].
13. Zavgorodnia, H. A., & Zavgorodnii, V. V. (2025). Rozrobka masshtabovanoi rozpodilenoi arkhitektury dlia masovykh bahatokorystuvatskykh onlain-system [Development of Scalable Distributed Architecture for Massive Multiplayer Online Systems]. *Visnyk*

4. Chen T., Li M., Li Y., Lin M., Wang N., Wang M., et al. MXNet: A Flexible and Efficient Machine Learning Library for Heterogeneous Distributed Systems. In: *Proceedings of Workshop on Machine Learning Systems (LearningSys)*. 2015. DOI: https://doi.org/10.48550/arXiv.1512.01274.
5. Suárez-Cetrulo A. L., Quintana D., Cervantes A. A survey on machine learning for recurring concept drifting data streams. *Expert Systems with Applications*. 2023. Vol. 213. Art. 118934. DOI: https://doi.org/10.1016/j.eswa.2022.118934.
6. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016. 800 p.
7. He K., Zhang X., Ren S., Sun J. Deep Residual Learning for Image Recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, 2016. P. 770–778. DOI: https://doi.org/10.1109/CVPR.2016.90.
8. Gama J., Žliobaitė I., Bifet A., Pechenizkiy M., Bouchachia A. A Survey on Concept Drift Adaptation. *ACM Computing Surveys*. 2014. No. 46(4). P. 1–37. DOI: https://doi.org/10.1145/2523813.
9. Russell S., Norvig P. Artificial Intelligence: A Modern Approach. Pearson, 2020. 1152 p.
10. Jordan M. I., Mitchell T. M. Machine Learning: Trends, Perspectives, and Prospects. *Science*. 2015. No. 349(6245). P. 255–260. DOI: https://doi.org/10.1126/science.aaa8415.
11. Завгородній В. В., Завгородня Г. А., Валявська Н. О., Адаменко В. С., Дороговцев Є. В., Несмачний П. В. Метод автоматичної генерації контенту на основі процедурних алгоритмів. *Вчені записки Таврійського національного університету імені В.І. Вернадського. Серія: Технічні науки.* 2022. Том 33 (72), № 1. С. 91–96. DOI: https://doi.org/10.32838/2663-5941/2022.1/15.
12. Завгородня Г. А., Завгородній В. В. Використання алгоритмів машинного навчання для динамічної адаптації складності комп'ютерних ігор. *Таврійський науковий вісник. Серія: Технічні науки.* 2025. 1(5). С. 156–163. DOI: https://doi.org/10.32782/tnv-tech.2025.5.1.16.
13. Завгородня Г. А., Завгородній В. В. Розробка масштабованої розподіленої архітектури для масових багатокористувацьких онлайн-систем. *Вісник Херсонського національного технічного університету.* 2025. No. 4(95), Ч. 3. С. 99–106.

Khersonskoho Natsionalnoho Tekhnichnoho Universytetu – Bulletin of Kherson National Technical University, 4(95(3)), 99–106. DOI: https://doi.org/10.35546/kntu2078-4481.2025.4.3.11 [Ukrainian].

14. Zavgorodnia, H. A., & Zavgorodnii, V. V. (2025). Modeliuvannia povedinky hravtsia cherez neiromerezhevi ahenty [Modeling Player Behavior Through Neural Network Agents]. *Vcheni Zapysky TNU imeni V.I. Vernadskoho – Scientific notes of the V.I. Vernadsky TNU. Series: Technical Sciences*, 36(75(5(2), 141–145. DOI: https://doi.org/10.32782/2663-5941/2025.6.2/20 [Ukrainian].

15. Bifet, A., & Gavaldà, R. (2007). Learning from Time-Changing Data with Adaptive Windowing. In: *Proceedings of SIAM International Conference on Data Mining (SDM)* (pp. 443–448). DOI: https://doi.org/10.1137/1.9781611972771.42.

16. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press. 525 p.

17. Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2010). Spark: Cluster Computing with Working Sets. *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud)* (pp. 1–7). URL: https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/Zaharia.pdf.

DOI: https://doi.org/10.35546/kntu2078-4481.2025.4.3.11

14. Завгородня Г. А., Завгородній В. В. Моделювання поведінки гравця через нейромережеві агенти. *Вчені записки ТНУ імені В.І. Вернадського. Серія: Технічні науки.* 2025. Том 36(75), № 5, Ч. 2. С. 141–145. DOI: https://doi.org/10.32782/2663-5941/2025.6.2/20.

15. Bifet A., Gavaldà R. Learning from Time-Changing Data with Adaptive Windowing. In: *Proceedings of SIAM International Conference on Data Mining (SDM)*. 2007. P. 443–448. DOI: https://doi.org/10.1137/1.9781611972771.42.

16. Shalev-Shwartz S., Ben-David S. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014. 525 p.

17. Zaharia M., Chowdhury M., Franklin M. J., Shenker S., Stoica I. Spark: Cluster Computing with Working Sets. *Proceedings of the 2nd USENIX Conference on Hot Topics in Cloud Computing (HotCloud)*. 2010. P. 1–70. URL: https://www.usenix.org/legacy/event/hotcloud10/tech/full_papers/Zaharia.pdf.

