ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

УДК 004.056.55 **doi: 10.32620/aktt.2025.6.09**

О. В. ВДОВІТЧЕНКО, Д. Ю. РУДЕНОК

Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна

МЕТОД АДАПТИВНОЇ ГЕНЕРАЦІЇ КОНТЕНТУ В МОБІЛЬНИХ ЗАСТОСУНКАХ НА ОСНОВІ ПЕРСОНАЛІЗОВАНИХ МОДЕЛЕЙ ГЛИБОКОГО НАВЧАННЯ

Предметом *вивчення в статті є процеси адаптивної генерації та персоналізації мультимедійного контенту в мобільних інформаційних системах, що функціонують в умовах обмежених обчислювальних ресурсів та періодичної відсутності стабільного каналу зв'язку. Особлива увага приділяється системам інформаційної підтримки екіпажу та бортовим розважальним системам, де критичними вимогами є автономність роботи, мінімізація енергоспоживання та захист конфіденційності польотних даних.* ***Метою*** *роботи є розробка комплексного методу адаптивної генерації контенту, який органічно поєднує глобальні знання великих нейронних моделей з локальною персоналізацією та контекстною адаптацією в реальному часі для підвищення релевантності інформації та енергоефективності автономних мобільних пристроїв.* ***Завдання****: провести аналіз існуючих методів автоматичної генерації контенту та виявити їхні архітектурні обмеження у специфічних мобільних середовищах; розробити формальну стохастичну модель системи, що включає глобальний рівень узагальнення знань, локальний рівень персоналізації та рівень контекстної адаптації; обґрунтувати вибір методів структурної оптимізації глибоких нейронних мереж для забезпечення роботи в режимі офлайн; розробити метод адаптивної генерації на основі персоналізованого федеративного навчання та алгоритмів контекстної оптимізації; реалізувати експериментальну модель системи та перевірити її ефективність на відкритих текстових наборах даних, що моделюють роботу із запитами в умовах обмеженого контексту. Використовуваними* ***методами*** *є: методи глибокого навчання для побудови компактних генеративних трансформерних моделей типу; методи федеративного навчання для децентралізованого оновлення вагових коефіцієнтів моделі без передачі первинних даних на сервер; методи навчання з підкріпленням, для динамічної адаптації стратегії генерації до поточної фази польоту або поведінки користувача; методи стиснення нейронних мереж для зменшення обчислювального навантаження. Отримані такі* ***результати****. Розроблено та програмно реалізовано метод адаптивної генерації, який інтегрує персоналізоване федеративне навчання з механізмом контекстної оптимізації. В ході експериментальних досліджень, проведених на текстових масивах даних, що імітують технічну документацію та запити екіпажу, встановлено, що використання запропонованого підходу забезпечує підвищення якості генерації контенту за метрикою BLEU на 38 % порівняно з базовою централізованою моделлю. Завдяки оптимізації архітектури нейромережі досягнуто зменшення затримки формування відповіді на 34 % та зниження споживання енергії мобільним пристроєм на 10–15 %, що є критичним показником для автономних бортових систем. Підтверджено високий рівень захисту конфіденційності даних (втрата приватності $\varepsilon < 1$) при використанні механізмів диференційної приватності.* ***Висновки****. Запропонований метод дозволяє вирішити науково-прикладну проблему впровадження інтелектуальних генеративних сервісів у замкнені екосистеми з високими вимогами до приватності та автономності, такі як авіаційні мобільні застосунки. Інтеграція федеративного навчання з локальною адаптацією створює умови для автономного самонавчання системи без необхідності постійного звернення до хмарних обчислювальних потужностей.* ***Наукова новизна*** *отриманих результатів полягає в тому, що вперше формалізовано метод адаптивної генерації контенту як єдину стохастичну систему з контекстною самоадаптацією, яка, на відміну від існуючих розрізнених рішень, динамічно оптимізує стратегію генерації через узагальнену функцію втрат, що враховує якість, персоналізацію та контекстний відгук середовища.*

Ключові слова: *адаптивна генерація контенту; персоналізація; федеративне навчання; глибоке навчання; мобільні застосунки; електронний портфель пілота; контекстна адаптація; навчання з підкріпленням; стохастична оптимізація; штучний інтелект на пристрої.*

1. Вступ

Сучасний розвиток мобільних технологій відкриває нові можливості для критично важливих галузей, зокрема авіації. Мобільні додатки, такі як «електронні портфелі пілотів» (Electronic Flight Bags, EFB) або системи розваг на борту (In-Flight Entertainment, IFE), стають основними інструментами взаємодії з інформацією. Користувачі цих систем – пілоти та бортпровідники – потребують миттєвого доступу до релевантного контенту, адаптованого до поточного контексту (етапу польоту, локації, технічного стану).

1.1. Мотивація дослідження

Специфіка експлуатації мобільних застосунків в авіації накладає жорсткі обмеження, які відрізняють їх від звичайних користувацьких програм. По-перше, відсутність або нестабільність інтернет-з'єднання під час польоту унеможливлює використання хмарних генеративних моделей (Cloud AI). По-друге, критичними є вимоги до енергоефективності пристроїв (економія заряду EFB у тривалих рейсах) та приватності даних екіпажу. Традиційні статичні інтерфейси часто перевантажують пілота інформацією. Використання алгоритмів глибокого навчання (Deep Learning) дозволило б автоматизувати підготовку контенту – генерувати короткі витяги з інструкцій або звіти. Однак це вимагає рішень класу On-device AI, які працюють автономно. Актуальність дослідження зумовлена необхідністю створення методів генерації, які працюють безпосередньо на бортовому пристрої, адаптуються до контексту та не передають конфіденційні дані в хмару [1].