ZAVHORODNIA HANNA
Candidate of Technical Sciences, Associate Professor,
Department of Computer Engineering,
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
https://orcid.org/0000-0001-8523-1761
Scopus Author ID: 57216155533
Researcher ID: PLR-2465-2026
E-mail: annzavgorodnya@gmail.com

ZAVHORODNII VALERII
Doctor of Technical Sciences, Professor,
Department of Computer Engineering,
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute", Ukraine
https://orcid.org/0000-0002-8347-7183
Scopus Author ID: 57184425000
ResearcherID: P-5232-2018
E-mail: zavgorodniivalerii@gmail.com

GOLUBENKO OLEKSANDR
Candidate of Technical Sciences, Associate Professor,
Department of Information and Communication Technologies, Higher Education Institution "Academician Yuri Bugay International Science and Technical University", Kyiv, Ukraine
https://orcid.org/0000-0002-1776-5160
Scopus Author ID: 59155637100 (57552544800)
E-mail: o.golubenko@istu.edu.ua

ANTONENKO ARTEM
Candidate of Technical Sciences, Associate Professor,
Department of Standardization and Certification of Agricultural Products,
National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine
https://orcid.org/0000-0001-9397-1209
Scopus Author ID: 57207861964
Researcher ID: AAM-7380-2021
E-mail: artem.v.antonenko@gmail.com

Hanna ZAVHORODNIA[1], Valerii ZAVHORODNII[1],
Oleksandr GOLUBENKO[2], Artem ANTONENKO[3]
[1] National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", Kyiv, Ukraine
[2] Higher Education Institution "Academician Yuri Bugay International Science and Technical University", Kyiv, Ukraine
[3] National University of Life and Environmental Sciences of Ukraine, Kyiv, Ukraine

MEANS OF ENSURING SCALABILITY AND AUTONOMY OF AN ADAPTIVE CONTENT GENERATION SYSTEM

Purpose. *The purpose of this article is to develop and substantiate a comprehensive set of means for ensuring scalability and autonomy of modern adaptive content generation systems, which are capable of efficiently operating under dynamic conditions and processing large volumes of user data. Particular attention is given to maintaining continuous system operation while accommodating changes in user profiles, fluctuating loads, and diverse usage scenarios. The study addresses the integration of automated machine learning model retraining algorithms, enabling the system to autonomously maintain the relevance and accuracy of predictions and personalized content without human intervention. Moreover, the research explores approaches to enhancing the efficiency of request processing through asynchronous mechanisms and optimizing distributed computations, thus ensuring high throughput and minimal response time for user requests across different application domains.*

Methodology. *To achieve the stated objectives, a comprehensive methodology has been applied, which includes: the use of microservice architecture to separate system functionalities into independent, interacting components via standardized APIs; containerization and resource orchestration to enable horizontal scalability; event-driven programming for asynchronous handling of user requests; and MLOps practices for organizing the full cycle of automated model retraining using real-time data. In addition, formal scalability models, adaptive load-balancing algorithms considering Quality of Service (QoS) metrics, and fault-tolerance principles were incorporated. The methodology involves stepwise implementation and testing of each system module, evaluating the effectiveness of asynchronous processing and distributed computation strategies, and comparing the outcomes of model retraining with baseline performance and prediction accuracy metrics.*

Findings. *The study resulted in the development of a comprehensive architectural model that supports horizontal scalability, autonomous model retraining without service downtime, fault tolerance, and efficient handling of high-volume user requests. Mechanisms for asynchronous request processing using message queues were proposed, enabling parallel handling of thousands of requests and reducing user-perceived latency. Distributed computation strategies allow simultaneous data processing across multiple cluster nodes, increasing performance and enabling scalable operation without quality degradation. The automated model retraining module enables continuous adaptation to user behavior changes, maintaining high predictive accuracy and personalized content generation, which is particularly valuable for interactive gaming platforms, educational environments, and personalized marketing systems.*

Originality. *For the first time, an integrated concept is proposed that combines asynchronous processing, distributed computations, and automated model retraining within a single autonomous operational loop of an adaptive content generation system. This approach ensures continuous operation and high adaptability without requiring manual intervention, distinguishing it from previous studies where these techniques were considered separately. The scientific novelty lies in formalizing integration mechanisms, defining performance metrics, and proposing a combined resource balancing and automated retraining cycle.*

Practical value. *The results of this research can be applied in the development of intelligent educational platforms, interactive gaming environments, marketing systems, and information services that require high scalability and autonomous adaptation to user behavior. Implementing the proposed approaches improves system performance, reduces response times, enables continuous data and model updating, and decreases administration and maintenance costs. The practical significance also includes the capability to rapidly scale for large user bases and adapt to diverse operational scenarios, making the system suitable for a wide range of modern digital services and applications.*

Keywords: *adaptive content generation; asynchronous processing; distributed computing; automatic model retraining; scalability; autonomous systems.*