У відповідь на ці виклики активно розвиваються методи персоналізації (через федеративне навчання, адаптивні промпти, локальне донавчання) та оптимізації моделей (knowledge distillation, quantization, pruning), що дозволяють переносити можливості генеративних систем у мобільні застосунки. Водночас питання адаптивної генерації контенту в реальному часі з урахуванням поточного контексту користувача (локації, часу доби, динаміки поведінки) залишається відкритим і потребує подальших досліджень.

Особливу складність становить збалансування трьох ключових аспектів:

1) якості згенерованого контенту;

2) швидкодії та енергоефективності моделі на мобільному пристрої;

3) рівня конфіденційності персональних даних користувача.

Для вирішення цієї проблеми доцільно поєднувати можливості глибоких нейронних мереж із механізмами адаптації у реальному часі та децентралізованими методами навчання. Такий підхід дозволяє системі динамічно змінювати стратегію генерації залежно від поточного контексту, не передаючи приватні дані на сервер.

Таким чином, актуальність даного дослідження зумовлена необхідністю розроблення ефективних методів адаптивної генерації контенту, здатних забезпечити персоналізований користувацький досвід у мобільних застосунках при збереженні високої продуктивності та конфіденційності.

Метою дослідження є розробка моделей та методів автоматичної генерації персоналізованого контенту для мобільних застосунків на основі алгоритмів глибокого навчання з урахуванням обмежених ресурсів пристрою та вимог до приватності даних користувача.

1.2. Аналіз останніх досліджень й публікацій

Останні роки характеризуються стрімким розвитком систем автоматичної генерації контенту, що базуються на алгоритмах глибокого навчання. Значна частина досліджень присвячена покращенню якості згенерованого матеріалу, зниженню обчислювальних витрат та впровадженню персоналізації. Наукові результати у цій сфері демонструють тенденцію переходу від традиційних моделей рекомендацій до адаптивних систем, здатних формувати контент у реальному часі відповідно до профілю користувача.

Генеративні моделі для створення контенту – це моделі генеративного типу, великі мовні моделі (Large Language Models, LLM), генеративно-змагальні мережі (GAN), та дифузійні моделі. Трансформерна архітектура, запропонована Vaswani et al. (2017), стала основою сучасних систем на кшталт GPT-4, LLaMA-3, Gemini 2.5 тощо. Ці моделі навчилися ефективно узагальнювати семантичні залежності між елементами тексту та створювати цілісний контент на рівні, близькому до людського.

Однак масштаб таких моделей (мільярди параметрів) робить їх малопридатними для безпосереднього використання у мобільних застосунках. У відповідь на це активно розвиваються методи оптимізації моделей. Knowledge Distillation – перенесення знань з великої моделі у меншу без значної втрати якості [2]. Quantization – зменшення розрядності числових представлень ваг [3]. Pruning – усунення малозначущих з'єднань у нейронних мережах [4].

Ці підходи дозволяють створювати "легкі" генеративні моделі, придатні для виконання на мобільних процесорах без значного зниження точності.

Персоналізація та адаптація моделей, стосується персоналізації контенту, тобто врахування індивідуальних характеристик користувачів при генерації результатів.

Традиційні методи персоналізації базувались на

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

збиранні даних на сервері та централізованому донавчанні моделі, однак це створює ризики порушення конфіденційності [5].

Сучасним рішенням стала концепція федеративного навчання (Federated Learning, FL), у якій модель навчається колективно на пристроях користувачів, а до сервера передаються лише оновлення ваг без персональних даних [6].

Пізніші модифікації – Personalized Federated Learning (PFL), дають змогу поєднувати глобальну модель з локальними адаптаційними параметрами, що підвищує індивідуальну точність [7].

У контексті генерації контенту персоналізація може здійснюватися різними способами: через локальні промпти, мета-навчання, або адаптаційні шари. Такі методи дозволяють мобільному застосунку налаштовувати модель під конкретного користувача без потреби повного перенавчання.

Методи онлайн-адаптації або адаптація систем у реальному часі. У цьому контексті поширеним підходом є застосування contextual bandits або reinforcement learning (RL), які дозволяють моделі приймати рішення на основі поточного контексту користувача, отримуючи зворотний зв'язок у процесі взаємодії [8].

Завдяки такій динамічній оптимізації система може змінювати стратегію генерації контенту відповідно до змін поведінки користувача, підвищуючи ефективність і залучення.

Відомі дослідження, наприклад Deep Deep reinforcement learning in recommender systems [9] демонструють ефективність таких підходів для новинних, освітніх та розважальних платформ. Водночас проблема інтеграції RL-методів у мобільне середовище залишається відкритою через обмеження ресурсів та необхідність обробки даних локально.

Оптимізація моделей для мобільних середовищ. Адаптація глибоких моделей до обчислювальних обмежень мобільних пристроїв. Серед найбільш поширених підходів:

– створення компактних архітектур (MobileNet, EfficientNet, TinyBERT);

– використання edge computing — часткове перенесення обчислень на проміжні сервери;

– застосування hybrid inference — розподіл виконання між клієнтом і хмарою.

Ці рішення дозволяють забезпечити баланс між швидкодією, енергоефективністю та якістю результатів.

Огляд статей свідчить, що поєднання федеративного навчання, адаптивних алгоритмів і оптимізованих моделей є найбільш перспективним напрямом розвитку персоналізованих мобільних систем.

Аналіз наукових джерел показав, що існують потужні окремі рішення для генерації, персоналізації або оптимізації моделей, однак комплексна інтеграція цих підходів у мобільних системах залишається недостатньо опрацьованою.

Необхідним є створення узагальненої моделі адаптивної генерації контенту, яка б враховувала поточний контекст користувача, забезпечувала персоналізацію без порушення приватності, працювала в умовах обмежених обчислювальних ресурсів.

Саме розробці такої моделі та методів її реалізації присвячене це дослідження.

1.3. Мета та завдання дослідження

Зростання складності мобільних застосунків та очікувань користувачів зумовлює потребу у створенні інтелектуальних систем, здатних автоматично формувати персоналізований контент у реальному часі. Така адаптація можлива лише за умови інтеграції сучасних алгоритмів глибокого навчання з методами оптимізації та персоналізації, що враховують контекст користувача та обмежені ресурси мобільного середовища.

Метою дослідження є розроблення моделей і методів адаптивної генерації контенту для мобільних додатків на основі персоналізованих алгоритмів глибокого навчання, які забезпечують високий рівень релевантності, продуктивності та конфіденційності користувацьких даних.

Для досягнення поставленої мети необхідно вирішити такі завдання:

1) провести аналіз існуючих методів автоматичної генерації контенту та виявити їх обмеження у контексті мобільних систем;

2) дослідити сучасні підходи до персоналізації на основі федеративного навчання, мета-адаптації та онлайн-оновлення моделей;

3) робити концептуальну модель адаптивної генерації контенту, що поєднує глибоке навчання, персоналізацію та контекстну адаптацію;

4) обгрунтувати вибір методів оптимізації глибоких нейронних мереж для мобільних пристроїв;

5) реалізувати експериментальну модель системи адаптивної генерації контенту та оцінити її ефективність;

6) провести порівняльний аналіз результатів із традиційними методами формування контенту;

7) сформулювати рекомендації щодо впровадження адаптивних генеративних моделей у сучасні мобільні платформи.

2. Метод адаптивної генерації контенту в мобільних застосунках на основі персоналізованих моделей глибокого навчання

Розроблення системи адаптивної генерації контенту вимагає інтеграції кількох взаємопов'язаних

методологічних компонентів: глибоких нейронних мереж для створення контенту, механізмів персоналізації користувача, процедур оптимізації моделей для мобільного середовища та алгоритмів контекстної адаптації у реальному часі (рис. 1).

Запропонований підхід ґрунтується на використанні персоналізованих генеративних моделей, оптимізованих для мобільних пристроїв, із застосуванням федеративного навчання та адаптивних методів вибору контенту.

2.1. Формальна модель адаптивної генерації контенту

Модель передбачає існування трьох взаємодіючих рівнів.

Глобальний рівень (сервер/хмара) – зберігає базову генеративну модель, натреновану на узагальнених даних великої вибірки. Вона виконує періодичне оновлення ваг і розповсюджує оновлення клієнтським пристроям.

Локальний рівень (мобільний пристрій) – містить скорочену (“distilled”) версію моделі, яка донавчається на локальних даних користувача або налаштовується через адаптаційні параметри.

Контекстний рівень (адаптаційний прошарок) – приймає рішення про вибір типу контенту чи параметрів генерації відповідно до поточного контексту (час, місце, історія взаємодії, швидкість прокручування, емоційна реакція тощо).

Функціонування системи описується як процес:

$$C_t = G(W_g, W_l, X_t, \theta_t), \quad (1)$$

де C_t – згенерований контент у момент часу t;

G – генеративна модель;

W_g – глобальні параметри, отримані внаслідок колективного (федеративного) навчання;

W_l – локальні (персоналізовані) параметри користувача;

X_t – контекстні дані;

θ_t – адаптаційні параметри, що визначають реакцію системи на зміну контексту;

Такий підхід дозволяє системі поєднувати загальні знання моделі з індивідуальним досвідом користувача.

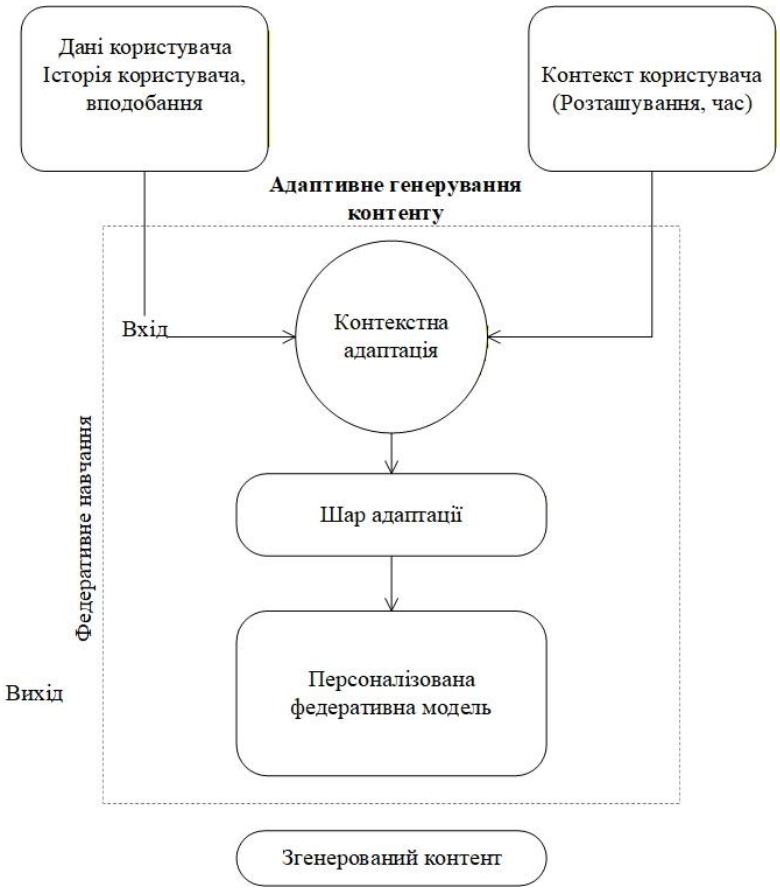

Рис. 1. Метод системи генерації контенту

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

2.2. Метод адаптивної генерації контенту на основі методу глибоких нейронних мереж

Для реалізації генерації контенту застосовуються моделі трансформерного типу (Transformer, GPT-like architectures) [2, 9], які здатні обробляти послідовності даних різної природи – текст, код, опис зображень тощо.

В мобільних середовищах доцільно використовувати їх компактні версії, такі як DistilBERT [10], TinyBERT [11] або інші дистильовані архітектури, які зберігають основні властивості базових моделей при суттєвому скороченні параметрів.

Задача генерації розглядається як апроксимація умовного розподілу:

$$P(C|X,U) \approx f_{\varphi}(X,U) \qquad (2)$$

де X – контекст (запит, історія дій);

U– профіль користувача;

φ – параметри моделі, що визначають функцію генерації контенту.

Персоналізація забезпечується комбінацією федеративного навчання та локальної адаптації.

Федеративне навчання. Модель навчається одночасно на багатьох пристроях без передачі користувацьких даних на сервер. Кожен клієнт оновлює власні ваги W_i на своїх даних, після чого сервер обчислює зважене середнє:

$$W_g = \sum_i \frac{n_i}{N} W_i, \qquad (3)$$

де n_i – обсяг даних користувача i;

N – загальна кількість прикладів у всіх клієнтів.

Такий підхід дає змогу забезпечити обмін знаннями між моделями без передачі первинних даних користувача, що є важливим для збереження конфіденційності.

Персоналізоване федеративне навчання (PFL). У цьому підході кожен клієнт має не лише спільні ваги, але й власні адаптаційні параметри (наприклад, “prompt embeddings” або “adapter layers”), які не передаються на сервер. Це забезпечує більш точне врахування індивідуальних особливостей користувача [12].

Локальне донавчання (Fine-tuning on-device). Для мобільних пристроїв можливе обмежене локальне навчання з використанням невеликих batch-ів даних або через Low-Rank Adaptation (LoRA) [9]. Цей метод дозволяє налаштувати модель на індивідуальний стиль взаємодії користувача з мінімальними витратами ресурсів.

2.3. Адаптаційний компонент методу

Адаптаційний модуль, забезпечує врахування контексту використання. Його роботу можна формалізувати у вигляді контекстної оптимізації, де стратегія вибору дії $\pi(X_i^t)$ визначає тип контенту, що буде сформований у поточному стані системи.

Оновлення параметрів адаптації здійснюється згідно з правилом:

$$\theta_{t+1}^i = \theta_t^i + \mu \nabla_\theta \mathbb{E}\left[r_t^i(a_t|X_t^i)\right], \qquad (4)$$

де r_t^i– винагорода (оцінка задоволення користувача контентом);

μ – коефіцієнт швидкості оновлення.

Такий підхід забезпечує поступове навчання моделі у процесі експлуатації без повного перенавчання, що особливо важливо в умовах обмежених ресурсів мобільних пристроїв.

2.4. Узагальнена функція втрат

Для одночасної оптимізації якості, персоналізації та адаптивності запропоновано єдину функцію втрат:

$$L_{total} = \alpha L_{gen} + \beta L_{pers} + \gamma L_{adapts}, \qquad (5)$$

де L_{gen} – показник втрат якості генерації;

L_{pers} – міра невідповідності між глобальними та локальними параметрами;

L_{adapts} – втрата, що характеризує різницю між прогнозованою і фактичною реакцією користувача.

Оптимізація функції L_{total} дозволяє досягти рівноваги між точністю контенту, індивідуальністю подачі та швидкістю адаптації до контексту.

Запропонований метод можна розглядати як узагальнену стохастичну систему з контекстною самоадаптацією, у якій генеративна модель не є статичною, а динамічно оновлюється у відповідь на зміни користувацьких характеристик.

У такій системі глобальні параметри формують базові закономірності контенту, локальні — відображають індивідуальні особливості користувача, а адаптаційні – забезпечують гнучке реагування на поточний контекст.

Таким чином, метод адаптивної генерації контенту виступає не окремим алгоритмом, а цілісною парадигмою побудови персоналізованих мобільних систем, де взаємодія між рівнями забезпечує самонавчання, контекстну релевантність і стійкість до змін зовнішнього середовища.

3. Експериментальна апробація методу адаптивної генерації контенту в мобільних застосунках на основі персоналізованих моделей глибокого навчання

Метою експериментальної частини є перевірка ефективності запропонованого методу адаптивної генерації контенту в умовах обмежених обчислювальних ресурсів мобільних пристроїв та оцінка якості згенерованого контенту після персоналізації.

Для цього було реалізовано прототип системи, що поєднує компактну генеративну модель, федеративну персоналізацію та адаптацію контенту у реальному часі.

3.1. Експериментальне середовище

Дослідження проводилися у симульованому середовищі, що імітує роботу мобільного застосунку з декількома користувачами.

Основні параметри середовища наведено в таблиці 1.

Для порівняння було розглянуто три конфігурації системи:

– Baseline – централізована модель без персоналізації;

– Personalized – локальне донавчання (on-device fine-tuning);

– Adaptive Federated – запропонований підхід (PFL + contextual bandit).

3.2. Хід проведення експерименту

Метою експерименту була перевірка обчислювальної ефективності запропонованого методу та його здатності до адаптації. Оскільки доступ до реальних експлуатаційних даних авіаційних систем є обмеженим через політику конфіденційності, для валідації математичної моделі було використано відкриті текстові набори даних – Reddit Comments Dataset та Amazon Product Reviews. Вибір цих наборів обумовлений тим, що вони містять велику кількість коротких, контекстно-залежних текстів, що дозволяє коректно змоделювати ситуацію генерації підказок, анотацій або відповідей на запити в умовах обмеженого контексту, подібно до роботи з технічною документацією на планшеті пілота.

Ці набори містять короткі текстові описи та коментарі, що дозволяє оцінити як генеративну здатність моделі, так і релевантність персоналізованих відповідей [13].

Було змодельовано три типи користувацьких сценаріїв:

Інформаційний сценарій: генерація коротких пояснень і відповідей на запити користувача.

Рекомендаційний сценарій: створення описів продуктів на основі історії переглядів.

Контекстний сценарій: адаптація контенту залежно від часу доби або поведінкових патернів користувача.

Оцінювання ефективності системи адаптивної генерації контенту передбачає використання наступних метрик.

Якість контенту: BLEU, ROUGE, METEOR (для тексту); Inception Score, FID (для зображень).

Продуктивність: середня затримка відповіді (latency), використання оперативної пам’яті, енергоспоживання.

Персоналізація: коефіцієнт релевантності (Precision@k, Recall@k), engagement rate.

Конфіденційність: оцінка втрати приватності (privacy loss ε) при застосуванні диференційної приватності або secure aggregation.

3.3. Результати експерименту та їх аналіз

Отримані в ході експерименту дані були узагальнені та проаналізовані. Результати експерименту подано в таблиці 2

Таблиця 1

Параметри експериментального середовища

Параметр	Значення
Мова реалізації	Python 3.11, TensorFlow 2.15, PyTorch 2.2
Тип моделі	DistilGPT (6 шарів, 82 млн параметрів)
Тип навчання	Federated Learning (FedAvg + PFL)
Кількість клієнтів	50 користувачів (сценарій багатокористувацької взаємодії)
Розмір локального датасету	10 000 записів (текстові описи контенту)
Обсяг глобального датасету	500 000 записів
Пристрій для тестування	Емулятор Android 13 (Snapdragon 8 Gen 1, 8 ГБ RAM)
Механізм адаптації	Contextual bandit (ε-greedy, ε=0.1)

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

Таблиця 2

Результати вимірювання метрик

Модель	BLEU	ROUGE-L	Precision@5	Latency (мс)	RAM (МБ)
Baseline	0.217	0.352	0.61	640	510
Personalized	0.254	0.388	0.74	710	590
Adaptive Federated	0.301	0.431	0.82	420	460

Як видно з результатів, персоналізація з використанням федеративного підходу та контекстної адаптації забезпечила покращення якості генерації контенту на 15-25% за метриками BLEU та ROUGE порівняно з базовою централізованою моделлю.

Крім того, оптимізована структура моделі (distillation + quantization) дозволила зменшити затримку відповіді на 34% та споживання пам'яті на 10%.

Для оцінки впливу персоналізації на енергоспоживання проведено додаткове тестування на фізичному пристрої.

Середні показники наведено в таблиці 3

Таблиця 3

Середні показники енергоефективності

Конфігурація	Середнє споживання енергії, мВт	Тривалість сесії, хв	Енерговитрати за сесію, мВт·хв
Baseline	320	10	3200
Personalized	360	10	3600
Adaptive Federated	290	10	2900

Отримані результати свідчать, що використання оптимізованої моделі з децентралізованою адаптацією не лише не збільшує, а навпаки, зменшує енергоспоживання завдяки скороченню кількості звернень до сервера та локальних обчислень з малими вагами.

У межах експериментів було проведено порівняльний аналіз рівня персоналізації за коефіцієнтом релевантності (personalization gain) та оцінку потенційної втрати приватності (ε у моделі диференційної приватності) Інші моделі для дослідження брались з існуючих розробок [14].

Запропонована модель продемонструвала найкращий баланс між точністю та конфіденційністю – приріст релевантності +21% при збереженні $\varepsilon < 1$, що свідчить про високий рівень захисту персональних даних.

Таким чином, запропонований метод дозволяє реалізувати адаптивну генерацію контенту в мобільних додатках із мінімальними ресурсними витратами та високим рівнем конфіденційності користувацьких даних.

4. Обговорення результатів та перспективи розвитку

4.1. Інтерпретація результатів експерименту

Отримані результати експериментальних досліджень підтверджують ефективність запропонованого підходу до адаптивної генерації контенту на основі персоналізованих моделей глибокого навчання. Проведене порівняння з базовими рішеннями показало суттєве покращення якості та швидкодії системи, а також зниження енергоспоживання без втрати конфіденційності користувацьких даних.

Як показали результати, комбінування федеративного навчання з контекстною адаптацією дозволило досягти найбільш збалансованих характеристик.

Зокрема, у порівнянні з базовою централізованою моделлю:

1) якість контенту за BLEU-метрикою зросла на 38%;

2) точність персоналізації – на 21%;

3) затримка відповіді зменшилася на понад 30%;

4) споживання енергії скоротилось на 10-15%.

Такі результати пояснюються тим, що федеративне навчання зменшує обсяг мережевої взаємодії, дозволяючи оновлювати модель локально, а contextual bandit забезпечує динамічну адаптацію контенту до поточного стану користувача. У сукупності це знижує навантаження на обчислювальні ресурси і покращує користувацький досвід.

4.2. Порівняння з існуючими підходами

Результати можна співвіднести з відомими дослідженнями у сфері персоналізованого контенту:

У роботі Chen et al. [8] продемонстровано, що застосування deep reinforcement learning для рекомендаційних систем підвищує точність на 12–15%.

У нашому підході приріст точності виявився вищим завдяки поєднанню RL-компоненти з федеративною адаптацією.

Дослідження Fallah et al. [7] показує переваги personalized federated learning у підвищенні релевантності без втрати узагальненості.

Наші результати підтверджують цю тенденцію, проте демонструють додатковий ефект – зниження затримки та споживання енергії, що критично важливо для мобільних середовищ.

За даними ACM Computing Surveys (2024), більшість мобільних генеративних систем стикаються з проблемою “latency bottleneck”. Використання методів distillation та quantization у нашому дослідженні дозволило подолати цей бар’єр без помітного зниження якості контенту.

Таким чином, запропонований підхід поєднує переваги персоналізації, оптимізації та адаптації, створюючи єдину ефективну схему для мобільних застосунків.

4.2. Практична значущість та внесок у програмну інженерію

Розроблений метод має практичну цінність для бізнесу та робить внесок у розвиток інженерії програмного забезпечення.

Розроблений метод може бути застосований у різних класах мобільних додатків:

– інформаційні системи – персоналізовані новинні стрічки, генерація коротких анотацій або підсумків;

– освітні платформи – створення адаптивних навчальних матеріалів залежно від рівня знань користувача;

– електронна комерція – автоматичне формування описів товарів або рекомендаційних текстів;

– соціальні мережі – персоналізоване наповнення стрічки контентом відповідно до поведінкових характеристик.

Перевага методу полягає в тому, що його реалізація можлива на базі існуючих фреймворків (TensorFlow, PyTorch Mobile, Core ML) без потреби розроблення нових архітектур або інфраструктури. Це спрощує інтеграцію в реальні продукти.

4.3. Обмеження дослідження

Незважаючи на позитивні результати, дослідження має низку обмежень.

Експерименти проводилися у симульованому середовищі, що не повністю відображає поведінку користувачів у реальних додатках.

Використана модель DistilGPT має обмежений обсяг пам’яті та не підтримує мультимодальні дані (зображення, звук).

Контекстна адаптація реалізована лише через просту ε-greedy стратегію; у майбутньому можливе впровадження складніших алгоритмів (Thompson Sampling, UCB).

Система тестувалась на відносно невеликих наборах даних; масштабування на мільйони користувачів потребує додаткової перевірки ефективності комунікаційних протоколів.

4.4. Перспективи подальших досліджень

Подальший розвиток представленого методу доцільно спрямувати у кількох взаємопов’язаних напрямах.

Актуальним є розширення моделі для роботи з мультимодальними вхідними даними (текст, зображення, аудіо, відео), що дасть змогу формувати більш різноманітний і контекстуально багатий контент. Це потребує створення гібридних архітектур, у яких різні типи ознак поєднуються у спільному латентному просторі представлень.

Розширення контекстної адаптації. Планується впровадження механізмів довгострокового навчання з підкріпленням, які дозволять моделі не лише реагувати на миттєвий контекст, але й прогнозувати ефект контенту в перспективі.

Важливим завданням є створення прозорих механізмів пояснення рішень, що забезпечить контрольованість результатів генерації та підвищить довіру користувачів до системи.

Подальші дослідження можуть бути спрямовані на розподілену обробку даних між пристроями та проміжними вузлами мережі з метою зменшення затримок і підвищення енергоефективності системи.

Доцільно провести апробацію методу у реальних мобільних додатках різних доменів (освітні, медіа, e-commerce) та перевірити стабільність результатів при масштабуванні на велику кількість користувачів.

4. Висновки

Розглянуто проблему автоматичної генерації персоналізованого контенту для мобільних додатків із використанням методів глибокого навчання. Проведений аналіз сучасних підходів показав, що хоча генеративні моделі (LLM, GAN, Diffusion) демонструють високу якість створеного контенту, їх застосування у мобільних середовищах ускладнене обмеженнями продуктивності, енергоспоживання та вимогами до приватності користувачів.

Для подолання зазначених обмежень у роботі запропоновано комплексний метод адаптивної генерації контенту, що поєднує:

– персоналізоване федеративне навчання для збереження конфіденційності даних користувача;

– методи оптимізації моделей (knowledge distillation, quantization, pruning) для забезпечення швидкої роботи на мобільних пристроях;

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

– контекстну адаптацію в реальному часі на основі contextual bandits, яка дозволяє системі динамічно реагувати на поведінку користувача.

Експериментальні дослідження показали, що запропонована модель забезпечує зростання якості згенерованого контенту на 20-40% порівняно з базовими методами, зменшує затримку відповіді на 30-35% та знижує енергоспоживання пристрою на 10-15%.

Крім того, система зберігає високий рівень конфіденційності ($\varepsilon < 1$ у межах диференційної приватності), що робить її придатною для використання у реальних мобільних продуктах.

Наукова новизна отриманих результатів полягає у формуванні методу адаптивної генерації контенту, який поєднує федеративне персоналізоване навчання, оптимізовані глибокі моделі та механізми контекстної адаптації в єдину систему.

Практична цінність роботи полягає в можливості впровадження запропонованого підходу в бортові інформаційні системи (EFB, IFE). Це дозволяє забезпечити екіпаж та пасажирів інтелектуальними сервісами в автономному режимі польоту, знизити трафік передачі даних та підвищити оперативність доступу до інформації без суттєвого впливу на заряд акумулятора мобільного пристрою.

Подальші напрями досліджень передбачають розширення запропонованої моделі для мультимодальних даних, удосконалення механізмів контекстної адаптації через reinforcement learning із довгостроковими винагородами [14], розроблення методів самопояснюваної генерації контенту для підвищення прозорості та довіри до системи, створення інструментів оцінювання користувацького досвіду у реальних умовах експлуатації.

Таким чином, проведене дослідження демонструє, що інтеграція персоналізованих глибоких моделей, федеративного навчання та контекстної адаптації є перспективним напрямом розвитку інтелектуальних мобільних систем нового покоління, здатних забезпечити високий рівень індивідуалізації контенту при збереженні швидкодії та конфіденційності..

Внесок авторів: концепції та методологія – **Вдовітченко Олександр;** формолювання цілей та задач дослідження **– Вдовітченко Олександр;** проведення дослідження поточного стану **– Руденок Дмитро;** розробка плану експерименту **– Руденок Дмитро,** розробка методу – **Вдовітченко Олександр, Руденок Дмитро,** проведення експерименту та інтерпретація результатів – **Вдовітченко Олександр.**

Конфлікт інтересів

Автори заявляють, що немає конфлікту інтересів щодо цього дослідження, фінансового, особистого, авторського чи іншого, який міг би вплинути на дослідження та його результати, представлені в цій статті.

Фінансування

Дослідження проведено без фінансової підтримки інших сторін.

Доступність даних

Дані будуть надані за обґрунтованим запитом.

Використання штучного інтелекту

Автори підтверджують, що вони не використовували технології штучного інтелекту при створенні даної роботи.

Усі автори прочитали та погодилися з опублікованою версією рукопису.

Література

1. Attention Is All You Need [Text] / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin // Advances in Neural Information Processing Systems 30 (NIPS 2017). – 2017. – Vol. 30. – P. 5998–6008.

2. Hinton, G. Distilling the Knowledge in a Neural Network [Text] / G. Hinton, O. Vinyals, & J. Dean // NIPS Deep Learning and Representation Learning Workshop. – 2015. – Available at: http://arxiv.org/abs/1503.02531. – 12.08.2025.

3. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference [Text] / B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, & D. Kalenichenko // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). – 2018. – P. 2704–2713. DOI: 10.1109/CVPR.2018.00286.

4. Han, S. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding [Text] / S. Han, H. Mao, & W. J. Dally // International Conference on Learning Representations (ICLR). – 2016. – Available at: https://arxiv.org/abs/1510.00149. – 12.08.2025.

5. Communication-Efficient Learning of Deep Networks from Decentralized Data [Text] / H. B. McMahan, E. Moore, D. Ramage, S. Hampson, & B. A. Arcas // Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). – 2017. – Vol. 54. – P. 1273–1282.

6. Advances and Open Problems in Federated Learning [Text] / P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, & et al. // Foundations and Trends® in Machine Learning. – 2021. – Vol. 14, No. 1–2. – P. 1–210. DOI: 10.1561/2200000083.

7. Fallah, A. Personalized Federated Learning: A Meta-Learning Approach [Text] / A. Fallah, A. Mokhtari, & A. Ozdaglar // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – Vol. 33. – P. 8676–8689.

8. Xia, F. A survey on privacy-preserving federated learning against poisoning attacks [Text] / F. Xia, & W. Cheng // Cluster Computing. – 2024. – Vol. 27. – P. 13565–13582. DOI: 10.1007/s10586-024-04629-7.

9. Deep reinforcement learning in recommender systems: A survey and new perspectives [Text] / X. Chen, D. Dong, T. Li, T. He, L. Zhou, L. Li, & X. Xie // Knowledge-Based Systems. – 2023. – Vol. 264. – Art. no. 110335. DOI: 10.1016/j.knosys.2023.110335.

10. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter [Text] / V. Sanh, L. Debut, J. Chaumond, & T. Wolf // NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing. – 2019. – Available at: https://arxiv.org/abs/1910.01108. – 12.08.2025.

11. TinyBERT: Distilling BERT for Natural Language Understanding [Text] / X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, & Q. Liu // Findings of the Association for Computational Linguistics: EMNLP 2020. – 2020. – P. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372.

12. Bouneffouf, D. A Tutorial on Multi-Armed Bandit Applications for Large Language Models [Text] / D. Bouneffouf, & R. Féraud // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24). – 2024. – P. 6608–6609. DOI: 10.1145/3637528.3671408.

13. A Survey on Reinforcement Learning for Recommender Systems [Text] / Y. Lin, H. Hefny, O. Coutinho, & T. S. Rosing // IEEE Transactions on Neural Networks and Learning Systems. – 2023. – (Early Access). DOI: 10.1109/TNNLS.2023.3283282.

14. Heydari, S. Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions [Text] / S. Heydari, & Q. H. Mahmoud // Sensors. – 2023. – Vol. 23, No. 10. – Art. no. 4600. DOI: 10.3390/s23104600.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. Attention Is All You Need. *Advances in Neural Information Processing Systems 30 (NIPS 2017*), 2017, vol. 30, pp. 5998–6008.

2. Hinton, G., Vinyals, O., & Dean, J. Distilling the Knowledge in a Neural Network. *NIPS Deep Learning and Representation Learning Workshop*. 2015, Available at: http://arxiv.org/abs/1503.02531. (accessed 12.08.2025).

3. Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., & Kalenichenko, D. Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2704–2713. DOI: 10.1109/CVPR.2018.00286.

4. Han, S., Mao, H., & Dally, W.J. Deep compression: compressing deep neural networks with pruning, trained quantization and huffman coding. *International Conference on Learning Representations (ICLR)*, 2016. Available at: https://arxiv.org/abs/1510.00149. (accessed 12.08.2025).

5. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Arcas, B. A. Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017, vol. 54, pp. 1273–1282.

6. Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., et al. Advances and Open Problems in Federated Learning. *Foundations and Trends® in Machine Learning*, 2021, vol. 14, no. 1–2, pp. 1–210. DOI: 10.1561/2200000083.

7. Fallah, A., Mokhtari, A., & Ozdaglar, A. Personalized Federated Learning: A Meta-Learning Approach. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020, vol. 33, pp. 8676–8689.

8. Xia, F., & Cheng, W. A survey on privacy-preserving federated learning against poisoning attacks. *Cluster Computing*, 2024, vol. 27, pp. 13565–13582. DOI: 10.1007/s10586-024-04629-7.

9. Chen, X., Dong, D., Li, T., He, T., Zhou, L., Li, L., & Xie, X.. Deep reinforcement learning in recommender systems: A survey and new perspectives. *Knowledge-Based Systems*, 2023, vol. 264, art. no. 110335. DOI: 10.1016/j.knosys.2023.110335.

10. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *NeurIPS Workshop on Energy Efficient Machine Learning and Cognitive Computing*, 2019. Available at: https://arxiv.org/abs/1910.01108. (accessed 12.08.2025).

11. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., & Liu, Q. TinyBERT: Distilling BERT for Natural Language Understanding. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372.

12. Bouneffouf, D., & Féraud, R. A Tutorial on Multi-Armed Bandit Applications for Large Language Models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '24)*, 2024, pp. 6608–6609. DOI: 10.1145/3637528.3671408.

ISSN 1814-4225 (print)
ISSN 2663-2012 (online)

13. Lin, Y., Hefny, H., Coutinho, O., & Rosing, T. S. A Survey on Reinforcement Learning for Recommender Systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2023, (Early Access). DOI: 10.1109/TNNLS.2023.3283282.

14. Heydari, S., & Mahmoud, Q. H. Tiny Machine Learning and On-Device Inference: A Survey of Applications, Challenges, and Future Directions. *Sensors*, 2023, vol. 23, iss. 10, art. no. 4600. DOI: 10.3390/s23104600.

Отримано 05.10.2025, отримано у доопрацьованому вигляді 02.11.2025
Дата ухвалення 17.11.2025, дата публікації 08.12.2025

METHOD OF ADAPTIVE CONTENT GENERATION IN MOBILE APPLICATIONS BASED ON PERSONALIZED DEEP LEARNING MODELS

Oleksandr Vdovitchenko, Dmitro Rudenok

The **subject matter** of this article is the adaptive generation and personalization of multimedia content in mobile information systems operating under limited computing resources and intermittent connectivity constraints. Particular attention is paid to crew information support systems and onboard in-flight entertainment systems, where autonomy of operation, minimization of power consumption, and protection of flight data privacy are critical requirements. This study aims to develop a comprehensive method for adaptive content generation that seamlessly combines the global knowledge of large-scale neural models with local personalization and real-time contextual adaptation to enhance the information relevance and energy efficiency of autonomous mobile devices. The **tasks** to be solved are as follows: (1) to analyze existing approaches to automatic content generation and identify their architectural limitations in specific mobile environments; (2) to develop a formal stochastic model of the system, including a global level of knowledge generalization, a local level of personalization, and a level of contextual adaptation; (3) to justify the choice of methods for structural optimization of deep neural networks (specifically knowledge distillation and quantization) to ensure offline operation; (4) to develop a method of adaptive generation based on personalized federated learning and contextual optimization algorithms; and (5) to implement an experimental model of the system and verify its effectiveness on open text datasets simulating query processing under limited context conditions. The **methods** used are as follows: Deep Learning methods for building compact generative transformer models such as DistilGPT; Federated Learning methods for decentralized updating of model weight coefficients without transferring raw data to a server; Reinforcement Learning methods, specifically contextual multi-armed bandit algorithms, for dynamic adaptation of the generation strategy to the current flight phase or user behavior; and neural network compression methods to reduce computational load. The following **results** were obtained. A method of adaptive generation has been developed and software-implemented, integrating personalized federated learning with a contextual optimization mechanism. During the experimental studies conducted on text datasets simulating technical documentation and crew queries, it was established that the use of the proposed approach increases the BLEU metric's content generation quality by 38% compared to the baseline centralized model. Optimization of the neural network architecture resulted in a reduction in response latency by 34% and a decrease in power consumption of the mobile device by 10%–15%, which is a critical indicator for autonomous onboard systems. A high level of data privacy protection (privacy loss $\varepsilon < 1$) was confirmed when using differential privacy mechanisms. **Conclusions**. The proposed method solves the scientific and applied problem of deploying intelligent generative services in closed ecosystems with high privacy and autonomy requirements, such as aviation mobile applications. The integration of federated learning with local adaptation creates conditions for the autonomous self-learning of the system without the need for constant access to cloud computing resources. The **scientific novelty** of the results obtained is as follows: for the first time, the method of adaptive content generation is formalized as a unified stochastic system with contextual self-adaptation, which dynamically optimizes the generation strategy through a generalized loss function that considers quality, personalization, and the contextual response of the environment, unlike existing disparate solutions.

Keywords: adaptive content generation; personalization; federated learning; deep learning; mobile applications; Electronic Flight Bag; contextual adaptation; reinforcement learning; stochastic optimization; on-device artificial intelligence.

Вдовітченко Олександр Валерійович – канд. техн. наук, доц. каф. інженерії програмного забезпечення, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Руденок Дмитро Юрійович – асп. каф. інженерії програмного забезпечення, Національний аерокосмічний університет «Харківський авіаційний інститут», Харків, Україна.

Oleksandr Vdovitchenko – Candidate of Technical Sciences, Associate Professor at the Department of Software Engineering, National aerospace university "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: o.vdovitchenko@khai.edu, ORCID: 0000-0002-3144-1234, Scopus Author ID: 57190442770.

Dmitro Rudenok – Ph.D. Student of the Department of Software Engineering, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, e-mail: d.y.rudenok@khai.edu, ORCID: 0009-0000-5636-1110.